Synthetic Biology Roadmap 2030

The Engine of Next-Generation Biomanufacturing

Synthetic Biology Development Strategy Research Group,
China Society of Biotechnology

SCIENCE PRESS

edp sciences

Synthetic biology has garnered widespread attention for its scientific merit and applicative value. It is regarded as a new paradigm in life science research ("build to learn"), and a core driving force for the iterative advancement of biotechnology and the transformative development of biomanufacturing ("build to use"). The techno-economic core thread of "Synthetic Biology – Biotechnology – Biomanufacturing – Bioeconomy is striking a new chord of the times for the common development of humanity.

Synthetic Biology Roadmap 2030 is the outcome of in-depth collaborative research by numerous scholars. Drawing fully upon relevant international roadmaps and strategic development reports, and integrating the latest advances in the field, it proposes for the first time a four-in-one disciplinary architecture for synthetic biology encompassing "theoretical connotations, enabling technologies, application prospects, and governance principles", while rationally forecasting the development goals and implementation pathways to 2030.

As China's contribution to the global landscape of synthetic biology roadmaps, this work will serve as a vital reference for research, education, technology transfer, public engagement, and public literacy in the field of synthetic biology. Furthermore, we hope it will provide decision-making support for national scientific research planning and deployment, disciplinary development and industrial growth of synthetic biology, empowering it to fuel new quality productive forces in biomanufacturing.

ISBN : 978-2-7598-4052-6

9 782759 840526

edp sciences
www.edpsciences.org

经典中国国际出版工程
China Classics International

SCIENCE PRESS

Synthetic Biology Development Strategy Research Group, China Society of Biotechnology

Synthetic Biology Roadmap 2030

The Engine of Next-Generation Biomanufacturing

SCIENCE PRESS

edp sciences

Printed in France

EDP Sciences – ISBN(print): 978-2-7598-4052-6 – ISBN(ebook): 978-2-7598-4056-4
DOI: 10.1051/978-2-7598-4052-6

Published by Science Press, EDP Sciences, 2026

Synthetic Biology Development Strategy Research Group

Research Group Leader:

Zhang Xian-En[1, 2]

Research Group Members (Arrange in alphabetical order):

Bai Wen-Qin[3]	Chen Fang[4]	Chen Peng[5]	Cui Zong-Qiang[6]
Dai Jun-Biao[7]	Dai Lei[8]	Dai Zong-Jie[3]	Du Li[9]
Feng Jin-Hui[3]	Feng Yan[10]	Fu Mei-Fang[8]	Fu Xiong-Fei[8]
Gan Hai-Yun[8]	Gao Cai-Xia[11]	Ge Yun[12]	Hao Zi-Yang[13]
Hu Qiang[1]	Hu Zheng[8]	Huang He[14]	Huang Jian-Dong[15]
Jiang Hui-Feng[3]	Jiang Jian-Dong[16]	Jin Cheng[17]	Jin Fan[8]
Lei Rui-Peng[18]	Li Bing-Zhi[14]	Li Chun[19]	Li Feng[6]
Li Jian[20]	Li Jin-Gen[3]	Li Xue-Fei[8]	Li Yin[17]
Li Yu-Juan[8]	Lian Jia-Zhang[21]	Lin Min[22]	Lin Zhang-Lin[23]
Liu Chen-Li[8]	Liu Hai-Tao[24]	Liu Hai-Yan[25]	Liu Tao[3]
Liu Wan[26]	Liu Xing-Guo[27]	Lou Chun-Bo[8]	Lu Yuan[19]
Luo Xiao-Zhou[8]	Ma Ying-Fei[8]	Miao Wei[28]	Pan Hong[8]
Pang Dai-Wen[29]	Qi Fei[8]	Qin Jian-Hua[24]	Qin Lei[19]
Qu Ge[3]	Shen Yue[30]	Shi Jia-Fu[14]	Shi Shuo-Bo[31]
Si Tong[8]	Song Hao[14]	Song Mao-Yong[32]	Sun Zhou-Tong[3]
Tang Lei-Han[33]	Tao Ting-Ting[24]	Wang Guo-Yu[34]	Wang Hao-Yi[35]
Wang Jie[36]	Wang Jin[22]	Wang Meng[3]	Wang Qin-Hong[3]

Wang Xiang-Xi[2]	Wang Xiao-Wo[19]	Wang Ya-Jie[33]	Wang Ya-Qing[24]
Wang Yong[37]	Wei Ping[8]	Wei Wen-Sheng[5]	Wei Xin-Li[17]
Wei Zheng[19]	Wu Bian1[7]	Wu Xiao-Lei[5]	Xiang Hua[3]
Xie Zhen[19]	Xiong Yan[26]	Xu Ping[10]	Yan Fei[8]
Yan Xing[37]	Yang Guang-Yu[10]	Yang Hui[38]	Yang Yi[39]
Yao Bin[40]	Ye Hai-Feng[41]	You Chun[10]	Yu Tao[8]
Yuan Shu-Guang[8]	Zhang Cheng-Cai[28]	Zhang Chong[19]	Zhang Li-Xin[42]
Zhang Xu[24]	Zhao Yong[1]	Zheng Hao[43]	Zheng Hong-Chen[3]
Zhong Chao[8]	Zhou Jing-Wen[44]	Zhou Ning-Yi[10]	Zhou Yong-Jin[24]
Zhou Zhi-Hua[37]	Zhu Hua-Wei[17]	Zhu Jian-Kang[36, 45]	Zhu Zhi-Guang[3]

Consultants (Arrange in alphabetical order):

Cao Zhu-An	Chen Jian	Chen Run-Sheng	Chen Wei
Chen Ye-Guang	Chong Kang	Deng Zi-Xin	Fan Chun-Hai
Gao Fu	Han Bin	Han Jia-Huai	He Fu-Chu
Jiang Gui-Bin	Kang Le	Li Jia-Yang	Li Lin
Liu Chang-Sheng	Ma Yan-He	Ouyang Qi	Rao Zi-He
Tan Tian-Wei	Tan Wei-Hong	Tang Chao	Tian Zhi-Gang
Yan Xi-Yun	Yang Huan-Ming	Yang Sheng-Li	Yuan Ying-Jin
Zhang You-Ming	Zhang Yu-Kui	Zhao Guo-Ping	Zhao Jin-Dong
Zhu Bing	Zhu Yu-Xian		

Working Group Leader:

Fu Xiong-Fei[8]

Working Group Members:

Li Yu-Juan[8]	Li Xue-Fei[8]	Qi Fei[8]	Wu Wei[8]
Huang Yi[8]	Li Min[2]		

1 Shenzhen University of Advanced Technology
2 Institute of Biophysics, Chinese Academy of Sciences
3 Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences
4 National Science Library (Chengdu), Chinese Academy of Sciences
5 Peking University
6 Wuhan Institute of Virology, Chinese Academy of Sciences
7 Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences
8 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
9 University of Macau
10 Shanghai Jiao Tong University
11 Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
12 Shenzhen Bay Laboratory
13 Capital Medical University
14 Tianjin University
15 The University of Hong Kong
16 Nanjing Agricultural University
17 Institute of Microbiology, Chinese Academy of Sciences
18 Huazhong University of Science and Technology
19 Tsinghua University
20 ShanghaiTech University
21 Zhejiang University
22 Biotechnology Research Institute, Chinese Academy of Agricultural Sciences
23 Guangdong University of Technology
24 Dalian Institute of Chemical Physics, Chinese Academy of Sciences
25 University of Science and Technology of China
26 Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences
27 Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences
28 Institute of Hydrobiology, Chinese Academy of Sciences
29 Nankai University
30 BGI Research
31 Beijing University of Chemical Technology
32 Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences
33 Westlake University
34 Fudan University
35 Institute of Zoology, Chinese Academy of Sciences
36 Southern University of Science and Technology

37 CAS Center for Excellence in Molecular Plant Science/Institute of Plant Physiology and Ecology, Chinese Academy of Sciences
38 Shanghai Institute of Materia Medica, Chinese Academy of Sciences
39 East China University of Science and Technology
40 Institute of Animal Science, Chinese Academy of Agricultural Sciences
41 East China Normal University
42 Henan University
43 China Agricultural University
44 Jiangnan University
45 Macau University of Science and Technology

Foreword One

Synthetic biology, an emerging branch of life sciences in the 21st century, has evolved into a highly prominent discipline within just two decades—something we might not have anticipated when first encountering the concept.

Life sciences have long advanced under the framework of reductionism. The proposal of the DNA double-helix model propelled life sciences into the era of molecular biology, then the Human Genome Project enabled scientists to systematically explore and study life activities and organisms starting from the genome. This marked a shift from experimental science to systems and predictive science. Furthermore, computational biology has built mathematical models of life activities based on quantitative systems biology. Against this backdrop, synthetic biology emerged. As a nascent field in life sciences, its development vividly embodies the integration of science and engineering.

Reflecting on the evolution of synthetic biology, I am amazed by its rapid progress. From the standardization of basic bioparts to the design and construction of biological circuits, and further to the optimization and regulation of biological systems, we have gradually achieved rational design and editing of life systems, opening up new frontiers in biotechnology. The innovative applications of synthetic biology have rapidly expanded into fields such as medicine, industry, agriculture, energy, environment, materials, and information, driving next-generation biomanufacturing and the future bioeconomy.

In China, the Ministry of Science and Technology has significantly enhanced the country's research capabilities in synthetic biology through strategic layout, from early-stage arrangements under the 973 Program and 863 Program to systematic support in the "13th Five-Year Plan" key projects and strengthened backing in the "14th Five-Year Plan". This progress has established China as one of the leading forces in the global synthetic biology community. In this context, Chinese Society of Biotechnology proposed and organized a strategic study on the development of synthetic biology towards 2030, offering forward-thinking insights and strategic planning. The compilation of this book, which embodies these efforts, holds profound strategic significance. Bringing together the wisdom of over 100 distinguished scholars from more than 40

universities, research institutions, and a group of industry professionals spanning natural sciences, engineering sciences, and social sciences, this book clarifies research directions and goals towards 2030 and proposes governance principles for synthetic biology development based on international scientific consensus. It thus carries significant academic reference value and industrial guiding significance.

Synthetic biology is far from reaching its peak. With the iterative development and the continuously expanding of enabling applications, it will undoubtedly play a central role in driving the future bioeconomy and provide innovative solutions for global sustainable development. The publication of this book will serve as a crucial reference for the future disciplinary construction and industrial advancement of synthetic biology.

Yang Sheng-Li
May 2024

Foreword Two

The great differences and close interconnections between the study objects of the natural sciences (physics, chemistry, biology, geology, astronomy) on the spatial and temporal scales determine both the differentiation of the disciplines and their intersection. From the late 19th century to the early 20th century, biology evolved from a "descriptive" phase to an "analytical" one, entering the research stage of "life sciences" which focused primarily on understanding the universal mechanisms of life activities. Bolstered by interdisciplinary integration and technological innovations, two major revolutions of "molecular biology" and "genomics" emerged in the mid-to-late 20th century.

At the beginning of 21th century, "synthetic biology" was redefined through the introduction of engineering concepts and research paradigms, opening up a new era of "convergent" research in the life sciences. In 2009, synthetic biology was in its infancy. The Royal Society and Royal Academy of Engineering in the United Kingdom (UK) have suggested that the National Academy of Sciences and National Academy of Engineering in the United States of America (USA) and the Chinese Academy of Sciences and Chinese Academy of Engineering organize and convene a symposium on synthetic biology involving the six academies of the above three countries. After two years of preparation, three conferences were held in 2011–2012 in London, Shanghai, and Washington D C. These conferences comprehensively discussed connotations, technologies, platforms, scientific and economic significance, related social, ethical, and cultural issues, as well as policies and governance—playing a pivotal role in shaping the field's development over the subsequent decade. Since then, mainly based on the fact that synthetic biology has continuously demonstrated its powerful "enabling" potential for life sciences and biotechnology, and its significant impact on the iterative upgrading and even "disruptive" breakthroughs of societal productivity. Countries and regions including the United Kingdom, the European Union, the United States, Canada and Australia have successively released 11 roadmaps on synthetic biology by 2023, mainly related to bioeconomy and macro-national strategies, covering semiconductor synthetic biology, microbiomics, biomaterials, national defense, and climate change, etc.

At present, China's synthetic biology research has achieved a series of major breakthroughs. For example, yeast chromosome synthesis and chromosome engineering, carbon dioxide resource utilization and synthesis of high-value compounds, analysis of biosynthetic pathways of a series of important natural products and synthesis and industrial transformation of artificial organisms, new gene editing technologies, and computer-aided design of new enzymes. However, we still lack a well-developed discipline system structure. This not only affects the national strategic planning and forward-looking layout, as well as the formation of an ecosystem conducive to the participation of the whole society in "convergence", but also affects the industry to create an enabling channel for "translational research" and establish a scientific and efficient management system based on the strengthening of scientific research on biosafety and ethical risk regulation.

The publication of this book is both a "timely rain" and a "guiding light", complementing the "Synthetic Biology Development Strategy of China Towards 2035" jointly released by the Chinese Academy of Sciences and the National Natural Science Foundation of China. This book is different from the international published roadmaps in that it innovatively proposes a multi-scale theoretical framework for synthetic biology, explaining the integration of the "white-box" of biological principles and the "black-box" of artificial intelligence. The application section highlights the core concepts of "build to learn" and "build to use" in synthetic biology. By reading this book, readers can gain an understanding of both the present and future of synthetic biology from four perspectives, namely basic theories, enabling technologies, application prospects, and capacity building with governance principles.

This book unites the ideological contributions and theoretical practices of many experts and scholars, and proposes for the first time that synthetic biology be constructed as an emerging disciplinary system, reflecting a clearer disciplinary development vein with characteristics of the times, which is of great scientific significance. I believe that the publication of this book will provide an important reference for the research deployment, platform construction, talent cultivation, international cooperation and industrial policy of synthetic biology, and will also become an important form of communication with international counterparts and the public.

Zhao Guo-Ping
May 2024

Preface

Synthetic biology is an emerging interdisciplinary cutting-edge field. Recognized globally for its scientific and industrial potential, synthetic biology is regarded as a new paradigm in life sciences ("build to learn") and a core driving force for the iterative improvement of biotechnology and transformative development of biomanufacturing ("build to use").

From 2009 to 2011, six prestigious academies—the Royal Society and Royal Academy of Engineering in the United Kingdom (UK), the National Academy of Sciences and National Academy of Engineering in the United States of America (USA), and the Chinese Academy of Sciences and Chinese Academy of Engineering—jointly organized a landmark series of international symposia on synthetic biology. The "Three Countries, Six Academies" symposium series focused on discussions on connotation, significance, scientific advancements, technological innovations, platforms and policy frameworks. Since then, major developed countries have formulated and published relevant development roadmaps and strategic reports. These roadmaps and reports have become the main reference for *Bold Goals for U.S. Biotechnology and Biomanufacturing* issued by the United States in 2023. The plan refers to synthetic biology as an "emerging biotechnology" and emphasizes that the achievement of this ambitious goal depends on "breakthroughs in synthetic biology and artificial intelligence".

China attaches great importance to the development of synthetic biology. The Ministry of Science and Technology, from the preliminary layout of the 863 and 973 Programs to the systematic layout of the 13th Five-Year Plan as well as the strengthened support of the 14th Five-Year Plan, has greatly enhanced China's research strength in synthetic biology, making China one of the major forces in the field of synthetic biology at the international level. In the new era, biomanufacturing and bioeconomy have been listed as national strategies and rapidly become the focus of the whole society. The National Development and Reform Commission has released China's first bioeconomy plan, *14th Five-Year Plan for Bioeconomy Development*, while the Ministry of Science and Technology has deployed a number of key R&D specialized programs, including

synthetic biology, green manufacturing, and the integration of biology and information technology (BT and IT integration). Some local governments have also established specialized programs for synthetic biology and biomanufacturing, with some regions setting up policy-supporting industrial parks, nurturing a large number of startups and even listed companies, attracting active participation from investors. The axis of "Synthetic Biology-Biotechnology-Biomanufacturing-Bioeconomy" is striking a new keynote for realizing the goal of building a world leader in science and technology and advancing the common development of humanity.

In order to make forward-looking thinking and strategic planning for the future development of the field of synthetic biology, the Chinese Society of Biotechnology has carried out a research on the development strategy of synthetic biology towards 2030, which is organized and implemented by the Division of Synthetic Biology.

The Division of Synthetic Biology, Chinese Society of Biotechnology, serves as the "Home of Synthetic Biology Professionals in China", bringing together the leading experts and scholars in the field of synthetic biology. In order to accomplish this major academic task, the Division of Synthetic Biology set up a research group and a working group. The research group includes more than 100 heavyweight scholars from more than 40 universities and research institutes and a group of business people. Based on the full reference to the synthetic biology roadmaps and strategy reports released by various countries, combined with the research progress in the field, the first draft was formed through repeated and in-depth discussions, while more than 30 scholars and senior experts from the Chinese Academy of Sciences and the Chinese Academy of Engineering were consulted to write this book.

The content of this roadmap is divided into four aspects under the "3+1" framework.

The first part is theoretical framework. So far, the theoretical part has not been reflected in the roadmaps that have been published internationally. Through years of scientific practice and academic activities such as the Xiangshan Science Conference, Chinese scholars have developed important ideas and consensus, which are discussed theoretically in this roadmap. This framework consists of two main parts: multi-scale theoretical framework of synthetic biology and artificial intelligence for synthetic biology. This is a distinctive feature of the roadmap.

The second part focuses on 12 key enabling technologies, including DNA sequencing, synthesis and assembly, gene editing, protein design, genetic circuits, chassis cells, cell-free systems, artificial multicellular systems, organoid engineering, unnatural

amino acids encoding and synthetic biosystems, biotic-abiotic hybrid systems, biofoundries, and bioparts data and information platforms. It provides a comprehensive discussion predicting the developmental milestones and objectives of enabling technologies until 2030. Compared with existing international roadmaps, it reflects new progress, with appropriate expansion of content and adjustment of some objectives.

The third part emphasizes the application prospects of synthetic biology, which includes two aspects. One is "build to learn", i.e. single-cell de novo synthesis, which is a concentrated manifestation of various synthetic biology enabling technologies. It is not a general application, but rather aims to achieve a highly challenging scientific and engineering goal. The other is "build to use", i.e. synthetic biology promotes biomanufacturing and bioeconomy, serving as a guide for the application of synthetic biology in fields such as industry, medicine, agriculture, food, environment, and information convergence.

The fourth part is governance strategy. It involves ethical considerations, legal regulation, talent cultivation, financial guarantee, academic organizations and international exchanges, and public science popularization of synthetic biology, etc., and proposes governance principles for the development of synthetic biology based on the consensus of international scientific and technological communities, so as to promote the healthy development of synthetic biology together with international counterparts and the public.

In terms of content, this is a complete roadmap for synthetic biology. Based on the many versions of the roadmap released by various countries, it proposes for the first time a four-pronged synthetic biology disciplinary architecture of "theoretical framework, enabling technologies, application prospects and governance strategy". We shared this architecture at the EBRC Global Forum 2.0 in 2023, which had been recognized by international peers. This roadmap is the Chinese version of the world's synthetic biology development roadmap, which will provide an important reference for the construction of synthetic biology discipline system, international communication and cooperation, deployment of national science and technology programs, and the development of the next-generation bio-industry.

This roadmap brings together the wisdom and hard work of many scholars and industry representatives and is gratefully acknowledged.

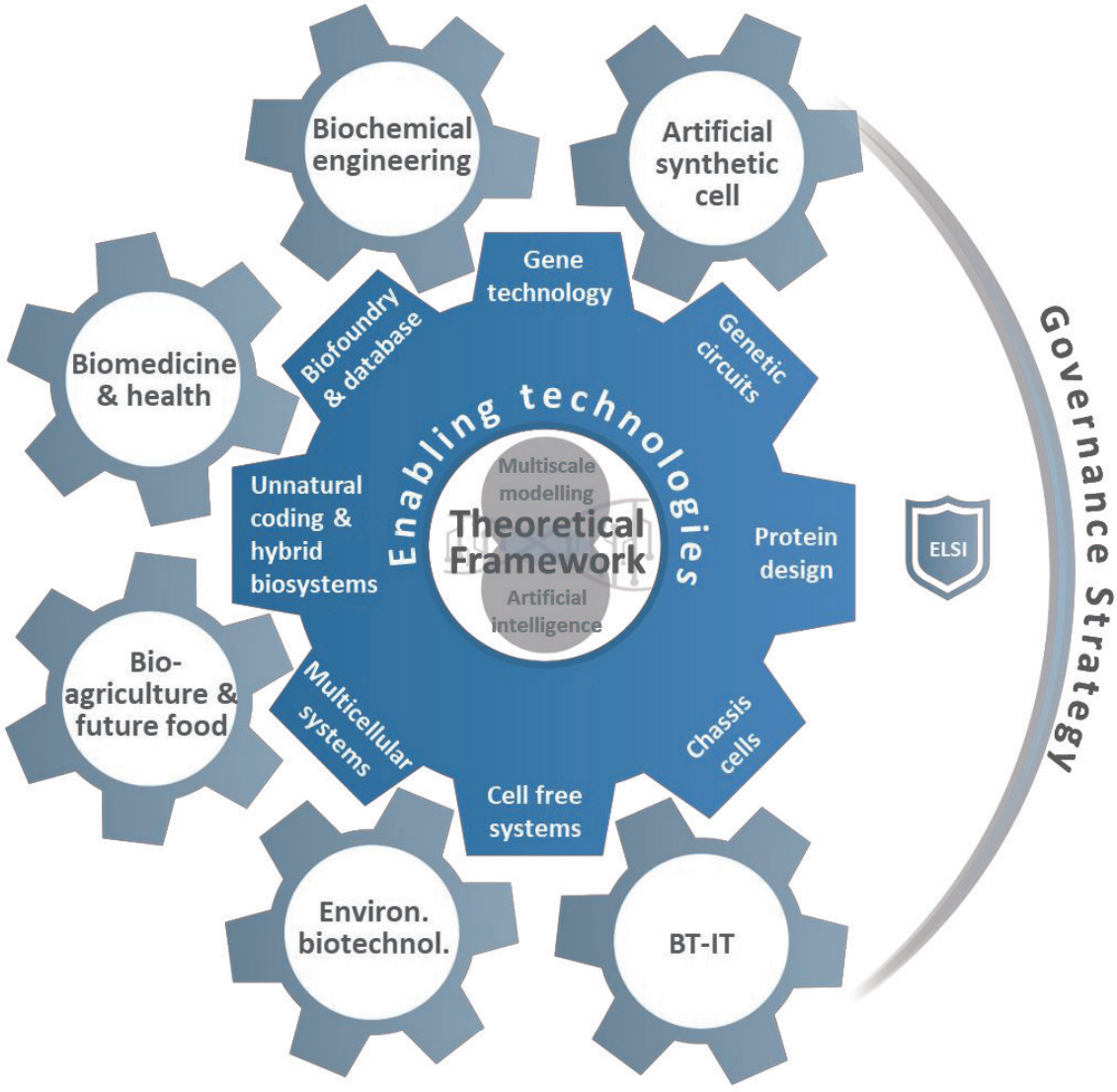
Back then, Dr. Yang Shengli, Dr. Ouyang Pingkai, Dr. Cao Zhu'an, etc., prospectively proposed the research of synthetic biology in the National 973 Program. Dr.

Zhao Guoping, on behalf of the Chinese side, participated in organizing and hosting the series symposia of "Three Countries, Six Academies" and made excellent contributions to the early practice, team building and base building with Dr. Yang Huanming, Dr. Ouyang Qi, Dr. Deng Zixin, Dr. Zhao Jindong, Dr. Ma Yanhe, Dr. Yuan Yingjin, and Dr. Liu Chenli.

Dr. Tang Leihan, Dr. Wang Xiaowo, Dr. Dai Junbiao, Dr. Xiang Hua, Dr. Feng Yan, Dr. Li Chun, Dr. Wei Ping, Dr. Wang Qinhong, Dr. Li Yin, Dr. Qin Jianhua, Dr. Chen Peng, Dr. Li Feng, Dr. Si Tong, Dr. Zhou Zhihua, Dr. Lin Zhanglin and other scholars invested significant efforts in the writing of various chapters of this book. They work closely with more than a hundred of scholars and business professionals to complete the main body of this roadmap. More than 30 consulting experts have provided important advice on this roadmap with their profound academic attainments and strategic thinking.

Dr. Gao Fu, Chairman of the Chinese Society of Biotechnology, and Dr. Zhang Hongxiang, Secretary General of the Chinese Society of Biotechnology, strongly supported this research. The working group consisted of personnel from Shenzhen University of Advanced Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Institute of Biophysics, Chinese Academy of Sciences, Shenzhen Institute of Synthetic Biology, and Shenzhen Synthetic Biology Association, whose high degree of responsibility and work efficiency are important guarantees for the completion of this roadmap.

Zhang Xian-En
May 2024


Biochemical engineering
Artificial synthetic cell
Biomedicine & health
Bio-agriculture & future food
Environ. biotechnol.
BT-IT
Gene technology
Genetic circuits
Protein design
Chassis cells
Cell free systems
Multicellular systems
Unnatural coding & hybrid biosystems
Biofoundry & database
Enabling technologies
Multiscale modelling
Theoretical Framework
Artificial intelligence
ELSI
Governance Strategy

Contents

1 Introduction

Synthetic biology, rooted in biosciences and converging with chemistry, physics, informatics, and engineering principles, aims to design and engineer natural or synthetic biological systems. It seeks to unravel the fundamental laws of life ("build to learn") and transform biological systems for engineering applications ("build to use"), earning its alternative designation as engineering biology. As a key to decoding life and a disruptive technology shaping the future, synthetic biology has unlocked the gateway for converting non-living matter into living systems. By enabling the rational design and editing of life forms, it drives iterative advancements in biotechnology, driving next-generation biomanufacturing and the future bioeconomy.

Over a century ago, French scholars proposed the concept of artificial cell synthesis. By the mid-20th century, researchers in the United States and China successively achieved the *in vitro* artificial synthesis of biomacromolecules such as DNA, RNA, and proteins. In the early 21st century, the integration of engineered bioparts into microbial chassis to construct genetic circuits marked a new era. Breakthroughs such as bistable genetic switches, gene oscillators demonstrated the feasibility of logical metabolic regulation and artificial redesign. Genome editing, gene module characterization and biological system modeling enriched the underlying technologies of synthetic biology. The artificial synthesis of microorganism genomes such as viral, bacterial, and yeast genomes revolutionized the large-scale engineering of life. Minimal genome development redefined our understanding of genomic function and chassis construction, while genetic code expansion and unnatural amino acid incorporation opened avenues for novel life forms and applications.

A series of breakthroughs in enabling technologies have provided brand-new means

for analyzing the laws of life and accelerated the engineering applications of synthetic biology. The AI-based protein structure prediction algorithm AlphaFold has provided a revolutionary technical approach for *de novo* protein design, demonstrating the immense potential of data-driven paradigms in life science research. Quantitative relationships between the topological structure and function of biological networks, analyzed using mathematical and physical models, has provided a theoretical framework for understanding and designing artificial genetic circuits. Synthesis of important plant-based drugs in yeast such as artemisinin precursors and opium demonstrates the immense potential for the efficient artificial synthesis of natural products. The emergence of technologies converting carbon dioxide into starch, glucose and lipids, etc., has opened new avenues for the resource utilization and high-value application of carbon dioxide. Large-scale synthesis of bio-based materials and raw materials demonstrates the immense potential of green biomanufacturing to replace traditional energy and chemical industries. DNA storage, nanobiodevices, synthetic biosensors, and other electronic life systems are gradually developing from concept to reality.

The rapid expansion of innovative applications of synthetic biology into fields such as medicine, industry, agriculture, energy, environment, materials, and information has accelerated the transformation and upgrading of traditional manufacturing industries. It has also expedited the process of enabling synthetic biology to empower the new quality productive forces in next generation of biomanufacturing. Synthetic biology will play a central role in the revitalization of the bioeconomy and provide a brand-new solution for promoting global sustainable development.

The enormous potential and broad prospects of synthetic biology have attracted extensive attention from countries around the world. The European Union, the United States, the United Kingdom, Canada, Australia, and other countries have successively released synthetic biology roadmaps and development plans. For example, *A Synthetic Biology Roadmap for the UK* released by the United Kingdom in 2012, *Semiconductor Synthetic Biology Roadmap* released by the United States in 2018, *Engineering Biology: A Research Roadmap for the Next-Generation Bioeconomy* released in 2019, *Microbiome Engineering: A Research Roadmap for the Next-Generation Bioeconomy* released in 2020, *Engineering Biology & Materials Science: A Research Roadmap for Interdisciplinary Innovation* released in 2021, and *Engineering Biology for Climate & Sustainability: A Research Roadmap for a Cleaner Future* released in 2022, *A National Synthetic Biology Roadmap: Identifying commercial and economic opportunities for Australia* released by

Australia in 2021, and *Engineering Biology: A platform technology to fuel multi-sector economic recovery and modernize biomanufacturing in Canada* released by Canada in 2020, etc.

China places great importance on the development of synthetic biology. China, the United States, and the United Kingdom jointly initiated the "Three Countries, Six Academies" synthetic biology series symposia in 2009-2011, which comprehensively discussed the positioning and development goals of synthetic biology from perspectives of scientific, visionary, technological, platform, and policy. From the preliminary planning of the 973 Program to the strengthened implementation of the 13th Five-Year Plan key projects, as well as the continued support under the 14th Five-Year Plan, the Ministry of Science and Technology, has significantly enhanced China's synthetic biology research capabilities. The National Development and Reform Commission, the Ministry of Education, the Chinese Academy of Sciences, and the National Natural Science Foundation of China have all focused on the layout of synthetic biology. Local governments such as Shenzhen, Tianjin, Shanghai, and other regions have actively promoted the development of synthetic biology through measures such as establishing specialized R&D programs, new R&D institutions, talent cultivation platforms, major infrastructure projects, and industrial development funds.

Over the past decade, China has not only laid solid foundations for the development of synthetic biology and established itself as a key contributor to its innovation and application of synthetic biology, but also achieved a series of significant outcomes in basic research, technological innovation and industrial application. Among them, four research achievements, including the ab initio design and chemical synthesis of eukaryotic brewer's yeast chromosomes, the artificial creation of '16-in-1' yeast chromosomes, *de novo in vitro* initio synthesis of carbon dioxide to starch, and the intracellular synthesis of carbon dioxide to glucose and fatty acids, were selected as the top 10 news stories on scientific and technological progress in China.

To promote the rapid and healthy development of synthetic biology, the Chinese Society of Biotechnology took the lead, with the Synthetic Biology Branch organizing and implementing the initiative, gathering Chinese experts in the field of synthetic biology to carry out the research on the development strategy of synthetic biology for 2030, resulting in the *Synthetic Biology Roadmap 2030*. This book aims to sort out the theoretical framework, improve the discipline system of synthetic biology, carry out technology prediction, provide references for synthetic biology innovation and

application, and enhance public participation and cognition to support the innovative development of synthetic biology.

This book comprises four sections: theoretical framework, enabling technologies, application outlook, and capacity building, under the "3+1" framework.

The theoretical framework consists of two main parts, the first is "multi-scale theoretical framework of synthetic biology", the second is "synthetic biology and artificial intelligence." Enabling technologies focuses on 12 technical directions, namely: DNA sequencing, synthesis and assembly, gene editing, protein design, genetic circuits, chassis cells, cell-free systems, artificial multicellular systems, organoid engineering, unnatural amino acids encoding and synthetic biosystems, synthetic hybrid biotic-abiotic systems, biofoundries, and biopart resource and information platforms. Application outlook encompass two major areas, one is "build to learn", i.e. single-cell *de novo* synthesis, the other is "build to use", i.e. the application of synthetic biology in fields such as industry, medicine, agriculture, food, environment, and information technology.

In addition, the development of synthetic biology requires simultaneous research on synthetic biology policies and ethical regulations. Governance strategy proposes to establish a set of ethical governance system in line with the law and stage of scientific and technological development, improve the mechanism of scientific and technological ethical governance with the participation of universities, scientific research institutes, enterprises, societies, associations, consortiums, as well as scientific researchers and the public, so as to promote the coordinated development of scientific and technological activities and scientific and technological ethics. At the same time, we should also actively carry out scientific and technological exchanges and cooperation to jointly address scientific and technological challenges and promote the high-quality development of synthetic biology and industry.

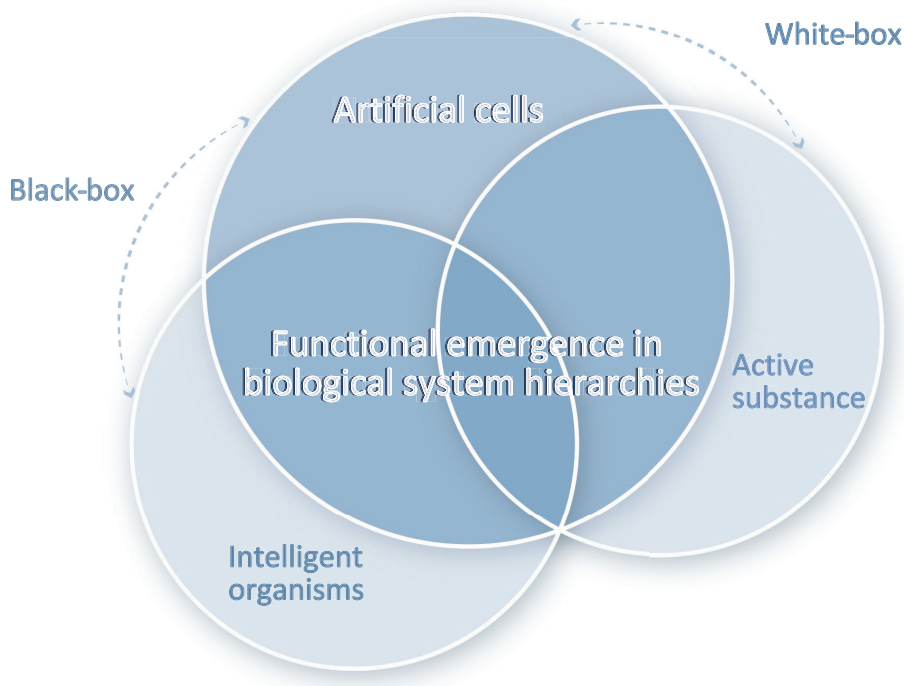
2 Theoretical Framework

Synthetic biology is an emerging interdisciplinary field initially driven by engineering technologies such as genetic circuit design and DNA synthesis. However, its theoretical foundation remains underdeveloped, and existing roadmaps lack a theoretical framework. As the scale of synthetic bioparts and their applications expands, their internal complexity and interactions with environments and hosts grow exponentially, posing significant challenges to predictable and rational design. Establishing a robust theoretical framework will accelerate the convergence of synthetic biology with traditional mathematical and physical sciences, breaking through the bottlenecks in creating more efficient, controllable, and robust intelligent biological systems. This, in turn, will lay the foundation for industrial-scale production.

To tackle these challenges, two primary methodologies have emerged for exploring biological functional modules (both natural and synthetic) and systematizing design principles. The first is the traditional quantitative biology approach, which employs quantitative characterization and mathematical modeling to construct knowledge-driven "white-box" models. These models excel at incrementally increasing complexity while establishing standardization. The second approach leverages artificial intelligence, using big biological data and machine learning to infer and construct data-driven "black-box models", which directly extract structures and correlations from vast datasets to inform component design. Together, these methodologies offer complementary pathways to address the growing demands of synthetic biology. To promote the deep integration of basic and applied research in the next phase, China hosted the "Quantitative Synthetic Biology" Xiangshan Conference in 2021, fostering key insights and consensus on the future direction of the field, which will be further organized and presented in this book.

The theoretical framework of this book consists of two parts. The first part focuses on theoretical prospects and planning based on “white-box” models, with the aims of establishing corresponding models and theories for life systems with specific functions and achieving the engineering goal of rational design and precise synthesis of biological systems from simple to complex. The second part focuses on synthetic biology research driven by artificial intelligence theory and technology, leveraging the integration of basic and applied research with knowledge and data to expand the function and application scenarios of synthetic biological modules.

Multi-scale Modeling of Synthetic Biology



Authors Tang Lei-Han, Liu Chen-Li, Fu Xiong-Fei, Li Xue-Fei

2.1 Multi-scale Modeling of Synthetic Biology

2.1.1 Abstract

Over billions of years of evolution, biological systems have developed an extraordinarily rich repertoire of functional modules. These modules interact in an ordered manner to sustain life's operations, reproduction, and adaptation to dynamic ecological environments. A core mission of synthetic biology is to systematically reorganize, modify, or redesign existing modules to enhance or expand specific functionalities, thereby advancing applications in industrial biosynthesis, healthcare and beyond. The fundamental units of these modules are proteins with specific binding capabilities. From proteins to molecular signaling pathways, metabolic networks, cytoskeletal structures, molecular machines, and membrane-bound or membraneless organelles, these components form the hierarchical hardware of life. Meanwhile, the activation and interaction of specific modules constitute the software that governs biological systems.

Theoretical and computational research should align with the near-term and mid-term developmental needs of synthetic biology. This involves focusing on the structural properties, self-organization, operation, and evolution of bioparts and processes, elucidating the interplay between material dynamics and information programming. By integrating insights from "black-box" models, researchers aim to refine design principles for functional components and progressively enhance the intelligence metrics of artificial elements and genetic circuits. Key research areas include protein functional design, regulation of signaling pathways, metabolic network engineering and homeostasis, membraneless organelle formation and design, and large-scale biological processes (e.g., microbial DNA replication and cell division).

2.1.2 Technical Overview

2.1.2.1 Organizational Principles and Evolutionary Theory of Biomacromolecular Assembly

Biomacromolecules are the foundation and key of life activities. The "sequence → structure → function" correspondence of proteins is the core issue in structural biology [1]. Conversely, the "function → structure → sequence" paradigm is the core issue in protein

design, which can be used to rationally modify existing proteins or guide the design of proteins from entirely new sequences. It serves as the theoretical foundation of synthetic biology and holds significant application potential [2]. In recent years, with the advancement of artificial intelligence technologies such as deep learning, methods such as AlphaFold can accurately predict protein structures based on sequences [3], and large-scale protein language models like Evolutionary Scale Modeling (ESM) can assist in predicting function. These developments indicate significant breakthroughs in the direction of "sequence → structure → function". Meanwhile, deep learning models like RFdiffusion and ProteinMPNN have demonstrated preliminary success in structure-based protein design, offering new avenues for "function → structure → sequence" reverse engineering. However, the vast majority of protein structure prediction models and protein sequence design models assume that protein structure remains static when the sequence is fixed. In reality, protein structure is dynamic. To further understand the relationship between protein sequence, structure, and function, and to improve the accuracy of protein function prediction and design, it is essential to explore the kinetic information and develop theoretical models that capture protein dynamics along the evolutionary trajectories. In nature, phenotype encoded by the genotype emerge through evolutionary processes. Because biological functions often serve as key constraints in the process of natural selection, evolutionary studies can provide valuable insights for constructing synthetic biology "white-box" models. Such models can guide the rational design of genotypes to achieve desired phenotypes [4].

2.1.2.2 Minimal Component Theory for Cellular Proliferation and Functional Synergy

Coordinating growth, replication, and division in unicellular organisms remains a complex theoretical challenge. Currently, computational models mainly rely on the operation modes of microorganisms under different physiological conditions. The corresponding theoretical models of functional coordination are typically constructed on optimization principles and demand extensive support from signal transduction pathways and genetic circuits [5]. Yet, many underlying molecular mechanisms remain only partially understood. The development of synthetic biology has provided the possibility of constructing genetic circuits with functional coordination that can achieve growth, replication, and division. This is a highly challenging but significantly meaningful research direction. By quantitatively characterizing and extracting the regulatory network

structures and principles underlying natural functional coordination, and by reconstructing artificial genetic circuits that embody these logical principles, synthetic biology can be substantially advanced. Moreover, it will lay a solid foundation for ultimately achieving the goals of creating synthetic life and designing life systems.

2.1.2.3 Foundational Theory of Cellular Function in Noisy Environments

The fundamental theoretical study of how cells execute life functions in noisy environments remains in its early stages. A central theme is the trade-off between accuracy and speed in molecular processes, alongside considerations of resource consumption and other performance metrics. On the one hand, the operation of complex genetic circuits is affected by both intracellular and extracellular noise, and a body of cybernetic studies has already been established in this area [6]. However, there are relatively few theoretical studies on information transfer based on statistical physics and stochastic thermodynamics, and the relevant theoretical studies are of great significance in revealing the efficiency, precision, stability, energy consumption and physical limits of information transfer of molecular circuits in physiological environments. On the other hand, the transport of molecular machines (such as proteins and complexes) and changes in their spatial distribution can affect the biochemical reactions within cells. These processes can consume energy and therefore operate under non-equilibrium conditions. Looking ahead, from the recent through 2030, significant progress is expected in relevant theoretical research informed by statistical physics [7, 8]. However, there remains a notable gap in theoretical studies that explicitly account for fluctuations during spatial transfer processes, which are likely to play a critical role in cellular function under noisy environments.

2.1.2.4 Self-Organization Emergence and Phase Transition Theory in Living Systems

There are abundant self-organization phenomena in living systems, such as clustered isotropic motions formed by flocks of birds and schools of fish through local self-organization. There have been theoretical and computational studies for such phenomena, but many other forms of self-organization and phase transition phenomena in non-equilibrium biological systems require new theories for explanation and development. For example, how is the widespread process of liquid-liquid phase separation within cells driven by protein-protein interactions and the interplay between proteins and their

intracellular environment? How are self-organization processes during biological development enabled by cell-cell interactions and communications [9]? The corresponding theoretical studies can provide theoretical guidance for quantitatively modifying and constructing artificial life systems.

2.1.2.5 Mechanisms and Characteristics of Intelligent Module Formation Under Selective Pressure

The stability and complexity of ecological environments remain both a central focus and a major challenge in ecological theory research [10]. Current research paradigms focus on observing changes in the composition of organisms within ecosystems, but methods for extracting systematic laws are still underdeveloped, and the quantitative, systematic study of experimental ecological systems requires further advancement. At the level of a single organism, adaptation to ecological environments can occur through the reshaping of internal gene regulatory networks, often driven by one or a series of specific genetic mutations. However, the stochastic and pervasive principles of these mutation processes remain to be revealed. Advancing theoretical studies in this area will be critical for uncovering the universal principles of adaptation, thereby deepening our understanding of microecological environments associated with disease. Such insights are also highly significant for synthetic biology, where they can inform strategies to modify and reconfigure microecological systems for biomedical and biotechnological applications.

2.1.2.6 Comparative Theoretical Study of Model Organisms in Fixed Environments and Intelligent Organisms in Changing Environments

Low-level organisms adapt to environmental changes via specific regulatory networks, and adaptive response principles (e.g., universality, specificity, category, critical logic, and robustness) remain central topics of investigation. There have been simulation studies of intelligent organisms based on deep learning, and it was found that these organisms can produce rule-specific behavioral changes under specific rule constraints to improve adaptability [11]. However, the interpretability of deep learning remains a significant challenge. Conducting comparative theoretical studies can help to understand the logic of learning and adaptation in real life organisms and improve the understanding of adaptive response principles.

2.1.3 Roadmaps

Current Status		
Predictive algorithms based on deep learning such as AlphaFold2 and MXfold2 for sequence-to-structure and biomacromolecular interactions are developing rapidly, but function-to-sequence theory and algorithm still need to be developed.		
Objective: To Develop the "White-box" Theory of Function→Structure→Sequence by Studying the Evolutionary Path of Molecular Sequences and the Synergistic Correlation of the Internal Structure of Molecules		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop the "white-box" theory of function→structure→sequence by investigating the evolutionary path of molecular sequences.	**Preliminarily formulate the "white-box" theory for single-molecule function of biomacromolecules.** • Explore the similarities and differences in large-scale conformational changes, catalysis and allosteric effects of biomacromolecules with the same class of functions and the evolution of their sequences.	**Extend and develop the "white-box" theory for multiple biomacromolecular interactions.** • In-depth investigate the functional structure and sequence basis of biomacromolecular interactions. • Study evolutionary mechanisms of intrinsically disordered protein aggregation.

Figure 1 Evolutionary theory roadmap for biomacromolecular assembly

Current Status		
Simulation algorithms for the top-down constructed minimal genome cells have been implemented, but theoretical framework that elucidate the coordination principles among different functional systems remain insufficient.		
Objective: To Propose a Theory of the Minimal Components and Minimal Interactions Required for the Proliferative Function of a Single Cell		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Iterative screen multifunctional synergistic regulatory network structures in single cells.	**Establish an iterative screening theory for the minimal components required for single-cell proliferation.** • Establish a theoretical framework for iterative screening of essential biological functions and combinations.	**Establish an iterative screening theory to realise the minimal regulatory network for single-cell proliferation.** • Develop a theoretical framework to iteratively screen network structures and explore robustness.

Figure 2 The theoretical roadmap of the minimal components and their interactions required for cell proliferation

Current Status		
A theoretical framework for non-equilibrium statistical physics to study the mutual constraints between free energy dissipation and functional robustness of interactions networks has been proposed. However, a unified theoretical framework for studying transient processes remains to be developed.		
Objective: To Advance the Foundational Theory of Cellular Processes by Extending Beyond Traditional Cybernetics and Establishing a Framework Grounded in Non-equilibrium Statistical Physics, Which Will Address the Dynamcis of Information Transfer, Molecular Machinery, Biosynthesis and Intracellular Transport Within Noisy Cellular Environments		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop statistical physics-based theories of information transfer.	**Establish a theoretical framework for information transfer in living systems.** • Clarify the physical limits of different genetic circuits in terms of information transmission.	**Establish a theoretical research framework for non-equilibrium statistical physics to study dynamic processes within cells.** • Clarify the quantitative effects of molecular machinery, biosynthesis and transport processes on cellular function and information transfer at the spatial and temporal scales of the cell.

Figure 3 Roadmap for the fundamental theory of life functions in cells under noisy environments

Current Status		
Recent progress has emphasized theoretical advances grounded in computer simulations. Frameworks describing the emergence of order in collective behavior have been established. However, a fundamental breakthrough in the underlying theory of self-organization in living systems remains outstanding, and exploration of universality classes is still limited.		
Objective: Establish a Phase Transition Theory Based on Non-equilibrium Statistical Physics for the Self-organization of Living Systems to Form Spatiotemporally Ordered Functional Structures		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop a theory of self-organized phase transitions based on new non-equilibrium systems of life.	**Establish new theory of statistical physics in specific biological phenomena.** • Establish a theoretical framework based on non-equilibrium statistical physics for the phenomenon of liquid-liquid phase separation and the corresponding condensate spacial structure in cells.	**Establish a unified theoretical framework for the phenomenon of self-organization in living systems.** • Establish a unified framework for the theoretical study of self-organized phase transitions and the definition of universality class for living systems with the phenomenon of ordered behavioural emergence.

Figure 4 Roadmap for self-organization and phase transition theory

Current Status		
There are computational studies of mutations or resistance mechanisms, however, the rationale behind the mechanisms is unclear and not systematically categorised.		
Objective: To Analyze the Mechanisms and Principles for the Emergence of Biologically Intelligent Modules Under Selective Pressures Such as Environmental Adaptation (Including Changing Environments) and Resource Competition and Identify the Universality Classes of These Mechanisms		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Reveal the mechanisms and universal principles of the formation process of the stability and complexity of biologically intelligent modules.	**Establish a theoretical framework for studying the mechanisms and principles of intelligent module formation.** • Study the process by which organisms adjust their gene regulatory networks through mutations to understand the principles of logic, stochasticity and universality of the process.	**Establish a theoretical framework for studying the construction of stable ecological environments.** • Uncover mechanisms and universal principles of system stability and complexity formation processes related to information flows and coping strategies at ecological scales.

Figure 5 Roadmap for intelligent module formation mechanism and characterization under selective pressure

Current Status		
Computational simulation approaches have been proposed to investigate the evolution of behaviour in intelligent organisms. Nevertheless, a comprehensive theoretical framework for the emergence of behavior remains undeveloped.		
Objective: Establish the Fundamental Principles Underlying the Emergence of Life function by Systematically Examing Both the Differences and Interconnections Between Living model Organisms and Intelligent Organisms Endowed with Learning Functions		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Explore the realisability of the principles of function emergence in synthetic lifeforms through the study of intelligent organisms with learning capabilities.	**Unravel the mechanisms of function emergence in intelligent organisms with learning capabilites.**	**Clarify the differences and connections between living model organisms and intelligent organisms with learning capabilites and establish a theory of optimal design of synthetic life-form functions.**

Figure 6 Roadmap for the comparative theory of living model organisms in fixed environments and intelligent organisms in changing environments

2.1.4 Technical Pathways

2.1.4.1 Evolutionary Theory of Biomacromolecular Assembly

Current Technologies: Mature computational methods, including deep learning-based sequence alignment algorithms and molecular dynamics simulations.

Objectives and Breakthroughs: Develop a "white-box" theory linking function → structure → sequence through studying evolutionary paths and physical properties of molecular sequences, from the perspective of intelligent soft matter.

Challenges: AI models lack sensitivity to the global effects of single-site changes in structure prediction; integration of heterogeneous biological data is inefficient; computational demands for biomacromolecular interactions exceed current capabilities; existing deep learning-based algorithms are insufficiently developed and suffer from poor interpretability.

Expected Progress Recently: Establish preliminary "white-box" models for biomacromolecular function.

Expected Progress by 2030: Extend and refine the "white-box" models for biomacromolecular interactions.

Potential Solutions

Use AI algorithms such as AlphaFold2 and MXfold2 to preliminarily predict the structures of biomacromolecules such as proteins and nucleic acids. By combining these with short-time molecular dynamics-simulated structural models, analyze the cross-species evolution of specific functional proteins, nucleic acids, and polysaccharides. Study the functions and evolution of these biomacromolecules in the context of large-scale conformational changes, catalysis, and aliasing, and explore the structural and sequence bases related to specific functions. From the function-structure-sequence perspective, propose foundational theories of function prediction for proteins, nucleic acids, and polysaccharides. Moreover, conduct in-depth research on the structural and sequence bases of complex biological functions such as biomacromolecular interactions and explore the evolutionary mechanisms underlying functions such as the aggregation of intrinsically disordered proteins.

2.1.4.2 Theoretical Framework for the Minimal Components and Their Interactions Required for Cell Proliferation

Current Technologies: Existing theoretical studies primarily focus on computational approaches. For instance, flux balance analysis is used to study cellular metabolic networks, while single-cell and multi-scale simulation techniques are applied in minimal genome research to explore the kinetic and homeostasis of intracellular biochemical reactions. However, there remains a lack of summaries of the underlying principles, and theoretical studies addressing the regulation and coordination among functional modules within the cell are still limited.

Objectives and Breakthroughs: To establish a theory of minimal components and their interactions required for single cell proliferation, and to iteratively identify and refine the regulatory network structures that enable multifunctional coordination within cells.

Challenges: The existence of numerous potential evolutionary trajectories, the vast search space of possible network configurations, and the uncertainty of universal laws governing cellular proliferation.

Expected Progress Recently: Develop theoretical frameworks for information transfer across time scales.

Expected Progress by 2030: Establish a complete theoretical framework for analysing and categorising the phenomenon of single-cell proliferation.

Potential Solutions

Based on the basic function of cell proliferation and known proliferation patterns, minimal component combinations can be identified through mathematical-physical modeling, with the corresponding combination laws abstracted and formalized. Subsequently, nonlinear dynamics analyses can be performed to summarize and classify the core structural features in a coarse-grained manner. From an evolutionary perspective, incorporating selection pressures will enable comparative analyses of the connections and differences among diverse network structures along the evolutionary pathways.

2.1.4.3 Foundational Theory of Cellular Function in Noisy Environments

Current Technologies: Significant progress has been made in theories and methods grounded in statistical physics. Notably, advances include the theory of energy and information transfer in biological oscillatory systems, as well as the quantitative relationship between energy dissipation rate and information flow in non-equilibrium steady states. Nevertheless, a unified theoretical framework that spans both spatial and temporal scales has yet to be established.

Objectives and Breakthroughs: In the context of noisy intracellular environments, the goal is to strengthen foundational theories of information transfer, molecular machinery, biosynthesis and transport. This involves: advancing cybernetics-based approaches, establishing a theoretical foundation grounded in non-equilibrium statistical physics and non-linear dynamics, and developing a comprehensive theory of information transfer informed by statistical physics.

Challenges: Theoretical and technical capabilities in non-equilibrium statistical physics are limited, as the process involves multiple temporal and spatial scales, making cross-scale theoretical research challenging.

Expected Progress Recently: Develop time-resolved theories of information transfer in living systems.

Expected Progress by 2030: Establish a theoretical framework for non-equilibrium statistical physics encompassing molecular machinery, biosynthesis, intracellular structural reorganization and transport processes on cellular spatiotemporal scales.

Potential Solutions

The approach is to advance non-equilibrium statistical physics and non-linear dynamics for specific biological systems, such as adaptive regulatory networks, by integrating synthetic biology construction and quantitative biology testing. For processes such as cell polarization, fine-grained short-time simulations and theoretical studies of molecular interactions will be used to extract key spatial and temporal constants. These constants will inform analyses of principles and mechanisms underlying the coupling between biomacromolecular interaction modes and cellular-scale biological functions. Furthermore, equilibrium theories and physical limits governing the relationship between

energy dissipation and function accuracy will be examined. Ultimately, this knowledge will guide the design of optimized dynamic processes.

2.1.4.4 Theory of Self-Organization and Phase Transitions

Current Technologies: For self-organized collective behaviors such as fish schools and bird flocks, there are existing phase transition theories based on non-equilibrium statistical physics. In addition, theoretical and computational frameworks exist for specific functional processes, including developmental dynamics. However, these approaches largely focus on individual phenomena within particular systems and lack the universality required to capture the broader principles of self-organization across living systems.

Objectives and Breakthroughs: To establish a phase transition theory based on non-equilibrium statistical physics that explains how living systems self-organize into spatiotemporally ordered functional structures. Building on this foundation, the goal is to develop a theory of self-organized phase transition programming in non-equilibrium biological systems.

Challenges: The mechanisms underlying liquid-liquid phase separation remain poorly understood, particularly due to the involvement of cross-scale interactions. The process spans multiple temporal and spatial scales, posing significant challenges for theoretical integration. Furthermore, identifying and characterizing the number and variability of universality classes in biological systems is inherently complex.

Expected Progress Recently: Establish a theoretical research framework, based on non-equilibrium statistical physics to describe liquid-liquid phase separation phenomenon in cells.

Expected Progress by 2030: Establish a unified theoretical framework that defines universality classes for functional units formed through self-organization in living systems, ranging from biomacromolecular assemblies to higher-order organizational structures.

Potential Solutions

For the liquid-liquid phase separation phenomenon, molecular dynamics simulations will be employed to quantitatively characterizing free energy input and dissipation,

thereby advancing phase transition theory within non-equilibrium statistical physics. Comparative analyses of self-organization across collective behaviors, developmental processes, and cellular differentiation will be conducted to identify commonalites and distinctions. Large-scale computational simulations will be used to extract key spatial and temporal constants, enabling the formulation of phase transition theories across multiple scales. Finally, by investigating the coupling of information and energy coupling across hierarchical levels, a unified cross-scale theoretical framework will be established. This framework will support the regulation and programmable control of self-organized behaviors in living systems.

2.1.4.5 Formation Mechanisms and Characteristics of Biological Intelligent Modules Under Selective Pressure

Current Technologies: Several phenomena illustrate the adaptive formation of biological intelligent modules under selective pressure. For instance, yeast can generate new interaction networks from existing protein complexes through mutation; cancer cells can remodel their regulatory networks to evade drugs and immune responses; and gut microorganisms, along with ecological species, establish stable coexistence systems by altering their interaction modes in response to environmental selection. Despite these observations, comprehensive theoretical studies remain limited.

Objectives and Breakthroughs: To elucidate the mechanisms by which biological intelligent modules emerge under selective pressures such as environmental adaptation. Key objectives include: analyzing the universality of emergence mechanisms and principles; developing a theoretical framework to capture the logic, stochasticity, and robustness of processes by which organisms reorganize internal gene regulatory networks through mutations and other adaptive strategies.

Challenges: The complexity of regulatory networks, characterized by numerous interconnected links, makes it difficult to pinpoint evolutionary sites that drive specific functional adaptations. The processes span diverse temporal and spatial dimensions, posing significant challenges for cross-scale theoretical integration.

Expected Progress Recently: Establish a theoretical framework describing yeast regulatory network evolution under selective pressure.

Expected Progress by 2030: For complex ecological systems such as tumour tissues, gut microbiota, and marine environments, establish a theoretical framework for studying the mechanisms and universality principles underlying the emergence of

stability and complexity across diverse environments.

Potential Solutions

For the yeast regulatory network, existing regulatory network information will be leveraged alongside functional requirements. Simulations involving random modification of network links, combined with a function-oriented steepest descent method, will be used to identify key sites for structural adjustment. This will enable the construction of a theoretical framework for regulatory network evolution under selective pressure, clarifying the logic, stochasticity, and robustness of the process. For tumor, intestinal, and marine microenvironments, synthetic biology and quantitative biology testing will be integrated with multi-scale model calculations. By summarizing the evolutionary laws of ecological environments, extracting critical interaction network structures, and identifying key spatiotemporal parameters, the aim is to distill the mechanisms and universality principles that govern the formation of functional stability and complexity in diverse biological systems.

2.1.4.6 Comparative Theory of Model Organisms in Fixed Environments and Intelligent Organisms in Changing Environments

Current Technologies: Recent computational advances include the development of algorithms that enable intelligent organisms with learning functions to undergo virtual evolution. These approaches represent promising steps toward modeling adaptive life processes. However, theoretical research in this domain remains underdeveloped, with limited frameworks to explain the underlying principles.

Objectives and Breakthroughs: The aims are to establish the foundational principles of life function emergence by analyzing both the differences and connections between model organisms and intelligent organisms with learning capabilities. Also, to explore the feasibility of applying these principles to synthetic lifeforms, particularly through the study of intelligent organisms with learning functions.

Challenges: Deep learning-derived laws often suffer from poor interpretability, limiting their explanatory power. Even for model organisms, the optimization theory of functional design remain incomplete.

Expected Progress Recently: Reveal the mechanisms by which intelligent

organisms with learning capabilities generate new biological functions.

Expected Progress by 2030: Clarify the distinctions and connections between model organisms and intelligent organisms with learning capabilities, and establish an optimal design theory for synthetic life-form functions.

Potential Solutions

Advancing statistical physics research integrated with deep-learning neural networks will be critical. This approach can uncover evolutionary laws governing neural network structures that determine the functions of intelligent organisms, while also elucidating the pathways and robustness of function formation. Because intelligent organisms with learning functions exhibit clearer components and interactions, comparative studies with standardized model organisms will help distill the fundamental principles of life function design. These principles can then be extended to synthetic life systems, providing a rational basis for the design and programming of novel biological functions.

2.1.5 Summary

The use of mathematics, physics, and related disciplines to model and computationally study complex life systems has a long history. However, due to the complexity and diversity of biological laws, much of the existing research focuses on specific subsystems, making it difficult to establish a systematic theory and limiting the scalability of current approaches. This part of the roadmap outlines the envisioned future directions, pathways, and key technologies for the development of functionalized complex systems from now to 2030. It emphasizes mechanisms underlying the emergence and evolution of life functions across multiple spatio-temporal scales, including biomacromolecules, single cells, and bioclusters. The ultimate goal is to advance the construction of foundational theories of life systems and to overcome existing data and methodological bottlenecks. The development of such theories is expected to provide robust theoretical support for the rational design of synthetic biology, thereby enabling more systematic and predictive approaches to engineering life.

References

[1] Bahar I, Lezon T R, Yang L W, et al. Global dynamics of proteins: bridging between structure and function. Annual Review of Biophysics, 2010, 39: 23.

[2] Pan X, Kortemme T. Recent advances in *de novo* protein design: principles, methods, and applications. Journal of Biological Chemistry, 2021, 296: 100558.

[3] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589.

[4] Tang Q Y, Kaneko K. Dynamics-evolution correspondence in protein structures. Physical Review Letters, 2021, 127(9): 098103.

[5] Zheng H, Bai Y, Jiang M, et al. General quantitative relations linking cell growth and the cell cycle in *Escherichia coli*. Nature Microbiology, 2020, 5(8):995-1001.

[6] Vecchio D D, Dy A J, Qian Y. Control theory meets synthetic biology. Journal of the Royal Society Interface, 2016, 13: 20160380

[7] Wang S W, Tang L H. Emergence of collective oscillations in adaptive cells. Nature Communications, 2019, 10: 5613.

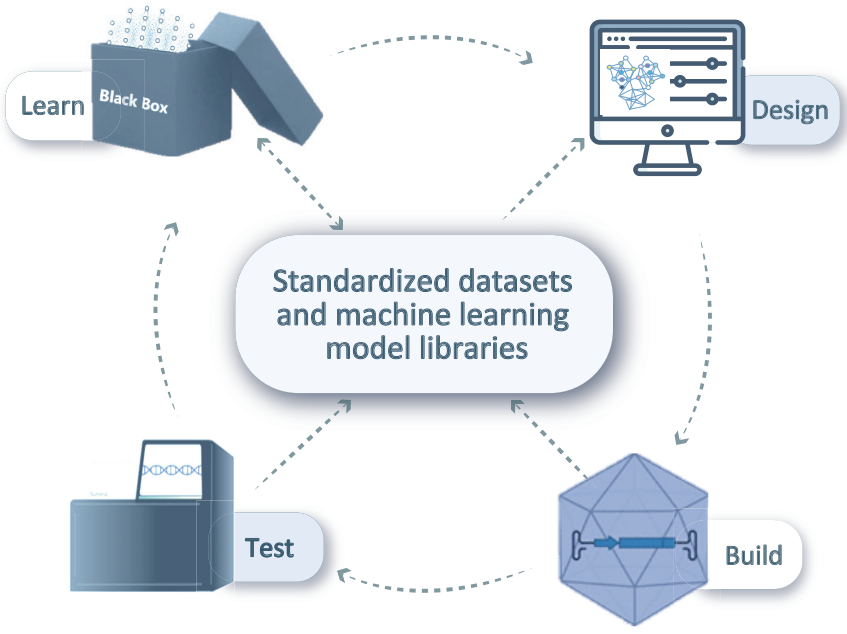
[8] Zhang D, Cao Y, Ouyang Q, et al. The energy cost and optimal design for synchronization of coupled molecular oscillators. Nature Physics, 2020, 16(1):95-100.

[9] Guan G, Wong M K, Zhao Z, et al. Volume segregation programming in a nematode's early embryogenesis. Physical Review E, 2021, 104: 054409.

[10] Landi P, Minoarivelo H O, Brg L H, et al. Complexity and stability of ecological networks: a review of the theory. Population Ecology, 2018, 60: 319-345.

[11] Gupta A, Savarese S, Ganguli S, et al. Embodied intelligence via learning and evolution. Nature Communications, 2021, 12: 5721.

Artificial Intelligence for Synthetic Biology



Authors Wang Xiao-Wo, Wei Zheng, Liu Hai-Yan

2.2 Artificial Intelligence for Synthetic Biology

2.2.1 Abstract

The construction of synthetic biological systems relies on profound understanding and precise modeling of living systems. Artificial intelligence (AI) technologies can effectively learn and model complex biological laws, predict the functionality of synthetic biological systems, and guide the design of artificial bioparts. This enables direct human comprehension and utilization of intricate biological mechanisms. Against the backdrop of rapidly accumulating high-throughput biological data, targeted research on AI theories and methods tailored to the knowledge and data characteristics of synthetic biology holds promise for overcoming key technical bottlenecks in the "design-build-test-learn" closed-loop development framework. Such advancements could comprehensively accelerate development processes while achieving cost reduction and efficiency improvement.

2.2.2 Technical Overview

2.2.2.1 Artificial Intelligence Technology

Artificial intelligence refers to machine-demonstrated intelligence, with its conceptual foundation established in the 1950s. Although numerous AI algorithms were subsequently developed, limitations in hardware performance and data availability constrained their practical application. Around 2010, breakthroughs emerged through Graphics Processing Unit (GPU) accelerated computing and the accumulation of massive datasets. Machine learning algorithms, particularly deep learning, achieved revolutionary progress, surpassing human-level recognition accuracy in image processing and enabling commercial applications across facial recognition, machine translation, autonomous driving, and other domains. Prediction represents a fundamental AI task, utilizing sample features to forecast specific attributes such as image classification, text translation, and protein structure prediction. Generation constitutes another core AI capability, creating novel samples with desired characteristics by leveraging probability distributions of target datasets, exemplified by image synthesis and enhancement, text generation, and DNA/RNA/amino acid sequence design. While application maturity varies across

domains and tasks, AI's scope continues to expand alongside sustained improvements in model performance.

2.2.2.2 AI Theories and Methods for Synthetic Biology

Although AI technologies share fundamental algorithmic frameworks across disciplines, their theoretical and methodological implementations differ significantly due to domain-specific knowledge and data characteristics [1]. First, AI typically requires large-volume, high-quality, and diverse training data-conditions rarely met by biological datasets. Second, biological data faces curse of dimensionality, where feature counts vastly exceed sample quantities compared to image-based applications. Third, the complexity of biological systems renders conventional AI performance metrics inadequate for evaluating true biological pattern recognition capabilities, complicating model training and deployment [1, 2]. Recent breakthroughs exemplified by AlphaFold [3] in protein structure prediction demonstrate that tailored AI innovations can overcome these challenges, which benefits from innovations in related artificial intelligence theories and methods, thereby developing neural networks specifically designed for amino acid sequences and protein conformations. This underscores the necessity for dedicated research into theories and methods aligned with synthetic biology's unique knowledge architecture and data properties to meet practical application demands [1, 2, 4-8].

2.2.3 Roadmaps

Current Status		
The field lacks theoretical modeling frameworks tailored to the intrinsic characteristics of synthetic biology research objects. Most synthetic biology problems still suffer from the absence of standardized datasets and effective model evaluation methodologies.		
Objective 1: Establish Standardized Benchmark Datasets for Biopart/Module-Function Relationships		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Refine the fundamental common issue in the domain, develop data standards for each challenge and create standardized biopart/ module-function datasets.	**Establish standardized datasets on 5 to 10 fundamental common issues.** • Master or develop high-throughput experimental techniques. • Create public data integration methodologies.	**Establish standardized datasets on 10 to 20 fundamental common issues.** • Reduce costs of high-throughput experimentation. • Generate experimental data specifically optimized for AI model training.
Objective 2: Develop AI Model Evaluation Framework for Biopart/Module Design		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance existing AI evaluation frameworks by incorporating biological knowledge and data and improve model interpretability and transparency to verify reliability.	**Establish standardized baseline evaluation metrics.** • Define application boundaries for evaluation metrics. • Develop model interpretation methods for validation.	**Build comprehensive evaluation framework.** • Develop metrics accounting for real-world impacts across diverse samples. • Integrate model mechanisms with biological knowledge in evaluation frameworks.

Objective 3: Construct Machine Learning Model Zoo for Biopart/Module Design		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop novel AI methods and model architectures addressing critical common issues to create specialized machine learning model zoo for synthetic biology design.	**Construct machine learning models on 5 to 10 fundamental common issues.** • Form interdisciplinary AI-synthetic biology teams for collaborative model development. • Promote cutting-edge cross-disciplinary research initiatives.	**Construct machine learning models on 10 to 20 fundamental common issues.** • Enable AI models to effectively integrate domain-specific biological knowledge. • Implement design space reduction strategies.

Figure 1 Roadmap for constructing knowledge-data co-driven machine learning model zoo and corresponding standardized datasets

Current Status		
The field lacks effective explainable artificial intelligence (XAI) methodologies, leaving knowledge from published biological literature and insights derived from AI models underutilized. Furthermore, existing knowledge and data-driven insights remain insufficiently integrated into biological system simulation and design workflows.		
Objective 1: Develop AI Technologies for Automated Biological Knowledge Mining from Literature		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop natural language processing machine learning models adapted to biological literature, make breakthroughs in automated construction of knowledge graphs for bioparts, and construct knowledge bases.	**Establish biological knowledge graphs and associated databases in major synthetic biology subfields, supported by domain-adapted models and software for automated knowledge extraction.** • Systematically and standardardly define biological functions and knowledge. • Innovatively design natural language processing (NLP) methods tailored to biological literature characteristics for automated knowledge graph extraction.	**Develop AI methods, achieve systematic, automated, and intelligent knowledge graph construction with summarization and classification accuracy comparable to manual annotation.** • Categorize functional modules in natural and synthetic biological systems. • Develop AI-powered text-cleaning tools with enhanced precision. • Optimize AI models to align extracted knowledge with existing biological knowledge frameworks.

Objective 2: Develop Interpretable AI Methods for Biological Knowledge Extraction		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop interpretable AI techniques to decode how models represent biological principles, enabling "white-box" model understanding.	**Develop interpretation methods for common machine learning models to elucidate operational mechanisms and learned knowledge of the model.** • Conduct interpretations according to the characterization of neural network and specific biological questions.	**Basically realize "white-box" model explanations and achieve automated extraction of critical biological patterns.** • Enhance model performance metrics. • Integrate cutting-edge model mechanism insights to accurately identify the patterns recognized by the model and automatically extract complex biological laws.
Objective 3: Create Knowledge-Data Co-Driven Biosystem Simulation Technologies and Software Platforms		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Establish a knowledge-data-driven biological system simulation framework based on artificial intelligence technology, effectively reduce simulation complexity and implement accurate simulation of cross-scale and large-scale systems.	**Initially establish biological system simulation frameworks co-driven by knowledge and data to simulate partial system functions.** • Implement hybrid digital twins of cellular functions combining knowledge-based mathematical models and data-driven AI. • Leverage deep learning to enhance simulation accuracy and speed of molecular interactions.	**Build modular whole-cell simulation platforms and systematic and standardized coarse-grained biomacromolecular simulation workflows.** • Fuse mechanistic models with AI architectures to refine the accuracy of cellular digital twins. • Design novel neural network architectures based on first principles for physicochemical and dynamic property simulation.

Objective 4: Construct Knowledge-Data Synergistically Driven Feature Representations and Applications of Biologically Complex Rules		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop representation learning techniques for synthetic biology data and knowledge, effectively integrate of knowledge into artificial intelligence design models, and enhance intelligent assisted design capability for complex synthetic biological systems.	**Obtain effective feature representations of synthetic biological data, enabling the utilization of the inherent patterns in the data for intelligent assisted design.** • Train AI models to capture high-dimensional feature representations of biological knowledge. • Combine mathematical models with AI for dimensionality reduction in spatiotemopral complexity.	**Develop more universal techniques for the dimensionality reduction of biological complex systems, develop targeted and interpretable AI technologies, and effectively utilize the rules contained in the biological knowledge base and data for intelligent assistance in design.** • Implement a universal dimensionality reduction algorithm using mathematical models and AI models. • Embed physicochemical/biological knowledge into models through interpretability frameworks.

Figure 2 Roadmap for developing knowledge-data synergy-driven reverse design technologies for synthetic biological systems

<table>
<tr><th colspan="3">Current Status</th></tr>
<tr><td colspan="3">While intelligent optimization of dry-lab/wet-lab closed-loop experimental iterations has been achieved in specific cases such as enzyme design, decision-making processes remain heavily reliant on empirical knowledge, limiting generalizability to other application scenarios.</td></tr>
<tr><th colspan="3">Objective 1: Establish AI-Driven “Design-Build-Test-Learn” Closed-Loop Pipelines</th></tr>
<tr><th>Expected Breakthroughs</th><th>Expected Progress Recently</th><th>Expected Progress by 2030</th></tr>
<tr><td>Develop machine learning-based adaptive experimental design methods to enable automated experiment planning and parameter optimization through active learning mechanisms.</td><td>Establish prototype adaptive experimental design platform for core synthetic biology challenges.
• Investigate strategies for incremental data utilization and model optimization.
• Develop machine learning-based adaptive experimental design frameworks.</td><td>Operationalize AI-driven dry-lab/wet-lab closed- loop platform as pipelines.
• Create active learning-enabled machine learning models for targeted applications.
• Enhance machine learning-driven adaptive experimental design.</td></tr>
</table>

Figure 3 Roadmap for AI-driven dry-lab/wet-lab closed-loop experimental systems

2.2.4 Technical Pathways

2.2.4.1 Build Knowledge-data Co-driven Machine Learning Model Zoo with Standardized Datasets

Current Technologies: Compared to domains such as computer vision and natural language processing, synthetic biology lacks robust artificial intelligence (AI) theories and methodologies tailored to biological knowledge and data. Many synthetic biology issues still lack unified formalized mathematical definitions and standardized datasets for machine learning (ML) model training, validation, and benchmarking. Critical relationships such as "sequence-structure-function" and "molecule-network- function" remain poorly characterized. As a result, most issues cannot be addressed in the same way as the "protein structure prediction" issues, where models and training methods like AlphaFold that are tailored to amino acid sequences and protein structures. While vast datasets exist for DNA, RNA, proteins, small-molecule drugs, and other biological entities, standardized benchmark datasets are scarce across most synthetic biology applications. Different research teams process and screen data based on their own objectives and experience, which increases the difficulty of applying AI technologies. Although AI methods have been applied to some synthetic biology issues, the absence of unified reliability assessment frameworks impedes comparative analysis of competing approaches. Many issues still lack highly effective methods, with model structures often borrowed from fields such as image processing and natural language processing, necessitating deeper understanding and innovation tailored to the characteristics of the research subject. Additionally, existing ML models face limitations in software compatibility, programming languages, and runtime environments, hindering modular reuse and scalability.

Objectives and Breakthroughs: Establish standardized benchmark datasets for biopart/module-function relationships. In-depth study the characteristics of knowledge and data within the field, abstract and refine fundamental common issues, establish standardized formal mathematical definitions, develop data standards for each fundamental issue, and create standardized bioparts/module functional dataset, accurately describe "sequences-structure-function" and "molecule-network-function" and other important relationships, covering the design and modification of gene elements, RNA

elements, protein components, chassis cell, etc. Data standards should be determined based on fundamental common issues, ensuring that these issues have relevance across multiple related specific applications and support the continuous expansion of datasets. This ensures that the artificial intelligence models trained on the standardized datasets can be easily transferred to specific issues.

Develop AI model evaluation systems for biopart/module design. Combine biological knowledge and relevant data characteristics to analyze discrepancies between model evaluation metrics and real-world biological performance. Identify weaknesses that hinder performance improvements and enhance the overall reliability of the evaluation system. Additionally, based on existing AI model evaluation systems, validate model reliability through "white-box" interpretation, and continuously optimize and improve the evaluation system based on biological experiment validation results.

Construct machine learning model zoo for biopart design. For fundamental common issues in synthetic biology, utilize existing biological knowledge and data to develop machine learning model structures, enabling the models to efficiently learn from involved biological laws, and accurately describe the "sequence-structure-function" and "molecule-network-function" relationships. Based on above, collect the most advanced published models, and reimplement them using two specified software frameworks to form a machine learning model library.

Challenges: Some fundamental common issues are constrained by experimental throughput and funding scale, making it challenging to generate a large volume of data in a single batch. Public data integration faces challenges from inconsistent experimental conditions, variable data quality, and missing preprocessing metadata. The cost of high-throughput experiments for some of the fundamental common issues is still too high, and the complexity and variety of background conditions in biological experiments make it difficult to obtain datasets that are compatible with AI methods, and AI models trained on these datasets are not representative and are only applicable to some specific conditions. Evaluation metrics tend to target only the majority of samples, and the large impact of some few misclassifications or small errors on research goals is difficult to measure.

Artificial intelligence models are often very complex and achieving a "white-box" interpretation of the model and representing it in an appropriate biological formalism are current challenges. Sequences designed by AI generative models are not as accessible to manual judgment of quality as images and text generated by AI models in other domains

and are difficult to assess computationally. Commonly used evaluation metrics in AI techniques (e.g., accuracy of classification models and R^2 of regression models) are often insufficient for the field of synthetic biology, where the diversity of data samples and the complexity of the problem affect the real-world performance of the model during the design process.

Artificial intelligence methods are less or less maturely applied to many synthetic biology problems, often due to a lack of referenceable benchmark models. Moreover, biological systems are too complex for AI methods to model them with a limited data set, and thus the performance is difficult to meet the application requirements.

Expected Progress Recently: Establish standard datasets for 5-10 fundamental common issues, establish standardized model evaluation metrics, and establish a machine learning model zoo.

Expected Progress by 2030: Establish standard datasets for 10-20 fundamental common issues, establish a model evaluation metric system, and establish a machine learning model zoo.

Potential Solutions

Master or develop high-throughput experimental techniques to generate a large diversity of data in a single run. Taking the problem of designing and modifying gene regulatory sequences in biopharmaceutical development and industrial chassis cell optimization as an example: for the design of promoter sequences, the gene expression levels of 100,000 synthetic sequences in cells can be evaluated using techniques similar to massively parallel reporter systems (MPRA); for enhancer design, techniques similar to STARR-seq can be used to evaluate the large number of synthetic enhancers' gene regulation performance. For the integration of public data, it is necessary to screen, clean and preprocess the raw data, develop data integration methods to normalize the data, eliminate the batch effect, and enable the combined use of data obtained from different laboratories. Taking the search for differentially expressed genes in human cells in the field of biomedicine as an example, single-cell RNA-seq data from different cell types in multiple laboratories are collected, processed using the same platform technologies and processes, and labeled with cell subpopulation types. Data involved in the structural design of proteins, such as enzymes and antibodies, should follow similar principles, and methods should be developed accordingly.

For key critical issues in the field, the cost of high-throughput experiments should be gradually reduced. If there are biotechnology bottlenecks, funds can be focused on generating sufficient datasets for a few model organisms or model cells. For example, in the U.S. ENCODE database, since the order of magnitude of the cost of each ChIP-seq data is about 10,000 CNY, the research target of one data corresponds to a cell type-antibody type combination, and the number of cell types multiplied by the number of antibody types can reach more than 1 million, making it difficult to achieve comprehensive data coverage. The specific strategy adopted is to cover as many antibody types as possible for a few commonly used cell line types, and to cover as many cell types as possible for a few common antibody types, so as to meet the basic training requirements of the model, ensure that the trained model can extract the association relationship between different dimensions of the same object, and then further migrate the model to other related application tasks. Similar principles can be followed in other fields, such as the design and modification of industrial chassis cells, the design of industrial enzymes, and the design of gene editing technologies.

At present, more interdisciplinary research teams in AI and synthetic biology should be established to support relevant cutting-edge topics, which will lead to the development of new model structures and training methods by AI experts with the assistance of synthetic biology experts, so as to train machine-learning models with excellent performance to meet the basic needs in key areas such as industrial chassis cell optimization, industrial enzyme design, and pharmaceutical development. Each fundamental common problem should have at least two basic models and be implemented under two specified machine learning frameworks.

The application scope of the existing evaluation metrics should be accurately limited, and a methodology system that can effectively utilize biological experimental data and knowledge databases to evaluate the models should be developed for synthetic biology tasks.

Adequately study the samples that have a significant impact on the results of the model design, analyze their causes, and integrate these factors into the evaluation of the objective function or AI model to ensure that the evaluation process presents the true impact of misclassification on the real world. For the “white-boxed” explanation of AI, it is necessary to compare the knowledge learned from the internal structure of the AI model with the biological knowledge related to the problem under study, based on the latest research on the mechanism of the AI model, to discover the biological rule

corresponding to the model structure and parameters. For generative models, some measure of the difference in distribution between the original and generated samples can be searched as a performance assessment indicator, and experiments can be used to verify their reliability.

Incorporate relevant biological knowledge into the AI model to effectively reduce the design space, so that the model can learn complex rules in a limited dataset to meet the design needs under specific conditions.

2.2.4.2 Develop knowledge-data Co-driven Inverse Design Technologies for Synthetic Biological Systems

Current Technologies: The discovery and application of biological knowledge rely on the understanding of complex biological system data, however, biological system data are characterized by multi-level, cross-scale, and high coupling, in which the biological knowledge embedded in them is difficult to be extracted directly. Knowledge that has already been mastered has not been summarized in a clear, complete, and systematic way, and thus is difficult to apply. From other fields, artificial intelligence is a powerful tool to address knowledge discovery and application, but due to the specificity of biological data, existing AI theories and methods are difficult to adapt to biological problems. Therefore, biological knowledge discovery and reverse design of synthetic biological systems based on AI mainly involve the following aspects: development of natural language processing methods to extract knowledge from the literature and construct a knowledge graph; development of interpretable AI models and methods that can effectively extract complex biological laws; and development of knowledge-data co-driven methods for simulation and design of biological systems. In terms of literature knowledge summarization, a large amount of biological knowledge is currently published in the form of literature, which lacks systematic organization, thus making it difficult to be applied in synthetic biology. As the cost of manual organization is too high, the development of artificial intelligence methods to extract normalized knowledge graphs from texts is an urgent breakthrough technology. In several biological function prediction problems, artificial intelligence models have made some breakthroughs, but because the models are too complex, the knowledge learned by the models is difficult to be transformed into biological knowledge that can be understood by human beings, such as a variety of bioparts, molecular physicochemical properties, and the operation law of biological systems, and thus the deepening of the understanding of biology is relatively limited and difficult to be directly

applied. For the simulation model of biological system, although it can effectively assist the design of synthetic biology and can also evaluate the design results and reduce the cost of experimental screening, it can only realize the preliminary simulation model of biological system at present. Due to the complexity of the biological knowledge system, the artificial intelligence model based on data and knowledge is still immature and does not make good use of the structured knowledge system to constrain the optimization space. It is difficult to effectively reduce the amount of data required. For the biopart/module design, the existing AI models are mainly data-driven models, and how to utilize the existing knowledge to guide the design of artificial bioparts is one of the key issues in synthetic biology.

Objectives and Breakthroughs: Develop artificial intelligence technology to automatically mine biological knowledge from literature. Based on the existing machine learning models in the field of natural language processing, the automatic extraction of knowledge graphs from literature has been initially realized. On this basis, biological functions and knowledge are defined in a systematic and standardized manner. The architecture of the relevant machine learning models is optimized to adapt them to the task of biological knowledge extraction, and through a certain degree of data cleaning, the construction of the relevant biological knowledge base can be achieved. And after a certain amount of data cleaning, the construction of relevant biological knowledge database can be realized .

Develop biological knowledge extraction methods based on interpretable AI technology, deeply analyze the internal mechanism of AI models, study how the internal structure of the model summarizes and expresses biological knowledge, and realize the extraction of complex biological laws from AI models.

Develop knowledge-data co-driven biological system simulation technology and software platform, establish a knowledge-data co-driven biological system simulation framework based on mathematical mechanism and AI technology, effectively reduce the complexity of the simulation, and realize long-time and accurate simulation of cross-scale large-scale biological systems. At the cellular level, it is necessary to develop a digital twin cell simulation model. At the molecular level, it is necessary to develop a long-term, large-scale simulation method for the dynamics of biomacromolecule interactions.

Construct knowledge-data co-driven feature representation and application of biological complex laws. For the laws in the spatiotemporal evolution process of specific life systems, make use of the advantages of mathematical and artificial intelligence

models to reduce the dimension of the high-dimensional spatiotemporal complex laws. Master the working mechanism of artificial intelligence models and incorporate the good representation of biological knowledge into the artificial intelligence models for designing bioparts and biological systems, so as to maximally minimize the design space and improve the success rate of design.

Challenges: Many basic issues lack a more systematically standardized definition of biological functions and knowledge, and existing machine learning models are difficult to adapt and effectively extract biological knowledge from literature automatically. The current natural language processing technology is still difficult to realize high-precision knowledge graph extraction, and it is difficult to match the extracted knowledge graph with the existing biological knowledge system, and the reliability of the extracted knowledge is difficult to be guaranteed.

The existing generalized interpretation methods in fields such as image and text are not necessarily applicable to problems in synthetic biology. Most AI models are complex, and their internal operation mechanism is not clear enough, which is a black-box for the user, thus it is easy to cause misinterpretation.

Different from the operational contexts of simple system functions, subjective biases in understanding biological systems can lead to simulations that do not match the actual situation. Constructing cellular digital twins involves interactions across multiple temporal and spatial scales, making model construction difficult. For the simulation of macromolecular interaction dynamics, there are diverse processes for coarse-grained simulation of interaction dynamics between biological macromolecules. The increase in the number of components involved in biological systems may lead to problems such as difficulty in modeling or solving mathematical and physical models. Currently, there is a lack of methods for constructing digital twin models of cells that can achieve complex functions, and the construction of visual simulation platforms requires multidisciplinary cooperation. For the simulation of macromolecular interaction dynamics, we mainly face the problems of lack of standardization of coarse-grained pathways and the diverse classification of protein interaction modes.

Part of the biological knowledge is too complex, and it is difficult to summarize its laws from the data, or to represent it formally by methods such as model interpretation. The theory of dimensional analysis in spatiotemporal evolutionary systems is lacking. The types of laws in different life complex systems vary greatly, and how to standardize and automate classification and processing for different data or types of laws is one of the

key challenges. Deep learning models are one of the most widely used AI techniques, but they are mostly based on data learning, and it is difficult to embed physicochemical and biological knowledge into the models and achieve better design results.

Expected Progress Recently: Establish biological knowledge graphs and related databases in major subfields of synthetic biology, with adaptable models and software capable of automatically extracting knowledge from relevant literature. Common machine learning models all have interpretable methods to help people understand the mechanism and the knowledge the models have learned. Establish a preliminary knowledge-data co-driven biological system simulation framework to simulate biological system functions. Obtain effective feature representations of synthetic biological data and use the patterns contained in the data to conduct intelligent-assisted design.

Expected Progress by 2030: Develop artificial intelligence methods to achieve systematic, automated, and intelligent construction of knowledge graphs, with induction and classification accuracy reaching the level of manual annotation. Achieve basic white-boxing of machine learning models, enabling the automated extraction of important biological knowledge or patterns. Establish a whole-cell modular simulation platform and a systematic, standardized coarse-grained simulation process for biological macromolecules. Develop more general methods for reducing the dimensionality of complex biological systems, advance targeted explainable artificial intelligence technologies, and effectively utilize the rules embedded in biological knowledge bases and data for intelligent design assistance.

Potential Solutions

Multi-disciplinary cross-cooperation is crucial. For example, through in-depth collaboration between biologists and experts in computer natural language processing, systematic and standardized definitions of biological functions and knowledge can be established. Then, relevant machine learning models can be developed and optimized to adapt to the task of extracting relevant biological knowledge. Additionally, software targeting specific problems can be developed. For instance, in the case of enzyme metabolic pathways, information on enzyme-catalyzed pathways can be obtained from the literature, and accurate annotations of gene functions can be achieved.

Firstly, based on the main functions in traditional biology, the functional bioparts/modules in existing living organisms are sorted out. Secondly, the functional

bioparts/modules in living organisms are added, merged, and expanded by combining the existing synthetic biology functional modules. Under the constraints of these structured knowledge, more accurate text cleaning tools are developed based on AI technology, and the quality of the samples is strictly controlled in order to improve the accuracy of the model. Optimize the AI model to promote the adaptation of the extracted knowledge to the existing biological knowledge system.

Interpretations should be tailored to the characteristics of neural networks and specific biological problems. For example, for the analysis of DNA/RNA sequence-related models, researchers usually analyze which motifs and combination rules are learned by the model, so interpretation algorithms need to be developed to extract the motif combination rules and match the existing motif databases, such as JASPAR. For the design of industrial enzymes or other proteins, it is necessary to study the key sites in the amino acid sequence, the secondary structure and the logic of interactions.

Good knowledge extraction is based on excellent models, and a weak predictive performance results in low reliability of the extracted knowledge. Therefore, firstly, it is necessary to improve the performance of the model. Secondly, it is necessary to combine the latest research results of model mechanisms to accurately identify the patterns recognized by the model and gradually improve the automation of model recognition pattern extraction.

Based on the biological knowledge system to construct the mathematical model of biological system, use the data to construct the artificial intelligence model assembly to form a knowledge-data co-driven biological system simulation framework, and initially realize the digital twin of cell function. When introducing the gene regulatory network, in order to reduce the difficulty of constructing the mathematical model, the key interactions and relaxation time constants of different processes can be extracted through short-time single-molecule layer surface simulation, and different processes can be reasonably coarsely grained to reduce the amount of computation, so as to realize the introduction and portrayal of the multilevel gene regulatory network in the digital cell. When simulating biomacromolecule interactions, 2-5 special cases of multiple protein interaction modes at the second level are selected to try to realize the coarse-graining simulation and compared with the all-atom simulation to study and summarize the common laws in the coarse-graining process. For some of the mathematical and physical models with unsatisfactory simulation results but with experimental data, deep learning models can be trained to improve the simulation accuracy and simulation speed.

Promote the integration of mechanism model and artificial intelligence model to improve simulation accuracy. Starting from the first principle, a new neural network architecture can be developed to simulate the physics and kinetic properties of molecules, etc., with full consideration of the existing knowledge and laws, so that the mechanistic model is highly integrated with the deep learning model. For the construction of cellular digital twins, based on the development of a whole-cell simulation model of digital twins with gene regulatory networks, enrich and improve the functions of the digital cells, and at the same time join hands with computer vision and other related teams to realize the interactive visualization and computation of the digital cells, so as to enhance the virtual assisted design capability of synthetic organisms. For the simulation of macromolecular interaction dynamics, it is necessary to systematically organize and classify protein interaction modes, and select 5-10 protein interaction special cases for each type to carry out coarse-graining processing, establish the standard of coarse-graining simulation process for each type of protein interaction, and ultimately summarize whether there exists a unified system of coarse-graining simulation process for the known types of protein interactions.

Computer simulation using numerical resolution of partial differential equations and discrete models is used to classify the spatiotemporal evolutionary laws during bacterial spatial expansion, tumor evolution, etc. Define new order parameters, and identify the key interactions and their quantitative relationships with order parameters for specific classes of laws. Complex spaces laws embedded in high-dimensional data can also be dimensionally reduced using AI models so that they can be understood in low-dimensional manifolds and better applied to design. In addition, the use of training AI models to construct representations of high-dimensional features is also an important research methodology, where knowledge and laws can be better represented in the new representation space, and the data of the components desired for design can be more easily distinguished from their performance in the new representation space, e.g., large-scale pre-training models such as BERT are widely used in biological sequence design and function prediction.

By integrating the existing dimensionality reduction processes based on time series data and developing dimensionality reduction processes based on spatiotemporal evolutionary data, we can try to integrate the former into the latter system, or develop targeted standards for the above two separately, so as to realize the seamless integration of data. For some of these dimensionality reduction processes, neural network methods

can be used. Bio-knowledge embedding models rely on the interpretation of the model and, as the accuracy of the model interpretation improves, bio-knowledge can be more accurately embedded in a way that is suitable for deep learning models. For example, in the deep learning model of DNA regulatory sequence function, it has become clear that the convolution kernel in the first layer of the convolutional neural network is learning the motif of the DNA sequences, and therefore the known motif from databases such as JASPAR can be used directly as the convolution kernel in the first layer, which enables better prediction results to be achieved.

2.2.4.3 Develop Artificial Intelligence-driven Dry-wet Closed-loop Experimental System

Current Technologies: Dry-wet closed-loop experimental iterative systems have achieved successful designs for some synthetic biology problems, but the decisions in them are mostly empirically dependent and difficult to reuse in other application scenarios. Adaptive experimental design can more effectively utilize the limited wet experimental opportunities to explore the grand sample space and find samples that meet the design goals with fewer iterations. However, the current machine learning-based adaptive experimental design methods are still relatively preliminary, mainly providing auxiliary design functions, still relying on manual judgment, and providing very limited help for automated experimental design. Therefore, high-level adaptive experimental design models for specific design problems urgently need to be researched and developed in order to make the dry-wet closed-loop experimental system realize pipeline operation and achieve better screening results with fewer experiments.

Objectives and Breakthroughs: Artificial intelligence-driven "design-build-test-learn" closed-loop development process, for specific problems, based on biological knowledge and existing experimental data, to develop machine learning models capable of active learning, automated experimental design and experimental parameter tuning, etc., to achieve a high level of adaptive experimental design and automated pipeline platform operation, and lower the barrier to entry for using the platform.

Challenges: The solutions to many synthetic biology problems rely on experience rather than artificial intelligence models trained on existing datasets and knowledge bases. The subsequent effects depend on the level of the initial manual design. After obtaining high-throughput experimental data, there is a lack of feasible methods and strategies on how to organically combine the newly obtained experimental data with existing datasets

and knowledge bases and train AI models. The performance of AI models often depends on parameter adjustment. If the model is not adjusted for parameters, it may exhibit different performances when applied to different datasets of the same problem. Models that require manual adjustment are not conducive to the formation of pipeline. Currently, there is still a lack of research on the adaptation between intelligent models and wet laboratory experiment process modules.

Expected Progress Recently: Preliminary construction of an adaptive experimental design system platform for core problems in synthetic biology.

Expected Progress by 2030: Artificial intelligence-driven platform for dry-wet closed-loop experimental development system forms pipeline.

Potential Solutions

Train AI model to assist in the design of the initial samples with respect to the characteristics of the problem and conditions such as existing data and knowledge. Develop incremental learning methods that can optimize and iterate the original AI design model using newly generated experimental data, thereby increasing the probability of designing an effective sample. On the one hand, machine learning models capable of active learning are developed for specific problems. On the other hand, more parameter variations are taken into account in the machine learning models to further improve adaptive experimental design. The intelligent model can optimize the key steps in the wet experiment and coordinate the various process modules of the wet experiment according to the feedback of the experimental results, so as to improve the overall experimental efficiency, safety and reproducibility.

2.2.5 Summary

The development of AI technology in synthetic biology is still in its infancy. Due to the uniqueness of biological data, relevant AI methods and theoretical studies are still relatively lacking. The developed methods and models have certain limitations, and the accuracy of predictive simulation of synthetic life systems still needs to be improved. With the continuous development of high-throughput technology and the accumulation of huge amounts of data, artificial intelligence technology has shown great acceleration effect on the development of synthetic biology, and has made certain breakthroughs in

improving the design capability of biological systems, learning and representation of complex biological rules, and automated experimental design of "dry-wet closed-loop", etc. The breakthroughs have enabled us to more effectively understand and manage highly complex biological laws, to reduce costs and increase efficiency in synthetic biology research and daily production, and to realize applications in the fields of medicine, agriculture and industry.

References

[1] Eslami M, Adler A, Caceres R S, et al. Artificial intelligence for synthetic biology. Communications of the ACM, 2022, 65(5): 88-97.

[2] Lopatkin A J, Collins J J. Predictive biology: modelling, understanding and harnessing microbial complexity. Nature Reviews Microbiology, 2020, 18(9): 507-520.

[3] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589.

[4] Chen Y, Banerjee D, MuKhopadhyay A, et al. Systems and synthetic biology tools for advanced bioproduction hosts. Current Opinion in Biotechnology, 2020, 64: 101-109.

[5] Gallup O, Ming H, Ellis T. Ten future challenges for synthetic biology. Engineering Biology, 2021, 5(3): 51-59.

[6] Kitney R I, Bell J, Philp J. Build a sustainable vaccines industry with synthetic biology. Trends in Biotechnology, 2021, 39(9):866-874.

[7] Zhao X Y, Zhang H, Li X F, et al. An evolutionary perspective on quantitative biological principles and synthetic life design. Synthetic Biology Journal, 2022, 3(1):6-21.

[8] Zhang T, Leng M T, Jin F, et al. Overview on platform for synthetic biology research at Shenzhen. Synthetic Biology Journal, 2022, 3(1):184-194.

Expanded Reading

EBRC. Engineering Biology: A Research Roadmap for the Next-Generation Bioeconomy. 2019.

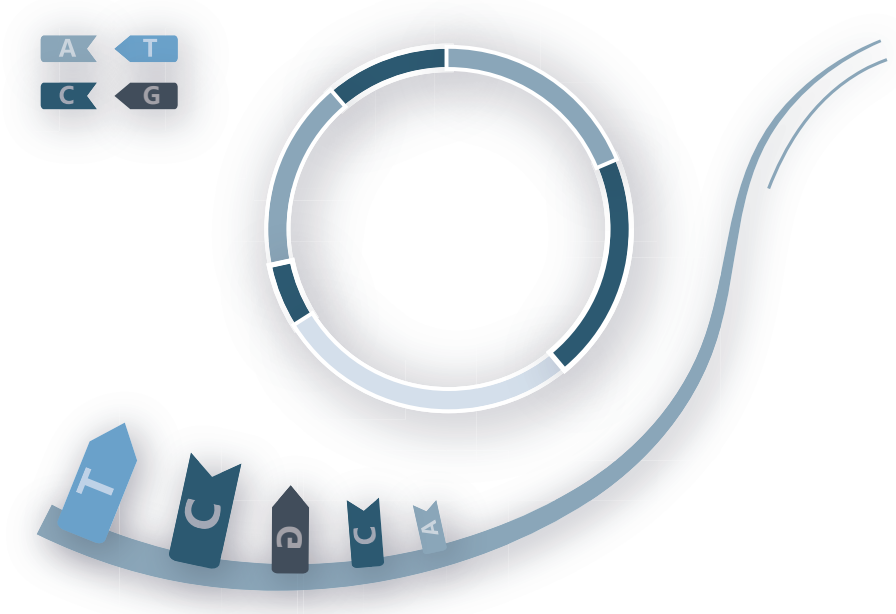
ERASynBio. Next steps for European synthetic biology: a strategic vision from ERASynBio.2014.

Synthetic Biology Leadership Council. Biodesign for the Bioeconomy: UK Synthetic Biology Strategic Plan 2016.2016.

3 Enabling Technologies

Enabling technologies refer to technologies that can be widely used to enhance existing technologies and gain high benefits. This chapter focuses on 12 essential enabling technologies, including DNA sequencing, synthesis and assembly, gene editing, protein design, genetic circuits, chassis cells, cell-free systems, artificial multicellular systems, organoid engineering, unnatural amino acids encoding and synthetic biosystems, synthetic hybrid biotic-abiotic systems, biofoundries, and biopart resource and information platforms, and a 2030-oriented prognosis for the iterative development of these technologies.

DNA Sequencing, Synthesis and Assembly



Authors Dai Jun-Biao, Shen Yue, Li Bing-Zhi

3.1 DNA Sequencing, Synthesis and Assembly

3.1.1 Abstract

Genomic deoxyribonucleic acid (DNA), as the carrier of genetic information, relies on underlying enabling technologies of "reading" (sequencing technologies) and "writing" (synthesis and assembly technologies) to support synthetic biology research and drive downstream industrial applications. Sequencing technologies enable the digital data interpretation of biological resources. The research frontier focuses on enhancing interpretation accuracy while achieving breakthroughs in efficiency, throughput, and cost, thereby broadening the range of detectable biomacromolecules. By excavating and functionally modifying genetic information, synthesis and assembly technologies deepen our understanding of life phenomena and facilitate the downstream application development of life science big data. A key research focus lies in advancing systematic DNA manipulation capabilities, including improvements in synthesis/assembly length, efficiency, and accuracy. Against this backdrop, this section systematically reviews cutting-edge developments in DNA sequencing, synthesis, and assembly technologies, identifies future directions and bottlenecks, and proposes effective strategies to advance underlying technologies supporting synthetic biology research and industrial translation.

3.1.2 Technical Overview

3.1.2.1 DNA Sequencing

In the 1970s, Rui Wu pioneered the position-specific primer extension strategy, which was refined by Frederick Sanger to develop the dideoxyribonucleotide chain termination method ("Sanger sequencing") in 1977, marking the advent of genomic sequencing. This method's core principle involves chain-terminating dideoxyribonucleotide incorporation combined with gel electrophoresis [1]. Post-2005, massively parallel sequencing (MPS) technologies emerged, propelling genomics into the high-throughput era and significantly accelerating scientific research and technological applications. Mainstream MPS technologies include pyrosequencing, sequencing-by-synthesis (SBS), and sequencing-by-ligation (SBL). The read length of MPS technologies typically ranges from 100 bp to 400 bp, and sequencing throughput has been dramatically enhanced (up to

6 Tb/Run) with approximately 99.7% accuracy [2]. Emerging since 2008, single-molecule sequencing technologies enable amplification-free, single-molecule real-time sequencing, thus achieving remarkable read lengths (up to 4 Mb) and direct detection of modified bases (e.g., 5-methylcytosine, 6-methyladenine). However, their accuracy (maximum 98%) has not yet reached the level of high-throughput sequencing technologies. Mainstream approaches currently follow two technical pathways: single-molecule fluorescence signal detection and nanopore electrical signal analysis [3,4].

3.1.2.2 DNA Synthesis

DNA synthesis methods are broadly categorized into chemical and biological approaches. Chemical synthesis is relatively mature, with the phosphoramidite triester synthesis method being the most widely used oligonucleotide synthesis technique, involving cyclic steps of deprotection, coupling, capping, and oxidation [5]. Breakthroughs in enzymatic DNA synthesis have emerged, including methods based on terminal deoxynucleotidyl transferase (TdT), TdT-deoxyribonucleoside triphosphate (dNTP) conjugates (TdT-dNTP), and hybrid enzyme-mediated systems, though these remain largely at the proof-of-concept stage [6]. Instrument development has evolved through two critical phases since the 1990s: first-generation column-based synthesizers progressed to second-generation high-throughput chip-based synthesizers. Recent years have witnessed growing research investment in enzymatic synthesis technologies and equipment development, though these remain in early-stage technological maturation.

3.1.2.3 DNA Assembly

DNA assembly stands as a pivotal technology for synthetic biology research, particularly in artificial genome design and construction. Assembly techniques are stratified by fragment size into short- and long-DNA fragment assembly. DNA assembly typically employs enzymatic or *in vivo* methods such as ligase chain reaction (LCR) and polymerase cycling assembly (PCA). For 10–100 kb DNA fragments, diverse assembly methods, including BioBrick, BglBrick, Golden Gate, SLIC, SLiCE, LCR, CPEC and Gibson assembly, enable hierarchical construction. However, current *in vitro* assembly outputs remain insufficient for subsequent experimental demands. The assembly of ultra-large DNA fragments (100 kb to 1 Mb) requires microbial recombination systems. Representative examples include the RecA and Red/ET systems in *Escherichia coli*, the

BGM system in *Bacillus subtilis*, and homologous recombination-based technologies in *Saccharomyces cerevisiae*, such as TAR-coupled yeast transformation, the DNA assembler, and CasHRA [7]. To address *in vitro* amplification challenges, Masayuki et al. developed a high-fidelity technology for amplifying large DNA fragments (10 kb–1 Mb), achieving superior fidelity at the megabase scale [8,9].

To enhance gene synthesis throughput and reduce costs, miniaturized and automated gene synthesis technologies integrating DNA synthesis and assembly have also made new progress. In 2011, researchers developed a chip-based combinatorial enzymatic approach consolidating oligonucleotide library synthesis, amplification, error correction, and gene assembly on a single microchip, significantly streamlining workflows. Advances in oligonucleotide synthesis have enabled automated DNA assembly through docking silicon wafer reactors and enzymatic DNA splicing. Despite rapid technological progress, critical challenges persist in the efficiency and fidelity of ultra-long fragment assembly, as well as in the integration of automatic systems.

3.1.3 Roadmaps

Current Status		
Massively parallel sequencing: Maximum sequencing spot density of 6 M/mm²; 4 M/mm² sample density on an active sensor array, 300 bp paired-end read lengths; Single-molecule sequencing: 4 Mb read length, 98% accuracy, and 245 Gb maximum single-run throughput per flowcell. Spatial transcriptomics: 100 nm-resolution spatial transcriptome, 6.5 mm × 6.5 mm capture area, and 1,800 genes captured within 100 μm scale binning.		
Objective 1: High-Throughput, Low-Cost, Long-Read Genomic Data Acquisition		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Increase pixel density of detection units. Develope biochemical systems based on novel luminescent substrates and sequencing enzymes.	**Single-device throughput: 100 Tb/day; Cost: 100 RMB per whole genome; Read length: 5 Mb-level.** • Upgrade field-of-view and resolution using new generation semiconductor sensors. • Establish novel surface chemistry treatment for flowcell reuse. • Develope genome extraction and library preparation methods compatible with ultra-long fragment sequencing. • Optimize high-precision sequencing biochemistry and associated algorithms.	**Single-device throughput: Petabyte (Pb)-class/day; Cost: 10 RMB per whole genome; Read length: 10 Mb-level.** • Design new fluorescent dyes to enhance signal-to-noise ratio. • Develope single-channel sequencing technology based on bioluminescence. • Engineer sequencing enzymes with higher activity and stability; and engineer efficient sample loading methods. • Design novel nanopore proteins to improve raw signal quality.

Objective 2: Hightly-efficient Genomic Data Reading and Analysis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Large-scale front-end sequencing signal acquisition capability. Real-time high-throughput data analysis capability.	**Real-time sequencing and data analysis at Tb-level throughput.** • Develope high-density, large-array front-end acquisition chips. • Build high-throughput real-time analysis system architectures. • Implement high-performance processing algorithms (e.g., distributed computing, heterogeneous computing).	**Real-time sequencing and data analysis at 100 Tb-level throughput.** • Customize efficient data processing chips dedicated to sequencing workflows. • Explore novel computational chip technologies and high-efficiency algorithm.
Objective 3: High-Precision, Multi-Dimensional, and Multi-Omic Genomic Data Acquisition		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Increase gene capture and enlarge field-of-view at single-cell resolution. Enable simultaneous nucleic acid and protein information acquisition.	**Nanometer-scale spatial resolution. Simultaneous detection of nucleic acid and protein information.** • Establish related sequencing technologies at 500 nm spatial resolution and beyond. • Develope non-oligo dT capture techniques. • Increase gene capture efficiency in single cell level. • Create integrated nucleic acid and protein co-detection technologies.	**High-resolution and large-field spatiotemporal multi-omics technologies.** • Fabricate >10 cm-scale chips for capture. • Design compatible biochemical systems on chip. • Develop algorithmic tools and supporting equipment. • Construct spatiotemporal atlases for key model organisms or tissues.

Figure 1 DNA sequencing roadmap

Current Status		
DNA chemical synthesis: Error rate of 0.1%; DNA enzymatic synthesis: Length of 60 nt; DNA chip synthesis: Throughput of million-base level, yield at fmol-level, cost $\leqslant 10^{-3}$CNY per base; Modified nucleic acid synthesis efficiency: 97%–98%; *In vitro* DNA assembly length: >100 kb.		
Objective 1: High-Fidelity Synthesis of Long Single-Stranded DNA (300 nt to >1000 nt)		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance length and fidelity through optimized chemical synthesis systems and enzymatic approaches.	**High-fidelity de novo synthesis of 300–500 nt single-stranded DNA.** • Enhance single-cycle specific deprotection; reduce synthesis cycles. • Reduce error rate to <0.1%. • Increase reaction yield to >30%.	**High-fidelity de novo synthesis of >1000 nt single-stranded DNA.** • Discover novel enzymes and compatible synthetic monomers. • Improve chemical reaction efficiency to 99.95%. • Achieve yield >60% with error rate <0.1%. • Explore novel oligonucleotide units and hybrid enzyme systems.

Objective 2: High-Yield, High-Throughput, Low-Cost DNA Synthesis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Innovate synthesis principles and chip design to enhance yield, throughput, and application adaptability.	**pmol-level DNA synthesis with million-level throughput at low cost.** • Increase chip reaction site density per unit area. • Expand chip spatial reaction area; optimize chip physical structure. • Elevate synthesis throughput to million-level. • Achieve cost $\leqslant 10^{-5}$ RMB/nt at million-level throughput.	**nmol-level DNA synthesis with ten-million-level throughput at low cost.** • Develop novel chip materials and compatible surface modification processes. • Boost reaction-specific surface area for nmol-level yield. • Explore high-throughput parallel synthesis and controlled high yield product separation. • Achieve cost $\leqslant 10^{-8}$ RMB/nt at ten-million-level throughput.
Objective 3: Efficient Synthesis of Modified DNA and RNA		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop novel modified monomers and synthesis pathways to enhance efficiency of chemically modified nucleic acids.	**High-efficiency synthesis of modified nucleic acids.** • Design novel, high-performance phosphoramidite protecting groups. • Engineer novel chemically modified monomers. • Enhance the stability of modified nucleic acids. • Develop efficient mirror-image nucleic acid synthesis methods and supporting amplification systems. • Optimize biochemical systems to achieve $\geqslant$98% single-cycle synthesis efficiency for modified monomers.	**Efficient synthesis of optically pure chiral modified nucleic acids.** • De novo design of modified nucleic acid structures. • Develop novel high-efficiency synthesis methods. • Achieve >95% single-cycle synthesis efficiency and >98% selectivity.

Objective 2: High-Yield, High-Throughput, Low-Cost DNA Synthesis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance DNA sequence prediction/design capabilities and enable reliable construction of ultra-long DNA fragments via one-step or iterative elongation strategies.	**Reliable assembly of ⩾100 kb DNA fragments.** • Integrate sequence design algorithms to increase assembly success rate of long fragment DNA to 90%. • Establish ⩾ 3 reliable assembly systems for prokaryotic/eukaryotic model organisms. • Reduce design-to-construction cycle for 100 kb fragments to ⩽2 weeks.	**Reliable assembly of ⩾10 Mb DNA fragments.** • Implement computer-aided sequence design and adaptive one-step/iterative elongation assembly strategies. • Establish ⩾ 5 reliable assembly systems for prokaryotic/eukaryotic model organisms and non-model species. • Reduce design-to-construction cycle for 10 Mb fragments to ⩽1 month. • Enable site-specific or domain-specific chemical modifications (e.g., methylation).

Figure 2 DNA synthesis and assembly roadmap

3.1.4 Technical Pathways

3.1.4.1 DNA Sequencing

Current Technologies: Over the past decade, advancements in MPS technologies have driven the diversification and refinement of commercial product lines tailored to diverse user needs and application scenarios [10]. For national-scale genome projects, the most critical parameters are high throughput and low cost. State-of-the-art instruments currently deliver whole-genome data for over 100 individuals per machine daily. To further reduce per-base sequencing costs, sequencing spot density continue to rise, with leading technologies achieving densities of about 6 M/mm², and products targeting >10 M/mm² under development. For clinical applications, reducing turnaround time while maintaining sufficient data output is paramount. Emerging sequencing technologies based on CMOS sensors or single-molecule sequencing show significant potential in this domain. Nanopore-based single-molecule sequencing deciphers nucleotide sequences by detecting continuous electric current signals generated as base units on nucleic acid molecules pass through a protein nanopore. This technology offers read lengths far exceeding high-throughput methods (average above 10 kb). High throughput comparable to that of high-throughput sequencing is achieved through the continuous capture and sequencing of fragments by individual nanopores, coupled with the parallel operation of multiple sequencing units. With sequencing speeds of hundreds of bases per second and multi-chip parallelization, nanopore systems can complete individual whole-genome sequencing in hours, demanding ultra-fast data processing and analysis. According to public reports, the highest-throughput nanopore sequencer (e.g., Oxford Nanopore's PromethION) generates up to 10 Tb of data per run. Spatial omics, recognized as the 2020 "Method of the Year" by *Nature Methods*, enables *in situ* multi-omic profiling (genomic, transcriptomic, and epigenomic,etc.) within tissues. Current mainstream spatial transcriptomics technologies achieve 100 nm resolution with a 6.5 mm × 6.5 mm capture area yet remain inadequate for single-cell spatial analysis or large tissue sections. Recently, Stereo-seq technology (500 nm resolution, 10 mm × 10 mm capture area) has emerged, outperforming competitors in resolution, capture area, and gene detection capacity [11].

To meet the needs of exploring cutting-edge scientific issues in life sciences and medicine, the further development of sequencing technology will focus on basic performance (such as sequencing throughput, cost, and read length), application

requirements (such as portability, timeliness), and technological breakthroughs (such as sequencing accuracy, spatiotemporal dimensions, and multi-omics integration), etc.

Objectives and Breakthroughs: Read nucleic acid data with high throughput, low cost and long read lengths, and make breakthroughs in the biochemical systems of high-density large-array signal sensing and acquisition units, as well as high luminous efficiency substrates and sequencing enzymes. Achieve high-efficient genomic data reading and analysis, and break through the sampling ability of sequencing signals at the front end of large arrays and the real-time analysis ability of high-throughput data. Obtain high-accuracy, multi-dimensional and multi-omics genomic data, which improves the number of gene captured and sample area at the single-cell resolution, and enables the simultaneous reading of nucleic acid and protein information.

Challenges: The limitation of the spatial-bandwidth product of the fluorescence signal collection device leads to a mutual constraint of the optical field of view and resolution, which restricts the improvement of throughput. Currently, the cost of consumables for sequencing flow cells is high. Single-molecule sequencing at the 5 Mb level is often limited by the preparation process of the sequencing library. During the extraction and preparation of the genome, it is extremely vulnerable to various physical or chemical effects and may break. The relatively low sequencing accuracy and difficulty in long fragments library preparation are the biggest bottlenecks that currently limit the large-scale application of single-molecule sequencing.

Currently, the mainstream MPS technologies all rely on the “discontinuous polymerization sequencing method” to achieve the reading of nucleotide sequences. In each round of polymerization reaction, the “base extension” of different nucleic acid molecule copies collected by the same signal acquisition unit cannot be completely synchronous, and the signal-to-noise ratio gradually decreases as the number of sequencing cycle increases. The existing technologies based on fluorescence signal acquisition cannot meet the requirements for long read lengths due to factors such as the quenching of fluorescent molecules and the inactivation of sequencing enzymes as the phototoxicity increases. The activity and stability of sequencing enzymes, as well as the sample loading efficiency, are the main factors limiting the efficient acquisition of 10 Mb-level fragment data. The nanopore protein is a key component for converting physical sequence into electrical sequencing signals, and its performance determines the limit of the sequencing accuracy.

Due to the need to balance high throughput and low cost, there is a requirement to

extremely enhance the array density and scale of a single flowcell, which is thus limited by the size of the current-based nanopore reaction system. The existing data analysis processes and algorithms cannot meet the real-time processing requirements for terabyte (Tb)-level data. General computing platforms such as existing CPUs and GPUs contain excessive redundant designs, which restricts the improvement of their performance in dedicated scenarios for base prediction, and they are completely unable to meet the real-time analysis requirements for a throughput of 100 Tb.

The spatial resolution of currently established technologies (such as DBiT-seq technology, Visium technology, etc.) range from 20 to 100 nm, which is lower than the single-cell scale (<10 mm), and the area of the non-capture region accounts for 75%. Therefore, the obtained transcriptome information is discontinuous. The number of genes captured in a single cell is only several hundred, which is far lower than the number of genes obtained by current single-cell sequencing (5,000 to 8,000), and it is impossible to obtain nucleic acid and protein information simultaneously. Limited by various factors such as cost, hardware, and operational complexity, the detection area of the chips or similar capture tools provided by existing technologies is at the millimeter scale, and the analysis methods are limited to a two-dimensional plane, resulting in 3D information missing issue.

Expected Progress Recently: Achieve 100 Tb/day throughput, 100 CNY per whole genome, 5 Mb read lengths, Tb-level real-time analysis, nanometer-scale spatial resolution, and simultaneous nucleic acid-protein detection.

Expected Progress by 2030: Reach petabyte (Pb)-class daily throughput, 10 CNY per whole genome, 10 Mb read lengths, 100 Tb-level real-time analysis, and high-resolution, large-field spatiotemporal multi-omics.

Potential Solutions

Use semiconductor sensors for direct signal acquisition without traditional optics, combined with the development of commercial semiconductor sensor technology, to achieve simultaneous enhancement of field of view and resolution. Design special structures and surface treatment to solve the sample loading and sequencing flowcells reuse. Optimize genome extraction and repair, reduce mechanical force and chemical stimulation, and efficiently repair damage sites such as Abasic, Nick, Gap, Crosslink in fragments of native or introduced to enhance fragment integrity. Develop 2D, UMI or

other novel multi-copy biochemical schemes and supporting base recognition algorithms to improve sequencing accuracy by means of consensus sequence construction.

Enhance the efficiency of polymerization reaction, increase the copy number of nanoclusters or nanoballs, and develop novel fluorescent dyes with high luminous efficiency, anti-bleaching, water solubility, as well as adaptability to the parameters of optical components of the instrument. Develop single-channel sequencing technology based on bioluminescence, establish sequencing protocols, simplify sequencing process, and optimize sequencing reaction reagents to achieve rapid and accurate base sequencing. Further enhance the activity and stability of sequencing enzymes by protein engineering to support 10 Mb read length sequencing, and at the same time, develop loading methods based on magnetic beads, large/small molecule additives, or other physicochemical principles to increase the capture efficiency of ultra-long fragments of nucleic acid molecules. Develop methods for the accurate prediction of polymers' structures, and utilize structural and sequence databases to excavate nanopore proteins with higher resolving power. Engineer and modify them to optimize their sequencing performance, and to produce highly consistent and high-resolution sequencing results.

Explore the size limit of nanopore reaction system to maximize the density and scale of front-end sampling circuit arrays. Starting from the top-level architectural design, solve the problem of acquiring and transmitting highly concurrent data streams, and at the same time, develop high-performance analysis and processing algorithms, such as distributed computing and heterogeneous computing, to realize real-time base prediction of the data.

Based on the properties of base prediction algorithms, customize sequencing-specific data processing chips. At the same time, actively focus on emerging technologies such as in-storage computing, optical computing, quantum computing, etc., and explore new principles of computing chips, systems and algorithms, and base prediction efficiency to achieve an order of magnitude improvement.

Establish adaptive sequencing technology with 500-nanometer-level spatial resolution, develop subcellular mapping tools, and improve the performance of transcriptome information acquisition. Develop non-oligo dT capture technology to avoid the problem of limited capture due to polyA degradation of RNA molecules. Optimize the number of branched probes per unit area of the chip and the biochemical reaction system to reduce the spatial resistance of the chip surface modification, enhance the efficiency of the interfacial reaction, and increase the number of genes captured in a single cell. Develop the spatiotemporal multiprotein detection technology based on high-throughput

sequencing to make up for the gaps in the field.

Develop large-field arrays greater than 10 cm in 2D dimension, and develop adaptive biochemical techniques and supporting sequencing, slicing and scanning equipment. Establish algorithmic tools for the registration between sequencing data and structural features, analysis over various resolutions, integration with single-cell sequencing data, and multi-omics joint analysis, so as to meet the needs of constructing spatiotemporal omics maps of key model organisms and organs.

3.1.4.2 DNA Synthesis and Assembly

Current Technologies: In synthesis technologies and equipment development, chemical synthesis currently achieves a maximum single-step efficiency of 99.5%, producing DNA strands of 200–250 nt with yields of 29%–35% [12]. Low-throughput synthesizers based on this method exhibit error rates of 1‰-3‰ and costs of 0.05–0.5 CNY/nt. The error rate for high-throughput chip-based synthesis is relatively higher (5‰–12‰). Although the cost is reduced by 2–3 orders of magnitude compared to traditional low-throughput synthesis methods, further reductions are still needed to support emerging applications such as large-scale genome synthesis and DNA storage. Enzymatic synthesis methods achieve 97% single-step efficiency but are limited to *de novo* synthesis of 60 nt with high error rates (>10%) [13,14]. High-throughput chip-based synthesis technologies (e.g., photochemical, electrochemical, inkjet, or integrated circuit-controlled systems) are constrained to fmol-level yields. Additionally, the demand for synthesis technologies has expanded from conventional nucleic acid synthesis to the modified nucleic acids to meet the requirements for stability and targeting in the development of nucleic acid-based drugs, RNA vaccines, and related applications. For DNA assembly, *in vitro* enzymatic methods combined with automation enable reliable assembly of 5–10 kb fragments, but success rates and efficiency decline as fragment length and complexity increase. Constructing 100 kb–Mb-scale genomes still relies on *in vivo* recombination systems, which face challenges in stability, process complexity, low automation, requiring significant manual input, with unpredictable timelines, and predictive design capabilities still needing further improvement.

Rapid advancements in life sciences and synthetic biology are driving increased demands on existing DNA synthesis and assembly technologies in terms of synthesis length, fidelity, throughput, cost-efficiency, yield, product diversity, and stability.

Objectives and Breakthroughs: For high-fidelity synthesis of long single-stranded

DNA (300 nt to over 1000 nt), improve both length and fidelity by optimizing biochemical systems in chemical and enzymatic synthesis methods. For high-throughput and low-cost DNA synthesis with high yield (pmol-nmol level), improve the synthesis yield and throughput by innovating the synthesis principle, optimizing the research and development of key raw materials and chips, and explore the adaptability of synthetic products for diverse applications. For efficient synthesis of modified DNA and RNA, develop new types of modified monomers, and combine with the optimization of the synthesis process to improve the synthesis efficiency of modified nucleic acids, and lay out the synthesis of mirror-image DNA and supporting amplification technologies. For efficient and high-fidelity assembly of long DNA fragments up to the genomic level, enhance DNA sequence prediction and design capabilities, reliably construct ultra-long DNA fragments using one-step or continuous extension strategy, and introduce chemical modifications (e.g., methylation) at specific sites or regions.

Challenges: The synthesis efficiency of the existing chemical method rarely exceeds 99.5%, and the theoretical yield of the 500 nt DNA product is less than 10%. In the current chemical synthesis cycle, acidic deprotection conditions can cause base loss, increasing error rates. Enzymes exhibit low catalytic efficiency and poor specificity, while monomer stability and single-step reaction rates remain suboptimal. There are few types of non-template synthetic enzymes available, and technical approaches remain relatively limited. The current mainstream high-throughput dot-matrix synthesis chips increase the density of reaction sites by reducing the reaction area per unit. Due to the extremely small-scale biochemical reaction systems, DNA synthesis yields are limited to the fmol level. To increase the yield, PCR amplification is required, which may introduce and amplify errors, and the non-uniformity can negatively affect downstream applications. Approximately 70% of the chemical reagents, chip materials, and key instrument components for DNA synthesis are imported, making cost difficult to control.

The synthesis throughput corresponds to the number of reaction sites on the chip. Further increasing the density of the reaction sites depends on more complex semiconductor manufacturing technologies and printing nozzles capable of ultra-high printing accuracy for high-density arrays. These processing technologies demand stringent standards, present significant technical challenges, and make further cost reduction difficult. Products synthesized by existing dot-matrix chips exist as mixtures, making it difficult to manipulate individual DNA molecules and limiting the adaptability of downstream applications.

Currently, the types of available modified monomers are limited. Their stability

during synthesis and drug delivery is low, making them prone to degradation, which further reduces the synthesis efficiency and drug efficacy. Reactions involving modified monomers exhibited lower synthesis efficiency than those with conventional monomers, making it difficult to synthesize long fragments. The current synthesis length is 20–30 nt. Existing DNA synthesis methods cannot produce optically pure chiral nucleic acid molecules. Sequence construction strategies for long fragments lack support from computer-aided tools and rely heavily on the experience of experimental personnel. Assembly efficiency and success rate are greatly affected by DNA fragment sequences complexity (e.g., GC content, secondary structure, repetitive sequences), fragment quantity, and the compatibility and expandability of host cells with exogenous DNA. Consequently, technical versatility is low and costs and timelines are difficult to control.

The construction of ultra-long fragments at the chromosomal level is influenced not only by the ability to assemble long-fragment DNA but also by the iterative and continuous construction of recombination system within organisms. Consequently, construction capacity and success rates are further limited by the efficiency and diversity of assembly systems, and efficient site-specific introduction of chemical modifications remains unattainable.

Expected Progress Recently: Achieve high-fidelity *de novo* synthesis of 300–500 nt long ssDNA. Realize pmol-level, high-throughput, low-cost DNA synthesis. Enable efficient synthesis of modified and mirror-image nucleic acids. Achieve reliable assembly of ⩾100 kb DNA fragments.

Expected Progress by 2030: Attain high-fidelity *de novo* synthesis of ssDNA longer than 1000 nt. Achieve nmol-level, high-throughput, low-cost DNA synthesis. Master efficient synthesis of optically pure, chiral, modified nucleic acids. Achieve reliable assembly of ⩾10 Mb DNA fragments with site-specific chemical modifications (e.g., methylation).

Potential Solutions

The development of novel phosphoramidite protecting groups is proposed to enhance deprotection specificity in single cycles, reducing the error rate to below 1‰ and achieving industry-leading standards. Additionally, the establishment of new biochemical methods, such as dual- or multi-nucleobase monomer synthesis, aims to reduce the number of synthesis cycles and increase reaction yields to over 30%.

Efforts will be made to discover novel terminal transferases and to improve their catalytic efficiency and specificity through protein design. Furthermore, the development of novel, reversible blocking synthetic monomers that are highly compatible with terminal deoxynucleotidyl transferases will be developed, with reversible blocking removal efficiencies exceeding 99% per minute and chemical reaction efficiencies reaching 99.95%. This approach is expected to enable yields of over 60% for 1000 nt single-stranded DNA with an error rate below 1‰. The design of novel oligonucleotide units (e.g., <10 nt random single-stranded DNA libraries) and the screening of enzyme hybrid systems, including polymerases, ligases, and recombinases, will be implemented. Advanced technologies such as inkjet printing, acoustic liquid handling, and microfluidics will be employed to achieve efficient synthesis of long single-stranded DNA.

Optimization of chip surface modification processes will increase the density of reaction sites per unit area and expand the effective spatial reaction area, enabling the efficient synthesis of DNA products at the pmol level. Furthermore, the physical structure of the chips will be improved through the use of masking and micro/nano-fabricating techniques to create multi-level patterned chips, thereby increasing synthesis throughput to million-level scale. The development of high-quality domestic reagents, consumables, and high-performance core components will reduce costs. By achieving self-reliance across the entire supply chain, from upstream raw materials to complete equipment, costs will be reduced dramatically to 10^{-5} CNY/nt.

Inspired by first-generation column-based synthesis carriers, the development of new chip materials and compatible surface modification processes will expand the reaction surface area, enabling production at the nmol level. Innovative strategies will be explored, including the binding, recognition, and mechanical manipulation of synthetic products and carriers, to achieve high-throughput parallel synthesis and controlled separation of high-yield products. By integrating functional modules and automation, synthesis throughput will reach tens of millions, reducing the cost per base to 10^{-8} CNY/nt and enabling the domestic production of high-performance equipment.

The development of novel and efficient phosphoramidite protecting groups will be pursued, alongside with the construction of new chemically modified monomers. This will improve single-cycle synthesis efficiency to over 98% and reduce error rates. Additionally, chemically modified monomers will be designed to enhance the stability of modified nucleic acids, enabling the efficient synthesis of modified nucleic acids longer than 50 nt and improving targeting specificity.

De novo design of modified nucleic acid molecules will be undertaken, and novel high-efficiency methods for synthesizing optically pure, chiral, chemically modified nucleic acids will be developed. These approaches are expected to achieve single-cycle synthesis efficiencies above 95% and selection efficiencies exceeding 98%.

Enzymatic assembly techniques and genetic engineering systems for common prokaryotic and eukaryotic model organisms will be optimized, expanding the number of usable model organism assembly systems to three or more. Partial or full-process automation will be explored. Building on existing chemical methods, tools such as polymerases and ligases will be employed to further enhance the *in vitro de novo* synthesis of methylated primers. Additionally, integrating methylation modification with long-fragment assembly technologies will enable the delivery of methylated fragments into cells. Relevant efficiency data and knowledge bases will be accumulated, and computer algorithms will be developed to design and predict sequences for long fragments, improving the success rate of 100 kb DNA fragment assembly to 90% and reducing the operational cycle to two weeks or less.

Full utilization of genomics interpretation databases and prior knowledge will further expand the number of available model organism assembly systems. Moreover, valuable non-model biological systems (five or more) and adaptable genome design and construction software will be developed to establish capabilities for genome-level ($\geqslant$10 Mb) design and construction within predictable timeframes of one month or less.

3.1.5 Summary

The rapid advancement of the latest technological revolution has promoted significant breakthroughs in life sciences, driven by the swift development of genomics reading and writing capabilities. These advancements have deepened research in life recognition across multiple scales, dimensions, and depths, as well as in control and application. However, the limitations of reading and writing technologies have become apparent, constraining the future development of the field. In sequencing, it is essential to enhance the technical accessibility and affordability in diverse application scenarios and to expand the capacity for acquiring and analyzing multi-layered, multi-dimensional data in life systems. For synthesis and assembly, significant improvements are needed in the stability and cost-effectiveness of technologies that have already been validated in principle. Furthermore, the integration of artificial intelligence and bioinformatics will

accelerate the development of design and construction capabilities for various artificial biological systems with research and industrial applications, along with the creation of automated software and hardware infrastructure. This will facilitate accurate, tunable, engineered, and economical high-efficiency construction. Through the integrated development of genomics reading and writing capabilities, the field of genomics technology can achieve greater breakthroughs.

References

[1] Sanger F, Nickelsn S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 1977, 74(12): 5463-5467.

[2] Foox J, Tighe S W, Nicolet C M, et al. Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. Nature Biotechnology, 2021, 39(9): 1129-1140.

[3] Karst S M, Ziels R M, Kirkegaard R H, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nature Methods, 2021,18(2): 165-169.

[4] De Coster W, Weissensteiner M H, Sedlazeck F J. Towards population-scale long-read sequencing. Nature Reviews Genetics, 2021, 2(9): 572-587.

[5] Beaucage S L, Caruthers M H. Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis. Tetrahedron Letters, 1981, 22(20): 1859-1862.

[6] Eisenstein M. Enzymatic DNA synthesis enters new phase. Nature Biotechnology, 2020, 38(10): 1113-1115.

[7] Ellis T, Adie T, Baldwin G S. DNA assembly for synthetic biology: from parts to pathways and beyond. Integrative Biology, 2011,8;3(2): 109-118.

[8] Su'etsugu M, Takada H, Katayama T, et al. Exponential propagation of large circular DNA by reconstitution of a chromosome-replication cycle. Nucleic Acids Research, 2017, 45(20): 11525-11534.

[9] Mukai T, Yoneji T, Yamada K, et al. Overcoming the challenges of megabase-sized plasmid construction in *Escherichia coli*. ACS Synthetic Biology, 2020, 9(6): 1315-1327.

[10] Goodwin S, Mcpherson J D, Mccombie W R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews, 2016, 17(5): 333-351.

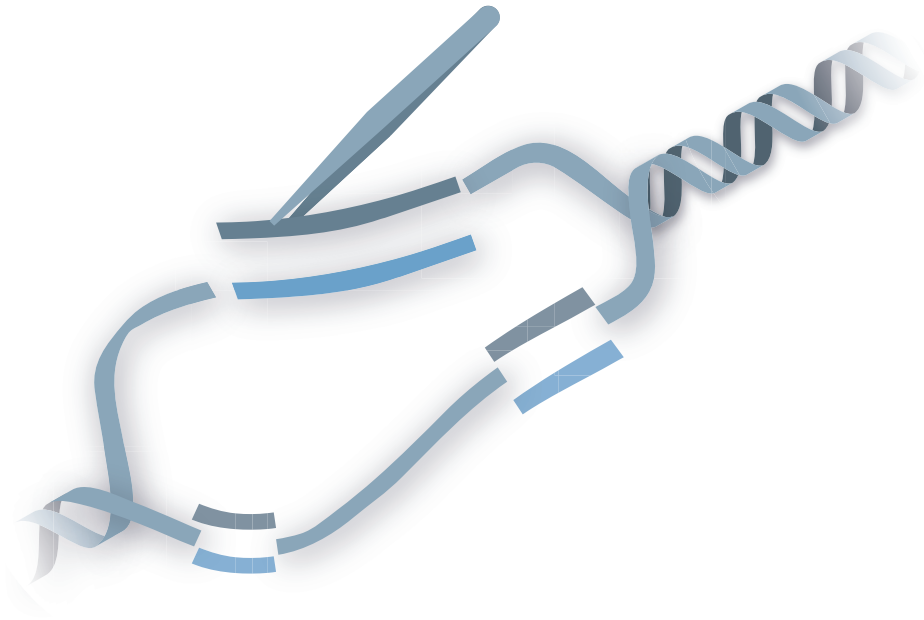
[11] Chen A, Liao S, Cheng M, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell, 2022, 185(10): 1777-1792.

[12] Caruthers M H. The chemical synthesis of DNA/RNA: our gift to science. Journal of Biological Chemistry, 2013, 288(2): 1420-1427.

[13] Palluk S, Arlow D H, De Rond T, et al. *De novo* DNA synthesis using polymerase-nucleotide conjugates. Nature Biotechnology, 2018, 36(7): 645-650.

[14] Lu X, Li J, Li C, et al. Enzymatic DNA synthesis by engineering terminal deoxynucleotidyl transferase. ACS Catalysis, 2022, 12(5): 2988-2997.

Gene Editing



Authors Xiang Hua, Gao Cai-Xia, Wei Wen-Sheng, Wang Hao-Yi, Yang Hui

3.2 Gene Editing

3.2.1 Abstract

Gene editing technology rewrites the genetic information by specifically deleting, replacing, inserting, or regulating the target genes, thus obtaining new functions or phenotypes. Gene editing is an important enabling technology in synthetic biology and plays a significant role in optimizing and reconstructing gene elements, circuits, systems, and networks. Currently, gene editing technologies mainly include three generations of underlying core technologies, namely the zinc finger nuclease (ZFN) technology, the transcription activator-like effector nuclease (TALEN) technology, and the CRISPR-Cas technology. In addition, there are two key core technologies developed based on the CRISPR-Cas system, namely base editing (BE) and prime editing (PE). Innovating the next-generation gene editing technology will further enhance its specificity, effectiveness, and safety, enable the writing of large DNA fragments, establish a new concept of the gene editing system, and achieve efficient delivery. Ultimately, it will enable the precise editing of any gene, fragment, or base in any species, providing important enabling technologies for life science research and the development of related industries.

3.2.2 Technical Overview

Currently, CRISPR-Cas gene editing technology, along with its derivative single base editing and prime editing systems, dominates the field of gene editing.

3.2.2.1 CRISPR-Cas Gene Editing Technology

The CRISPR-Cas system, derived from bacterial or archaeal adaptive immune systems, targets and cleaves invading foreign DNA or RNA viruses. From its initial discovery in 1987, the revelation of its DNA-targeting cleavage capability in 2007, to its development as a revolutionary genome-editing tool in 2012, the CRISPR system underwent extensive scientific exploration [1]. In 2013, the CRISPR-Cas9 system achieved gene editing in diverse eukaryotic cells, catalyzing a paradigm shift in genome engineering [2, 3]. In 2020, Emmanuelle Charpentier and Jennifer A. Doudna were awarded the Nobel Prize in Chemistry for their foundational work in discovering and elucidating

the DNA-targeting mechanism of CRISPR-Cas9. Over the past decade, CRISPR-Cas technology has been continuously optimized, and additional Cas nucleases—including Cas12a (Cpf1), Cas12b, Cas12i, Cas12f, Cas12j, and the RNA-targeting Cas13 (Type VI)—have been harnessed for genome editing [4, 5]. Current challenges for CRISPR-Cas technologies include mitigating off-target effects, improving delivery efficiency, and expanding applications in non-model species [6].

3.2.2.2 Base Editing Technology

Since the CRISPR-Cas editing technology based on DNA double-strand breaks poses potential risks in the field of medical applications, and the gene precise editing mediated by its homologous recombination repair is inefficient in plants, the team led by David R. Liu from the United States has successively developed the cytosine base editor (CBE) and the adenine base editor (ABE). Through the base editing technology that does not rely on DNA double-strand breaks, precise editing of some types of bases (i.e., base editing) has been achieved [7,8]. The CBE consists of two parts, namely the sgRNA and the fusion protein. The fusion protein is generally composed of a modified Cas9 protein (dCas9 or nCas9), a cytosine deaminase, and a uracil glycosylase inhibitor. The sgRNA pairs complementarily with the target site to guide the fusion protein to bind to the target site. The cytosine deaminase therein can transform the corresponding cytosine C in the non-target strand into uracil U through deamination. After DNA replication, U is further replaced by thymine T, and the uracil glycosylase inhibitor can inhibit the excision of U, ultimately achieving the precise editing from C to T. Similarly, the ABE can achieve the conversion from A to G. Recently, Chinese and foreign scholars have also established a new type of glycosylase base editor (GBE) based on CBE, which can achieve the specific substitution of C to G bases in mammalian cells. Similarly, based on ABE and the glycosylase that catalyzes the excision of inosine, adenine transversion editing from A to Y (Y = C or T) has been achieved. Currently, the base editing system still needs to be further improved in terms of the types of editing, the precision of the editing window, and its effectiveness in different species [9,10].

3.2.2.3 Prime Editing Technology

In 2019, David R. Liu's team integrated engineered nucleases with reverse transcriptases to develop prime editing (PE), a system capable of performing 12 types of

base substitutions, multi-base conversions, and small-fragment insertions or deletions in mammalian cells [11]. The system comprises a reverse transcriptase fused to nCas9 (H840A) and a prime editing guide RNA (pegRNA). Unlike standard sgRNA, pegRNA contains a 3′extension encoding a primer binding site (PBS) and a reverse transcription template. The nCas9 (H840A) introduces a nick in the non-target strand at the target site, exposing a free single-stranded region for PBS hybridization. The reverse transcriptase then synthesizes single-stranded DNA based on the provided template, after cellular repair, enables programmable edits of DNA sequence at 3′downstream of the protospacer adjacent motif (PAM). Despite broad applications, PE faces limitations such as large construct size, delivery challenges, inefficiency in large-fragment integration, and variable efficacy across species [12, 13].

3.2.2.4 Transposon-based Writing Technology

Beyond the above strategies, CRISPR-associated transposon (CAST) systems have emerged as promising tools for DNA writing since 2019. CAST leverages transposases to integrate target DNA fragments into specific genomic loci without double-strand breaks (DSBs) or reliance on HDR/non-homologous end joining (NHEJ) repair mechanisms, which is a promising technology for large DNA fragments writing. Early CAST systems were limited to prokaryotes, but recent optimizations in DNA-targeting activity and integration efficiency have enabled DSB-free, site-directed integration in human cells. However, functionality and efficiency remain areas for improvement [14].

3.2.3 Roadmaps

Current Status		
Three generations of foundational gene editing technologies—zinc finger nuclease (ZFN), transcription activator-like effector nuclease (TALEN), and CRISPR-Cas—have been established based on programmable protein modules targeting specific DNA sequences. Building on CRISPR-Cas, two pivotal core technologies, base editing (BE) and prime editing (PE), have further advanced the field. While CRISPR-Cas and its derivatives dominate current gene editing due to their simplicity and efficiency, challenges remain in specificity, precision, safety, and *in vivo* delivery efficiency, as well as editing efficiency in certain species.		
Objective 1: Enhance the Precision, Specificity, Safety, and *in Vivo* Delivery Efficiency of Existing Gene Editing Technologies		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Revolutionarily optimize existing CRISPR-Cas systems, base editors, and prime editors to address current limitations in off-target effects, delivery efficiency, editing efficiency, and application scope.	• Achieve editing of any gene and any specific single base pair in model organisms. • Enable highly efficient editing (>90%) of any gene in model organisms. • Realize simultaneous multi-locus editing/ regulation or combined gene editing and regulation in model organisms. • Establish novel gene editing delivery systems based on new concepts or strategies. • Utilize optimized editors in model organisms with no detectable off-target effects.	• Achieve editing of any gene and any specific single base in key non-model organisms (including animals, plants, fungi, industrial/environmental microbes, etc.). • Enable highly efficient editing (>60%) of any gene in key non-model organisms. • Improve delivery system specificity in specific tissues to achieve highly efficient *in vivo* gene editing or regulation (>90%) in targeted cells. • Realize quantitative, specific, and multiplexed editing in key model/non-model organisms or tissues/cells.

Objective 2: Develop Novel Concept-Based Gene Editing, Large-Fragment DNA Writing, and RNA Editing Technologies		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Discover non-CRISPR-Cas gene editing components through biological big data and rational design, elucidate their mechanisms and programmable targeting principles, and establish novel concept gene editing technologies, large-fragment DNA writing technologies, and RNA editing tools for efficient applications in animal, plant, and microbial cells.	• Develop original novel gene editing components, including nucleic acid-targeting and modification modules. • Decipher the structure-function mechanisms of novel nucleic acid-modifying components, explain its programmable principles and establish new principles and mechanisms for programmable gene editing. • Integrate programmable nucleic acid-targeting and modification modules to create non-CRISPR-Cas gene editing technologies. • Demonstrate at least one novel concept gene editing technology in model animals, plants, or microbial cells. • Establish a novel concept-based technology for precise writing of small DNA fragments (>5 kb) in eukaryotic genomes. • Achieve transcriptome-wide editing of any site and specific bases in key model organisms or experimental animals, with no detectable side-cut editing effect or off-target effects.	• Optimize novel gene editing systems to surpass CRISPR-Cas in compactness. • Achieve breakthroughs in gene editing components and theories, such as protein-independent DNA/RNA editing technologies. • Enhance the efficacy and specificity of novel gene editing underlying technologies and develop derivative core systems (e.g., base editing). • Apply novel underlying and derivative technologies in key model/non-model animals, plants, and microbes, achieving editing efficiencies comparable to CRISPR-Cas9. • Enable efficient and precise large-fragment DNA (>10 kb) writing in eukaryotic cells. • Realize transcript- and single-base editing in specific tissues of non-human primate models with no detectable off-target effects, and initiate clinical applications of RNA editing therapies.

Figure 1 Gene editing technology roadmap

3.2.4 Technical Pathways

Current Technologies: CRISPR-Cas gene editing has become the mainstream tool due to its simplicity of design, operational ease, and scalability. However, challenges such as off-target effects and PAM sequence limitations persist. Current strategies to mitigate these issues include rational design of Cas9 or high-throughput screening for high-fidelity variants (e.g., HiFi-Cas9), though these mutants often exhibit reduced on-target cleavage activity. Rational engineering has produced SpCas9 variants with expanded PAM recognition, such as SpRY (recognizing NRN, where R = A/G), achieving near-complete genome coverage. Nevertheless, optimizing, integrating, and validating diverse CRISPR-Cas systems to enable precise editing or regulation of more types of genes or base editors at any genomic locus across more model cell types remains a critical need [13, 15].

Existing tools enable DNA sequence editing and non-editing-based genome engineering, including gene regulation and chromatin remodeling. TALEN and CRISPR-Cas technologies introduce nicks or double-strand breaks (DSBs) at specific genomic sites, leveraging natural repair pathways for edits. Editing efficiency varies widely (2%-90%) depending on host and tissue type, with high-fidelity nucleases further reducing efficiency. In certain taxonomic groups, CRISPR-Cas systems remain inefficient or nonfunctional.

In addition to permanently altering the gene sequence, gene editing technology also includes achieving persistent gene repression and activation. By fusing site-specific DNA-binding proteins (zinc finger proteins, transcription activator-like effectors, and Cas proteins) with gene regulatory domains, the activation or repression of the desired genes can be carried out. Currently, up to six different genes can be regulated simultaneously, and the range of repression or activation in animal cells can reach hundreds of times. By constructing gene editing and regulatory systems based on the endogenous type I CRISPR-Cas system, it has been found that by adjusting the length of crRNA, the editing and regulation of different genes can be simultaneously achieved in prokaryotes.

Effective delivery is critical for *in vivo* gene editing. Three primary CRISPR-Cas delivery strategies exist: plasmid-based, Cas mRNA/sgRNA co-delivery, and Cas protein and sgRNA direct delivery. Key vectors include adeno-associated viruses (AAVs), extracellular exosomes, virus-like particles (VLPs), and lipid nanoparticles (LNPs). AAVs offer low immunogenicity, high infection efficiency and no pathogenicity, but are constrained by a 4.7 kb packaging limit, making it cannot package the SpCas9 gene and

its expression control elements (which have reached 4.7 kb) and sgRNA into the same AAV vector. Exosomes from O-blood group donors exhibit low immunogenicity and can encapsulate Cas9/sgRNA via electroporation, but scalable production and application protocols require refinement. VLPs, lacking viral genetic material, are safer alternatives compared to other virus-based delivery methods, which have also attracted much attention as potential gene drug delivery carriers, but remain under investigation. In March 2023, Zhang Feng's team engineered extracellular contractile injection systems (eCIS) to deliver Cas9 and base editors into eukaryotic cells, expanding delivery options [16]. Although the CRISPR system can be delivered into cells via different strategies, its sustained expression may lead to off-target-induced toxic side effects, necessitating precise control of gene editors over dosage and editing duration *in vivo*.

Current CRISPR technologies and CRISPR-derived technologies (base editing, prime editing) cannot efficiently achieve the precise insertion or replacement of large gene fragments in the genomes of eukaryotic cells. And CRISPR-associated transposon (CAST) systems struggle with the improvement in functionality and efficiency. Defense systems from bacteria and archaea, evolved through interactions with mobile genetic elements (e.g., transposons, plasmids, viruses or bacteriophages), offer untapped potential for novel genome editing tools due to their unique nucleic acid recognition, processing and foreign entity differentiation mechanisms. Similarly, eukaryotic regulatory proteins and transposons may inspire new editing innovations. RNA editing, meanwhile, awaits breakthroughs in novel principles and systems.

Objectives and Breakthroughs: Improve precision, specificity, safety, and *in vivo* delivery; achieve efficient editing of any gene *in vivo*; enable multi-gene, long-term regulation; and deliver gene editing tools to target cells in specific tissues and populations with controlled dosage and timing.

Develop gene editing technologies, large-fragment DNA writing technologies, and RNA editing technologies based on new concepts or systems. Develop underlying gene editing technologies and large-fragment DNA writing technologies for eukaryotic genomes based on new concepts, principles, or systems. Enable precise editing of any transcript or RNA base.

Challenges: There are still many unclear aspects regarding the performance of current gene (base) editing technologies and their biological foundations. The understanding of PAM sequence specificity, DNA off-target mechanisms, and on-target activities is not yet systematic and in-depth enough, and a systematic solution has not

been formed. The gene editing and its derivative technologies established in model organisms cannot be used or have low efficiency of use in many important non-model organisms. The understanding of DNA targeted recognition and cleavage, the influence of DNA supercoiling, as well as the double-stranded DNA break repair pathways and repair mechanisms is not profound enough. The ability to manipulate double-stranded DNA break repair in non-model cells still needs to be improved. There is a lack of quantitative understanding of transcriptional regulation, epigenetic mechanisms, and cross-regulatory interactions. Due to the presence of many repetitive sequences, the co-expression of multiple crRNAs or sgRNAs within the array can trigger genetic instability, and there are difficulties in the persistent regulation of multiple genes. There are deficiencies in the methodology for detecting the levels of gene editing and regulation at the whole-genome and single-cell resolution. Currently, CRISPR-Cas9, base editors, and prime editors are relatively large in size, making it difficult to achieve efficient delivery through conventional delivery systems such as AAV. The current delivery methods have low cell type specificity and low *in vivo* editing efficiency.

How to efficiently discover novel nucleic acid targeting elements and editing and processing elements with the potential for gene editing development is a technical bottleneck restricting the innovation of underlying gene editing technologies for new concepts, new principles or new systems. The working mechanism and programmability of potential editing elements remain unclear, which is a bottleneck restricting their systematic optimization and functional improvement. There are numerous systems that naturally possess a large capacity for gene fragment integration and editing, such as transposons and recombinases, which are widely distributed in the genomes of different species. However, the relevant basic research is weak, and there is a serious lack of related technology development work. There is insufficient ability to discover the sources of highly efficient gene large fragment writing elements and conduct optimized design, and there is insufficient breadth, depth and interdisciplinary integration promotion in basic research in related fields. At present, the molecular mechanism of the non-targeted RNA cleavage activity (collateral cleavage activity) inherent in Cas13 itself is still unclear; the efficiency and precision of using dCas13-ADAR proteins for *in vivo* RNA base editing need to be improved; the efficiency of using endogenous ADAR deaminases for RNA base editing also needs to be improved, and it has the limitation of only being able to perform A to I base editing; the types of editable motifs (such as GA motif) of RNA base editors still need to be broadened. The effectiveness of RNA editing in higher

animals such as primates needs to be further evaluated and improved, and strict safety assurance is required.

Expected Progress Recently: It is possible to achieve editing of any site in the genome, any gene or any specific single base pair in model organisms, and there is no detectable off-target effect; reveal the mechanisms that affect the efficiency of gene editing, and significantly improve the editing efficiency in the whole genome of important model organisms (exceeding 90%). Achieve persistent gene repression and activation, and specifically regulate the expression of genes in organisms. Improve existing gene editors or delivery systems to better adapt to intracellular delivery. Initially establish the underlying gene editing technologies based on new concepts, new principles or new systems, as well as the underlying technologies for large fragment replacement and writing in eukaryotic genomes based on new concepts, new principles or new systems. It is possible to achieve editing of any site in the transcriptome and any specific base in important model organisms or laboratory animals, and there are no detectable collateral cleavage editing effects and off-target effects.

Expected Progress by 2030: Enable editing of any gene or specific base pair in key non-model organisms (animals, plants, fungi, industrial/environmental microbes) with undetectable off-target effects. Achieve high-efficiency editing (>60%) in non-model organisms or cells. Implement long-term, multi-gene expression control across tissues and organisms. Enhance delivery specificity for efficient and targeted editing in specific tissues. Optimize novel concept/principle-driven gene editing systems to surpass CRISPR-Cas in specific performance metrics. Enable efficient and precise writing of large DNA fragments (>10 kb) in eukaryotic cells. Achieve transcript- and single-base editing in specific tissues of non-human primates with no detectable off-target effects, and initiate clinical trials for RNA editing therapies.

Potential Solutions

Screen to obtain or design and improve deaminases so that they can catalyze all possible nucleotide conversions. Improve prime editing tools to enhance their base editing efficiency. Design engineered nucleases and recombinases to achieve all possible base editing by controlling the repair of double-strand breaks. Improve gene editors to increase the on-target rate and reduce off-target effects. Optimize gene editors or editing strategies for the host DNA repair pathway to achieve minimal off-target effects.

Establish a set of high-fidelity gene editors that can specifically target the vast majority of sites in the genome to expand the coverage of editing sites.

Utilize the endogenous CRISPR-Cas systems of prokaryotic microorganisms to establish gene editing tools suitable for them, enabling genome editing and gene regulation of a wide variety of microorganisms. Continuously improve the gene editing tools and delivery vectors for important non-model economic plants, animals, microorganisms in industry, agriculture, medicine and the environment, and macrofungi, etc. Continuously optimize the editing tools for human and animal gene therapy, and while ensuring no detectable off-target effects, it is also necessary to overcome possible immune responses.

Fuse epigenetic effector proteins with Cas9 to reveal the influence of the higher-order chromatin structure on gene editing. Quantitatively and predictively understand the impact of the coupling of chromatin structure, DNA targeted cleavage, and repair on gene editing efficiency. Manipulate and quantitatively test the influence of the regulation of chromatin structure and the regulation of DNA repair efficiency on improving the efficiency of different gene editors. Manipulate and quantitatively test the influence of different combination methods, delivery methods, and expression methods of gene editors on gene editing efficiency. Comprehensively utilize the above factors that affect gene editing efficiency to achieve a significant improvement in gene editing efficiency among important model organism groups.

Use the editor with the highest insertion and deletion efficiency in model organisms. Establish an induction control system for the DNA break repair pathway with significantly high efficiency. Improve the nuclear-targeted delivery of DNA repair templates in model and non-model organisms. While delivering the gene editor, provide a reverse-transcribed RNA template coupled with sgRNA to increase the concentration of the repair template at the gene editing site. Develop more efficient delivery tools. Develop a specific delivery system for specific tissue cells to achieve gene editing of specific tissue cells *in vivo*.

Design genome editors that fuse different effectors and conduct quantitative characterization to obtain more effective CRISPRa (activation) and CRISPRi (inhibition) systems. Change the position, binding strength, and interaction mechanism of the RNP binding sites, etc., and conduct a systematic analysis of the gene regulation effects. Design transcription factors that can reliably generate the desired epigenetic traits. In the desired organism, express CRISPR components through tissue-specific promoters, or

effectively activate or inhibit genes through tissue-specific protein fusion domains.

It is possible to design a highly non-repetitive CRISPR toolbox, thereby designing multiple stable sgRNA arrays. Due to the different efficiencies of gene regulation through different sgRNA structures, design and implement a quantitative regulation network. Optimize single-cell gene editing, epigenetic modification (modifications such as histones, lncRNAs, nucleosomes, etc.), protein level, and metabolite level measurement technologies, and measure the efficiency of multi-gene editing or regulation; effectively and specifically deliver gene editing tools to target cells in specific tissues and specific communities, and be able to control their dosage and editing time.

Improve existing high-fidelity gene editors to reduce their size while maintaining their specificity. Develop smaller editors suitable for packaging and delivery by adeno-associated virus (AAV) and other vectors based on new compact CRISPR-Cas systems. Develop non-viral packaging and delivery systems that can directly deliver *in vivo* and maintain the activity of gene editing effector complexes (such as ribonucleoproteins, RNPs). Develop safe and effective large-capacity viral vector delivery systems (with a capacity of 10 kb or more) to simultaneously deliver editor DNA, single-guide RNA (sgRNA), and donor DNA molecules. Further optimize the RNA editing technology based on endogenous editing enzymes to improve the editing efficiency and expand the editing scope. Since there is no need to introduce exogenous editing enzymes or effector proteins, it avoids problems such as delivery and related immunogenicity caused thereby.

Develop and optimize the tropism/specificity of viral delivery systems on a large scale. Enhance their specificity through cell type-specific design of receptor interactions or other forms. Develop viral delivery technologies with a long half-life and low immunogenicity (this may be required for each organism of interest, including plant viruses and animal viruses). Develop reliable analytical methods for potential off-target sites, or regulate the activity of editors relying on cell types, small molecule regulation, etc., to enhance specificity and reduce or eliminate off-target effects.

Deeply mine the novel nucleic acid immune systems, transposons, repetitive sequences, and other new systems by utilizing large genomic databases, analyze the new principles of nucleic acid sequence targeted recognition, and develop original novel gene editing elements, including nucleic acid targeting elements, nucleic acid modification and processing elements, etc. Analyze the structure and functional mechanisms of novel nucleic acid modification and processing elements, reveal their programmability

principles, and establish new principles and mechanisms based on gene editing. Combine programmable novel nucleic acid targeting elements and processing elements to establish gene editing technologies with new concepts, principles, or systems, including but not limited to: the next-generation gene editing technology directly guided by proteins, the next-generation gene editing technology consisting only of nucleic acid components, a completely new CRISPR gene editing system that is free from patent restrictions, the underlying gene editing technology with independent intellectual property rights. Achieve gene editing in model animals, plants, and microbial cells using at least one novel concept gene editing technology.

By comprehensively integrating artificial intelligence analysis of large genomic data, etc., screen and obtain novel concept gene editing elements that exhibit advantages over existing CRISPR-Cas systems in terms of size. Make new breakthroughs in gene editing elements and gene editing theories. For example, by modifying natural RNA nucleic acid molecules or *de novo* synthesis and other means, develop ribozymes that do not rely on proteins and can target and cleave DNA and RNA, and develop novel RNA-based gene editing technologies. Based on the underlying technology of novel concept gene editing, use artificial intelligence deep learning and rational design to optimize its effectiveness and specificity, and establish a series of derivative core technologies, such as extended editing technologies like base editing, epigenetic editing, and RNA editing. Achieve gene editing in important model or non-model animals, plants, and microorganisms using the underlying technology of novel concept gene editing and its derivative technologies, reaching or approaching the editing efficiency of CRISPR-Cas9.

For different types of mobile genetic elements, such as recombinases, transposases, integrases, etc., further strengthen basic research, analyze their core molecular mechanisms, carry out functional modification and test on them to determine the functions that different types of elements can achieve in gene replacement or writing applications. For genomic databases, conduct in-depth annotation, evolutionary analysis, and functional prediction of mobile genetic elements, and establish a high-throughput functional screening and evaluation platform according to the functional characteristics of different types of elements to screen and obtain novel functional elements with the ability of genetic modification. By analyzing the molecular mechanism of large-fragment gene integration mediated by mobile elements, design modification plans, combine existing gene editing technologies with gene integration elements, and initially obtain the underlying technology for the replacement and writing of small fragments (>5 kb) in

eukaryotic genomic DNA.

Carry out in-depth annotation and mining of the continuously increasing novel genomics data (such as metagenomic data), continuously discover novel mobile genetic elements, and use artificial intelligence algorithms such as deep learning and high-throughput functional evaluation to reveal their functions and mechanisms, providing core elements for the underlying technology of brand-new gene writing. Based on the analysis and prediction of a large number of protein structures, use artificial intelligence deep learning and rational design to carry out completely new design and modification of newly mined mobile element proteins and important functional domains, and carry out design optimization for specific technical indicators, such as the size of the integrated fragment, fidelity, specificity, etc., to establish the underlying technology of innovative gene writing with excellent performance at the source. Based on the understanding of the mechanisms and pattern summarization of the gene integration functional elements existing in nature, establish the technical ability to design brand-new proteins at the source, generate protein structures and action mechanisms that have never appeared in natural evolution, and establish the source-innovative writing technology for large fragments (>10 kb).

Conduct high-throughput mutation screening of each functional domain of the Cas13 protein, and in combination with the crystal structure model of the Cas13 protein, determine the key amino acid sites related to the collateral cleavage activity. Optimize and combine different elements in existing RNA base editors to improve the efficiency of RNA base editing. Optimize the ADAR deaminase or other elements to reduce the off-target effects near the target sites and at the transcriptome level of the RNA base editors. Optimize the chemical modification of the guide RNA that binds to the endogenous ADAR enzyme to improve the efficiency of endogenous RNA base editing. Discover new ADAR proteins through bioinformatics or optimize the existing ADAR proteins to broaden the types of editable motifs of the RNA base editors.

Continuously track the latest microbial metagenomic databases to screen novel RNA editing proteins with smaller sizes. Through protein engineering modification and functional verification, obtain more efficient and precise compact RNA editing tools, enabling effective *in vivo* editing with a low-dose administration, and allowing for multi-site editing with a single administration to reduce the off-target risk potentially caused by the overexpression of the editing tools. Continuously optimize the delivery system of the RNA editing tools to obtain highly efficient tissue-specific delivery vectors.

The improvement of the delivery efficiency can further reduce the dosage of the editing tools, thereby reducing the off-target risk and the immunogenicity of the delivery vectors.

3.2.5 Summary

The new generation of gene editing technology will further achieve precision, specificity, miniaturization, and broad adaptability, break through technical bottlenecks such as the efficient writing of large DNA fragments and the efficient delivery of editing systems, and enable precise editing of any gene, fragment, or base in any important species and tissues. To innovate the new generation of gene editing technology, on the one hand, the existing gene or base editing technology based on CRISPR-Cas should be further optimized to enhance its specificity, precision, and safety. On the other hand, new systems, principles, and strategies for gene editing should also be innovated, and new generation underlying technologies for DNA and RNA editing should be developed to continuously promote the development of this field.

References

[1] Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, 2012, 337(6096): 816-821.

[2] Cong L, Ran F A, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science, 2013, 339(6121): 819-823.

[3] Wang H, Yang H, Shivalila C S, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell, 2013, 153(4): 910-918.

[4] Yang H, Gao P, Rajashankar K R, et al. PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease. Cell. 2016, 167(7): 1814-1828.

[5] Schuler G, Hu C, Ke A. Structural basis for RNA-guided DNA cleavage by IscB-omega RNA and mechanistic comparison with Cas9. Science, 2022:eabg7220.

[6] Knott G J, Doudna J A. CRISPR-Cas guides the future of genetic engineering. Science, 2018, 361(6405):866-869.

[7] Komor A C, Kim Y B, Packer M S, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature, 2016, 533: 420-424.

[8] Gaudelli N M, Komor A C, Rees H A, et al. Programmablebase editing of A·T to GC in genomic DNA without DNA cleavage. Nature, 2017, 551: 464-471.

[9] Rees H A, Liu D R. Base editing: precision chemistry on the genome and transcriptome of living cells. Nat Rev Genet, 2018, 19:770-788.

[10] Abudayyeh O O, Gootenberg J S, Franklin B,et al. A cytosine deaminase for programmable single-base

RNA editing. Science, 2019, 365(6451): 382-386.

[11] Anzalone A V, Randolph P B,Davis J R, et al Search-and-replace genome editing without double-strand breaks or donor DNA. Nature, 2019, 576:149-157.

[12] Zong Y, Liu Y, Xue C, et al. An engineered prime editor with enhanced editing efficiency in plants.Nat Biotechnol, 2022, 40: 1394-1402.

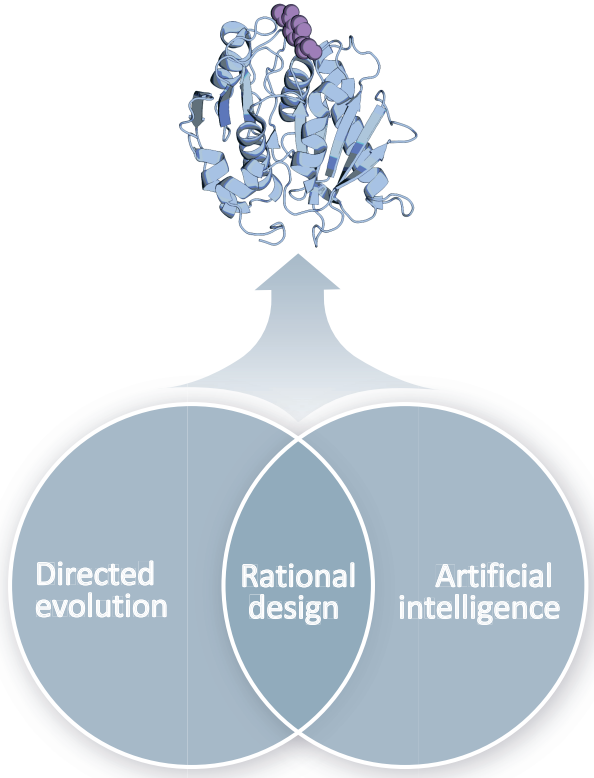
[13] Anzalone A V, Koblan L W, Liu D R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nat Biotechnol, 2020, 38(7): 824-844.

[14] Lampe G D, King R T, Halpin-Healy T S, et al. Targeted DNA integration in human cells without double-strand breaks using CRISPR-associated transposases. Nat Biotechnol, 2024, 42: 87-98.

[15] Wang J Y, Pausch P, Doudna J A. Structural biology of CRISPRCas immunity and genome editing enzymes. Nat Rev Microbiol, 2022, 20: 641-656.

[16] Kreitz J, Friedrich M J, Guru A, et al. Programmable protein delivery with a bacterial contractile injection system. Nature, 2023, 616:357-364.

Protein Design



Authors Feng Yan, Wu Bian, Sun Zhou-Tong, Yang Guang-Yu, Wang Xiang-Xi, Qu Ge, Wang Ya-Jie

3.3 Protein Design

3.3.1 Abstract

Proteins are the primary executors of biological functions and the foundational building blocks of biological systems. Designing versatile, functionally unique protein components is a key focus in synthetic biology. Protein design, as the core technology for creating high-quality protein components, integrates data-driven computational methods, structure-guided rational design, and directed evolution based on efficient screening to create novel functional proteins. Despite these advancements, challenges persist, including limited standardized protein data, data bias, incomplete understanding of structure-function relationships and regulatory mechanisms, and the lack of universal and efficient screening methods. Addressing these challenges requires developing high-quality protein databases, accurately elucidating structure-function relationships, and establishing efficient design and screening workflows, which are crucial steps to meet the demands of biological system engineering.

3.3.2 Technical Overview

3.3.2.1 Data-driven Computational Design

Protein-protein interactions play critical roles in nearly all biological processes. As high-quality protein component databases continue to be refined, data-driven computational design methods are poised to revolutionize protein component design. Future automated design platforms will generate vast amounts of standardized functional experimental data. Leveraging breakthroughs in structure prediction achieved by AlphaFold2, “white-box” models based on force field functions can deeply elucidate the reaction mechanisms and interaction relationships of various regulatory and catalytic components. Furthermore, artificial intelligence can be employed to decipher the fundamental “sequence-structure-function-interaction” mapping and underlying laws in proteins. By optimizing force field functions and AI algorithms, integrating “black-box” and “white-box” modeling approaches, and developing innovative computational design algorithms of protein components, we can construct next-generation intelligent design platforms. These systems will enable precise, multidimensional protein design and

optimization for tailored applications, thereby expanding their application scenarios.

3.3.2.2 Structure-driven Rational Design

Rational design of protein components primarily relies on structure-function relationships and protein-ligand interactions. By leveraging bioinformatics and three-dimensional structural data, researchers can create optimized mutant libraries for targeted modifications. This involves elucidating sequence-function mapping mechanisms, performing allosteric analysis of key active sites and regulatory sites in proteins, and decsiphering the molecular basis of the structure-activity relationship and ligand-selectivity in protein components. Developing novel methods and strategies enables computational simulations to identify key functional sites that influence performance and regulation. Additionally, eliminating redundant amino acid residues based on functional mechanisms can reduce mutant library size, providing a solid foundation for developing high-performance bioparts.

3.3.2.3 Ultra-high-efficient Screening-driven Directed Evolution

Directed evolution is a powerful, non-rational approach for designing functional and regulatory protein components, rapidly enhancing the functionality of bioparts by mimicking natural evolution through random mutation and high-throughput screening. However, conventional screening methods are limited to library sizes of approximately 10^3 to 10^5, significantly restricting evolutionary efficiency. Recent advances in ultra-high-throughput screening technologies, such as flow cytometry, droplet microfluidics, and *in vivo* strategies like growth-coupled and phage-assisted continuous evolution, enable screening libraries exceeding 10^7 variants. Developing innovative high-throughput, high-sensitivity protein component screening platform can expand screening library capacity to explore a broader protein sequence space, facilitating the generation of highly efficient functional components for industrial applications and advancing our understanding of ligand recognition and bioparts' molecular evolution.

3.3.3 Roadmaps

Current Status		
A number of databases with predictable, controllable, and assembled bioparts information, alongside core computational tools for protein design have emerged.		
Objective 1: Construct High-Quality, Standardized Protein Component Databases Using Next-Generation Technologies		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop a protein component database containing over 10 million elements, facilitating high-throughput, automated assembly of synthetic protein components, with a unit assembly capacity of 10^4 per month.	**Establish an open, shared, high-throughput, automated, and digital innovative platform for protein component.** • Develop a database and cloud platform to improve protein query efficiency. • Organize and classify protein components using bioinformatics, identifying characteristic sequences responsible for catalytic and regulatory functions. • Implement a combined strategy of plasmid preservation and DNA storage to enhance the automation of component preservation and extraction.	**Establish an automated, high-throughput system for component assembly and testing, completing the standardized assembly and testing of over 1 million protein components.** • Standardize protein component information and develop platforms for standardized and automated synthesis and assembly. • Develop assembly methods and technologies compatible with automated systems for standardized protein components. • Innovate gene synthesis technologies and construct super-heterogeneous expression hosts to achieve efficient expression of functional protein components.

Objective 2: Mine and Analyze Protein Databases to Understand Evolutionary Relationships within Protein Sequence Space and Develop Quantitative Theories for Component Structure, Function, and Molecular Evolution		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop a core technology system for protein component design, study the correspondence of evolution-sequence-structure-function, and enable rapid creation of high-performance proteins.	**Develop data-driven intelligent protein design technologies.** • Innovate technologies for large-scale protein data mining and refined evolutionary-sequence correspondence models. • Study constraints on protein evolvability and established a research framework for protein interactions.	**Develop *de novo* design technologies for full protein sequences, functional fragments, and key sites at different levels.** • Innovate advanced algorithms, such as deep learning, to enhance protein data processing capabilities. • Achieve a leap from natural protein sequences (input) to artificial protein sequences (output), develop a new method for protein design.

Figure 1 Roadmap for high-quality protein component characterization, data analysis, and rational design

Current Status		
Physical model-based computational technologies have seen several successful applications in protein component design. Nevertheless, data-driven design approaches are still at a nascent stage.		
Objective 1: Develop Novel Algorithms for Data-Driven Protein Component Design		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Construct high-quality, standardized libraires of rigid protein backbone and catalytic component characterization databases to serve as foundational resources for data-driven design algorithms targeting protein interaction regulators and catalytic components.	**Establish a standardized experimental characterization database for protein components and developed backbone design platforms.** • Improve workflows for data collection and cleaning. • Optimize the design of three basic backbone structures to facilitate the construction of rigid modular backbones. • Utilize high-throughput sequencing and screening technologies to generate high-quality, standardized experimental characterization data for protein components.	**Design backbones for large protein complexes with numerous subunits and develop innovative technologies for data-driven protein design.** • Develop generalized backbone modular assembly technologies, including deep learning and other advanced machine learning techniques. • Extract intrinsic properties embedded within large biological process datasets.

Objective 2: Decode Fundamental Principles from Protein Component Data and Develop Innovative Computational Design Theories and Algorithms		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Advance the integration of machine learning with "white box" models based on force field functions.	**Decode the "sequence-structure-function" relationships of protein components.** • Investigate universal patterns in sequence-structure-function relationships. • For proteins with specific functions that have been characterized, directly construct standardized module interfaces. • For proteins with unclear mechanisms but known structures, use "hallucination" techniques to transplant functional regions and perform standardized interface detection tasks.	**Integrate "white-box" and "black-box" models to achieve standardized interface assembly for mega functional proteins.** • Decompose functional proteins into multiple subfunctions or substructures; Assemble complete protein functions from modular components. • Quantify the performance of each layer of the model. • Test and validate using "white-box" models, systematically refining black-box model parameters based on feedback.

Objective 3: Establish Spatial Structural Design-Specific Multi-Module Standardized Interfaces for Precise Multidimensional Protein Component Design and Expanded Application Scenarios		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve the design of multiple independent protein interfaces.	**Establish a backbone interface design platform to develop foundational functional logic components.** • Enhance interface adaptability testing by considering adaptability as a key attribute during module selection. • Optimize interfaces through a combination of computational algorithms and biochemical experiments.	**Use pre-designed interfaces to assemble mega complexes.** • Design interface locks to stabilize complexes with fixed structures. • Incorporate interface lock proteins before adding proteins with repetitive interfaces, locking already assembled proteins before adding proteins with repetitive interfaces. • Enable the reuse of interfaces during the assembly of mega complexes.
Achieve intelligent, robust, and efficient customized artificial protein component design.	**Provide design strategies for protein components tailored to different application scenarios.** • Combine computational algorithms with biochemical experimental testing. • Establish a cyclical optimization framework integrating intelligent algorithms and biochemical testing.	**Enable intelligent protein component design for specific needs.** • Fully consider the complex, multi-level regulatory characteristics of biological systems. • Implement interventions at different regulatory levels. • Optimize the overall compatibility of artificial protein components with their chassis environments.

Figure 2 Roadmap for protein component design technologies based on machine learning, physical models, and other computational techniques

Current Status
In vitro screening systems are advancing through label-free mass spectrometry detection technologies and *in vitro* transcription-translation systems compatible with toxic proteins, with continuous breakthroughs in applicable systems. *In vivo* screening has been expanded primarily through the application of diverse enzymatic systems and the development of new intracellular continuous directed evolution systems. Nevertheless, developing efficient and sustainable directed evolution tools for eukaryotic organisms remains an unmet challenge.

Objective 1: Develop Universal and Efficient Screening Technologies		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve breakthroughs in novel protein component screening methods that accommodate diverse functions, with a focus on extensively developing label-free screening techniques utilizing natural substrates.	• Develop universal screening technologies capable of processing 10 million clones daily. • Utilize innovative fluorescent probes, biosensors, and fluorescence coupling methods to expand the application scope and improve the performance of fluorescence detection. • Create new mass spectrometry devices and sample pretreatment systems tailored for high-throughput protein component screening.	• Develop universal screening technologies capable of processing hundreds of millions of clones per day. • Innovate higher-throughput flow cytometers and droplet-based screening devices. • Create automated droplet-based mass spectrometry detection systems.

Objective 2: Develop High-Throughput, Automated *In Vivo* Protein Component Evolution Technologies		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop tools for targeted, large-window (greater than 1 kb), high-mutation-rate continuous directed evolution, compatible with various chassis cells.	**Enhance targeting precision of intracellular mutations, expand editing windows, and increase mutation rates.** • Broaden nucleases libraries to improve targeting versatility and sequence recognition coverage. • Develop novel polymerase-deaminase fusion proteins with high specificity, reducing non-target mutations to below 10^{-8}. • Create efficient *in vivo* continuous directed evolution technologies specifically for eukaryotic organisms. • Integrate *in vivo* directed evolution technologies with high-throughput screening to improve experimental throughput.	**Develop novel *in vivo* directed evolution strategies to improve evolutionary efficiency and throughput.** • Create advanced gene-editing technologies with enahnced specificity and broader editing windows. • Develop high-throughput platforms that combine continuous culture and screening, enabling simultaneous mutation and screening—realizing "continuous directed evolution for most intracellular mutation tools." • Achieve parallel testing of thousands of target proteins, with hundreds of rounds of mutation and screening per day.

Figure 3 Roadmap for the development of universal directed evolution and high-efficiency screening technologies

3.3.4 Technical Pathways

3.3.4.1 Characterization of High-quality Protein Components, Data Analysis, and Rational Design

Current Technologies: Currently widely used databases include the NCBI database (containing approximately 1.4 trillion base pairs as of June 2022), UniProt database (with 568,002 protein annotations as of August 2022), and PDB database (containing 194,011 biomacromolecular 3D structures as of July 2022). With advances in synthetic biology, new-generation databases—such as BioBrick and BioFab—have emerged, integrating predictable, regulatable, and assembly-ready bioparts. Establishing open-access international subcenters for protein components featuring large capacity, automation, and digitalization, along with improved collaborative data-sharing mechanisms, is essential for foundational research into protein function.

Protein sequences determine their three-dimensional structures, which in turn govern cellular functions. The rapid development of computational technologies, particularly artificial intelligence, has revolutionized protein engineering. Notably, on July 31, 2022, DeepMind announced that its AI program AlphaFold2 had predicted over 200 million protein structures across approximately 1 million species, covering nearly all cataloged proteins. By integrating mainstream tools like AlphaFold and RoseTTAFold and developing novel underlying systems to study evolution-sequence-structure-function relationships, we can accelerate the rapid design and creation of high-performance proteins.

Objectives and Breakthroughs: Develop high-quality, standardized protein component databases leveraging next-generation technologies such as high-throughput sequencing, deep mutational scanning, microfluidics, and AI. Conduct comprehensive database mining to elucidate evolutionary relationships within protein sequence space and establish quantitative theories for structural, functional, and molecular evolution.

Challenges: The enormous volume of data presents storage difficulties. Redundant data must be eliminated to enhance data quality, and protein characterization data require standardization. Currently, data collected from different laboratories lack uniform standards, resulting in high synthesis cost. Additionally, accurately uncovering the underlying evolutionary principles within vast sequence spaces remains difficult. Precise modeling of the relationship between protein evolvability and function is challenging.

Extracting the organization and spatial arrangement rules of protein functional modules is complex. Limited computational power hampers sequence processing capabilities, and the inherent complexity of the protein sequence space results in insufficient prediction accuracy.

Expected Progress Recently: Establish an open, shared innovative platform for protein components featuring high-throughput, automation, and digitization capabilities. Develop technical systems for high-throughput, automated, and digital mining, design, construction, testing, analysis, and modeling of protein components. Build a comprehensive protein component database containing over ten million entries to support most biological design needs. Advance data-driven intelligent protein shaping technologies by mining and analyzing the database, predicting protein structures, and uncovering evolution-sequence-structure-function relationships. Break through existing sequence-structure theoretical frameworks by studying the direct, end-to-end mapping between protein sequences and functions and establishing high-precision potential energy models.

Expected Progress by 2030: Build automated high-throughput systems for the assembly and testing of over one million standardized protein components. Utilize advanced CPU/GPU computing and deep learning to develop hierarchical *de novo* design technologies of protein complete sequences, functional fragments and key sites. Implement neural networks capable of hierarchical abstracting multi-dimensional amino acid sequence space to learn the evolutionary relationships of protein sequences, enabling a single-step transformation from natural protein sequences (informational input) to functional artificial sequences (functional output).

Potential Solutions

Build a cloud-based protein component database utilizing distributed processing, distributed databases, cloud storage, and virtualization technologies to design a protein component database and cloud platform to enhance query efficiency. Apply bioinformatics techniques such as structure-function-based sequence alignment and evolutionary analysis to classify and organize elements from diverse sources and families or share the same or similar functions. Use redundancy removal methods to identify representative functional elements within large isozyme datasets. Conduct standardized functional characterization of protein components tailored to different chassis

microorganisms and applications.

Establish a unified standard for protein components and develop a standardized, automated synthesis and assembly platform. Create a high-throughput assembly system capable of producing up to 10^4 protein components per month. Innovate gene synthesis technologies to reduce costs and construct advanced heterologous expression host for efficient functional protein production.

Develop advanced data mining tools to explore functional relationships within large protein datasets, especially among distant evolutionary relatives. Establish high-precision potential energy functions and models linking evolution and sequence, enabling quantitative analysis of protein evolvability and function. Establish a protein interaction research system to explore interaction networks between proteins, proteins and nucleic acids, or proteins and small molecule substrates, reveal the spatial arrangement patterns and collaborative mechanisms of amino acid residues responsible for functional execution in proteins, extract their intrinsic mechanisms for functional realization, and form a protein functional module dataset.

Implement deep learning algorithms to enhance the data processing capabilities and reduce computational resource requirements. Perform self-training of predictive models through experimental verification and classification refinement, continuously improving the prediction accuracy. Establish new theoretical frameworks for protein design, increasing the success rate and scalability of functional protein engineering.

3.3.4.2 Development of Protein Component Design Technologies Based on Machine Learning and Physical Models

Current Technologies: Biological systems are inherently complex, and current challenges include inconsistent data characterization methods and a lack of large-scale specialized datasets. Additionally, the long-standing academic practice of omitting negative data has skewed the statistical understanding of protein interaction regulators and catalytic components, resulting in an inaccurate statistical picture, limiting their utility for large-scale model training. To address data scarcity, machine learning pretraining methods leverage unannotated high-throughput sequencing data to map protein adaptive landscapes or employ low-parameter models to infer evolutionary information from homologous sequences, enabling high-accuracy predictions of pathogenic mutations [1]. Concurrently, protein-protein interaction pairs are designed and validated in large-scale experiments, with affinity data for interaction elements obtained

through high-throughput experimental validation. By inputting experimentally measured data and theoretical computational models into the designed deep neural network model, to learn sequence-structure-affinity relationship, enabling more accurate predictions of how changes in protein sequences affect their functions.

Structure determines function. Synthetic biology now enables the integration of enzymes associated with biochemical reaction nodes from diverse pathways to construct *in vivo* systems for decomposing and coupling complex functionalities. To advance protein design, tools rooted in physicochemical principles—such as Rosetta [2] and ABACUS [3]—have been developed, which have established a series of computational strategies, demonstrating broad applicability. For example, the University of Washington's David Baker team used Rosetta to design *de novo* proteins to function as molecular switches for pH-triggered conformational changes [4,5], dual-input logic gates [6], and, more recently, complex protein nanomaterials with tailored systemic properties via top-down reinforcement learning[7]. Their cryo-EM structures closely matched computational models, and designed icosahedral capsid nanoparticles exhibited bioactivity, enhancing vaccine responses and inducing angiogenesis.

Designing scaffolds and functional elements requires standardized interfaces for effective coupling. The design of protein interfaces depends on the design of the binding interface. Currently, the design of interaction interfaces has entered an era of rapid development. For example, the mini-binder design based on RIF-DOCK [8], RFDesign based on deep learning [9], etc., can all achieve the design and optimization of the interaction interfaces between proteins. In addition, Rosetta Antibody Design [10] based on the replacement of antibody scaffolds and variable regions can be classified as the design of interactions between proteins. Current algorithms focus on modeling the explicit or implicit sequence-structure relationship and intra-molecular interactions, with some exploring direct generation of functional proteins. Since protein functions often involve interactions with other molecules (e.g., small metabolites, nucleic acids), extensive biochemical experiments are necessary for optimization. Advances in deep learning have led to breakthroughs in predicting properties that are difficult to model physically, such as the solubility prediction of heterologous proteins in *E. coli* and the prediction of the immunogenicity of chimeric antibodies. When artificially modifying metabolic pathways, the performance of protein components predicted by deep learning has also been used to optimize the modeling of complete metabolic pathways [11].

Objectives and Breakthroughs: Develop data-driven protein design algorithms.

Build standardized rigid backbone libraries and catalytic component databases. Uncover underlying principles within protein data to advance computational design theories and algorithms. Integrate machine learning techniques with force field function-based "white-box" models for enhanced accuracy. Establish standardized interfaces featuring specific multi-modules for spatial structure design to achieve precise design of protein components in a multi-dimensional space, design of multiple independent protein interfaces, as well as intelligent, robust and efficient customized design of artificial protein components.

Challenges: Current challenges include the lack of standardized characterization methods across laboratories, hindering large-scale model training. Proteins are inherently flexible, and as the number of subunits increases, inaccuracies in scaffold length and spatial angles can lead to linear amplification of deviations in the overall complex structure. Preventing such amplified errors remains a significant challenge. Additionally, the scarcity of comprehensive biological databases and inherent biases contribute to overfitting and underfitting in machine learning model training.

The functional mechanisms of most protein components remain poorly understood. There is a lack of confidence models for key protein regions that influence enzymatic activity, and machine learning prediction models often lack generalizability across different tasks. Moreover, assembling proteins with complete functions within a standard unit volume of space is currently unfeasible. Models based on "black-box" approaches have weak interpretability and fail to establish effective links with "white-box" models that explain structure-function relationships.

Artificially designed interfaces can significantly impact protein function. When designing protein interfaces, it is necessary to fully consider the trade-off between the proportion of interface functions and the standardization plasticity of interfaces. When piecing together multiple different modules, the specificity of the interface will determine the upper limit of the logical complexity of module assembly. However, when the number of subunits forming the complex exceeds the existing number of interfaces, there will be a situation where no interfaces are available. The design efficiency of artificial protein components is relatively low, and it is necessary to consider the mutual influence among multi-level regulations in the artificial design pathway.

Expected Progress Recently: Establish standardized experimental databases and backbone design platforms. Optimize fundamental scaffold structure to achieve the construction of the basic rigid module scaffold, providing essential information for the

digital-driven design of protein components. Analyze the "sequence-structure-function" relationship of protein components; build an interface design platform for scaffold/scaffold and scaffold/functional elements and lay the foundation for functional logic assembly elements to achieve specific docking of different unit elements; provide design strategies for protein components required in different application scenarios.

Expected Progress by 2030: Leverage the experimental characterization database of standardized protein components and basic scaffold elements to design the scaffold of giant protein complexes with a large number of subunits. Develop innovative data-driven protein design technologies that integrate "white-box" model and "black-box" model to achieve standardized interface assembly of large-scale functional protein units. Use the designed interfaces to realize the assembly design of super-large complexes. The intelligent design of protein components serves the major strategic needs of the country.

Potential Solutions

Refine data collection and cleaning workflows by maximizing the use of raw literature for data verification and correction. Adopt structured, machine-readable data formats to construct distinct protein component characterization datasets tailored to specific properties, mitigating data inconsistency and large-scale missing. Continuously accumulate high-quality, standardized protein characterization data through experimental techniques such as high-throughput sequencing and screening.

Design universal assembly backbone modules using standardized protein skeleton units for repetitive stacking. For instance, to address angle measurement challenges, construct interfaces that enable backbone proteins to assemble into helical coil structures. Amplifying such structures allows logical signal detection, thereby resolving measurement difficulties. Leverage standardized protein component databases to develop novel machine learning methods that extract intrinsic ontological features embedded in big biological data.

For proteins with specific functions that have currently been analyzed and studied thoroughly, directly construct standardized module interfaces; for proteins whose action mechanisms are unclear but whose structures have been analyzed, transplant the known functional regions through methods such as hallucination, and establish standardized interface detection tasks. Deeply explore the universal laws of the sequence-structure-function relationship of protein components based on the force field

function. Use artificial intelligence models such as the transfer learning model to generate and extract deep features from the input protein component sequences, so as to perform a variety of prediction tasks based on the sequences.

The minimization design of functional proteins refers to using the least amount of amino acid sequences to perform the functions of a complete protein. A functional protein can be disassembled into multiple sub-functions or sub-structure units, and multiple modules are used to achieve the assembly of the complete protein function. Develop and verify the model systematically, quantify the performance related to each layer of the model, and test and verify it with a "white-box" model. Adjust the parameters of the black-box model systematically according to the feedback results.

When constructing functional modules, add interface compatibility screening. Consider it as an attribute of the functional module and take it into account when selecting and matching modules. Optimize the interface by combining computational algorithms and biochemical experiments. After designing a large number of interfaces, redesign the interfaces that have the least impact on other proteins to achieve the expansion, improvement, and optimization of the advantageous interfaces.

Carry out interface compatibility design at the original protein interface and design an interface lock to lock the complex that has formed a fixed structure. During the process of assembling proteins, add different assembly elements in the order of the assembly logic of the complex. First, add the interface lock protein. After locking the assembled protein, introduce the repetitive interface elements. At this time, the repetitive utilization of the interface during the assembly process of the super-large complex can be realized.

Combine computational algorithms with biochemical experimental tests to build a cyclic optimization design framework for intelligent algorithms and biochemical tests and improve the design efficiency of proteins.

Account for the multi-layered regulation and tightly coupled complexity of biological systems. Systematically intervene at different regulatory levels to achieve intelligent global compatibility and optimization between artificial protein components and chassis environments.

3.3.4.3 Development of High-Throughput Screening and Directed Evolution Technologies for Universal Protein Components

Current Technologies: The ultra-high-throughput screening method based on droplet microreactors for single-cell systems is a recent innovation in high-throughput

screening technology for detecting target protein components and metabolites. This system achieves efficiencies 3 to 5 orders of magnitude higher than traditional 96-well plate methods, enabling rapid and efficient screening of protein components and metabolites, and significantly improving the breeding efficiency of engineered bacteria. For example, a research team at the University of Cambridge disrupted cells by introducing cell lysis reagents during microdroplet preparation, conducting directed evolution screening of sulfatase with a throughput of approximately 1,000 per second, and successfully increasing the enzyme's activity and expression level sixfold [12]. Similarly, a team at Shanghai Jiao Tong University developed a dual-color fluorescence screening system based on a microfluidic chip. Using this system, they performed direct evolution of the thermophilic esterase AFEST. After multiple rounds of random mutagenesis and site-directed saturation mutagenesis, they obtained mutants with significantly improved activity and stereoselectivity, confirming the effectiveness of the dual-color droplet microfluidic screening system [13]. However, since screening of single-cell microreactors typically requires converting metabolites into detectable fluorescent signals—and most metabolites lack suitable fluorescent coupling strategies, this greatly limits the application of such systems in metabolite screening of engineered bacteria. To overcome this limitation, label-free detection methods such as Raman spectroscopy and mass spectrometry have been introduced into single-cell microreactor systems. Nonetheless, these detection methods are still in developmental stages and are not yet mature enough for widespread practical application.

Intracellular directed evolution—mutating specific genes within cells without the need for cloning or transformation—significantly shortens the experimental cycle of directed evolution. Early methods primarily relied on single/double-stranded DNA recombination techniques such as MAGE [14]. In recent years, the intracellular directed evolution tools developed based on CRISPR-Cas (such as CasPER [15], CHAnGE [16], MAGESTIC [17], CREATE [18], etc.) have greatly improved the efficiency of gene recombination in prokaryotic and eukaryotic cells. The phage-assisted continuous evolution system PACE is one of the most representative cases in recent years. It separates the mutation library from the host, which can avoid the introduction of a large number of background mutations [19]. However, the above methods all have various limitations. Recombination technologies are prone to introducing harmful mutations into the host genome, resulting in genetic instability. CRISPR-based systems offer a limited mutation window and a high off-target rate, resulting in a high level of background

mutations. The PACE system needs to rely on the relatively complex "lagoon" reactor technology, cannot act on proteins that cannot be coupled with the expression of pIII protein, and is also difficult to be realized in hosts other than *Escherichia coli*. These factors limit their widespread application. Therefore, key challenges in continuous *in vivo* directed evolution include improving targeting mutation accuracy, designing appropriately sized mutation windows, and achieving precise recognition and termination of target genes to prevent interference with downstream genetic elements.

Objectives and Breakthroughs: Develop novel *in vitro* high-throughput protein screening technologies. Advance automated, high-throughput *in vivo* protein evolution systems.

Challenges: The lack of effective fluorescent coupling methods for most protein components limits the detection sensitivity and application scope of microdroplet screening systems. Additionally, current high-throughput mass spectrometry techniques have relatively low automation levels and high costs, restricting their use for large-scale screening of numerous protein components. Throughput and speed bottlenecks in microfluidic screening devices further limit daily processing capacity beyond 10 million samples. Moreover, the destructive nature of mass spectrometry detection makes it incompatible with the non-destructive requirements of microdroplet systems.

In the realm of intracellular directed evolution, there is a significant need for tools with high targeting accuracy, high mutation rates, large mutation windows, and ease of operation. While CRISPR systems are currently the primary enzymes used, there is an urgent demand to develop new programmable nucleases with precise recognition capabilities and low off-target activity. These novel nucleases should also work efficiently and cooperatively with existing editing components *in vivo*.

Expected Progress Recently: Develop fluorescence-based *in vitro* screening (e.g., FACS, FADS) for >10 protein components at 10 million clones/day. Achieve label-free screening (e.g., high-throughput mass spectrometry) at >10,000/day. Optimize intracellular directed evolution tools for targeted, large-window (>1 kb), high-efficiency mutagenesis across chassis cells.

Expected Progress by 2030: Develop *in vitro* screening methods utilizing fluorescent labeling techniques such as Fluorescence-Activated Cell Sorting (FACS) and Fluorescence-Activated Droplet Sorting (FADS). Aim to achieve a screening capacity of 10 million clones per day for 100 types of protein components, and 100 million clones per day for 10 types. Additionally, develop label-free screening methods, such as

high-throughput mass spectrometry, with a throughput exceeding 100,000 clones per day. Building on novel programmable nucleases, such as Argonaute nucleases with precise target recognition and cleavage capabilities that bypass sequence recognition restrictions, test their catalytic activity and elucidate their molecular mechanisms. Establish a synergistic system by coupling these nucleases with components like T7 polymerase, deaminases, and others. Propose a new theoretical framework for *in vivo* directed evolution, aiming to surpass the limitations of existing CRISPR technology and develop a new generation of directed evolution tools.

Potential Solutions

Innovative fluorescent probes, biosensors, and other fluorescence coupling methods will be adopted to broaden the applicability and enhance the performance of fluorescence-based assays. Additionally, specialized high-throughput mass spectrometry equipment and sample preparation systems will be developed to facilitate rapid screening of protein components. Furthermore, advanced high-throughput screening platforms, such as faster flow cytometers, microdroplet screening systems, and automated microdroplet mass spectrometry detection systems, will be developed to meet the increasingly stringent requirements of high-throughput screening.

Existing Cas enzymes will be engineered to broaden their recognition sites and enhance their specificity. High-targeting CRISPR-Cas intracellular directed evolution tools will be utilized to develop a high-throughput, automated biosynthesis platform for DNA synthesis, transformation, and screening. The mutation rate will be increased, and the editing window will be expanded by improving the overall experimental throughput. For technologies with large editing windows but limited targeting accuracy, such as the T7 RNA polymerase-cytidine deaminase complex, orthogonality within cells will be improved through protein engineering and promoter modifications to reduce off-target effects. Additionally, intracellular directed evolution techniques will be integrated with automated biosynthesis platforms to enable the mutation screening of hundreds of target proteins and multiple rounds of mutation within a single day.

Bioinformatics analysis will be used to deeply mine microbial-derived defense system nucleases, with a focus on characterizing evolutionary representative protein components to enhance their performance and reduce off-target effects. The synergistic effects of novel nucleases combined with T7 polymerase, deaminases, and other enzymes

will be investigated. Optimization of the random mutation system will be achieved through rational fusion design and protein backbone engineering. Additionally, continuous, high-throughput cultivation platforms will be developed to support efficient intracellular directed evolution, enabling mutation screening of thousands of target proteins and conducting hundreds of rounds of parallel mutations within a single day.

3.3.5 Summary

In recent years, the exponential growth of protein sequences in databases and the rapid development of cutting-edge technologies in life sciences have provided an unprecedented repository of bioparts. Effectively utilizing this vast resource data and precisely elucidating the relationship between protein structures and functions presents both a paramount challenge and a transformative opportunity. Future research aims to establish a large-capacity, automated, and digital protein component center using next-generation technologies such as high-throughput sequencing, deep mutational scanning, and microfluidics. This initiative will address the prevailing issues of data scarcity and bias in existing protein databases, providing foundational elements for innovative protein design and engineering, and enabling deeper insights into the structural and functional mechanisms of protein components.

Building on this foundation, leveraging rapidly developing artificial intelligence technologies will facilitate the development of novel algorithms for independent and innovative protein design. Standardized interfaces with specific multi-modules based on spatial structure design can be established, fostering a deeper integration of the “black-box” model and the “white-box” model. This integration will deepen the understanding of the “sequence-structure-function-interaction” mapping and underlying laws of proteins, enable the design of multiple independent protein interfaces, and pave the way for a new generation of intelligent computational platforms for protein engineering. These advancements will expand application scenarios for protein components, achieve leapfrog development in protein engineering, and accelerate the industrial realization of engineering biology.

References

[1] Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary

data. Nature, 2021, 599: 91-95.

[2] Leman J K, Weitzner B D, Lewis S M, et al. Macromolecular modeling and design in Rosetta: Recent methods and frameworks. Nat Methods, 2020, 17: 665-680

[3] Xiong P, Wang M, Zhou X, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat Commun, 2014, 5: 5330

[4] Langan R, Boyken S, Ng A H, et al. *De novo* design of bioactive protein switches. Nature, 2019, 572: 205-210

[5] Boyken S, Benhaim M, Busch F, et al. *De novo* design of tunable, pH-driven conformational changes. Science, 2019, 364: 658-664.

[6] Chen Z, Kibler R, Hunt A, et al. *De novo* design of protein logic gates. Science, 2020, 368: 78-84.

[7] Lutz I, Wang S, Norn C, et al. Top-down design of protein architectures with reinforcement learning. Science, 2023, 380: 266-273.

[8] Cao L, Coventry B, Goreshnik I, et al. Robust *de novo* design of protein binding proteins from target structural information alone. Nature, 2022, 605: 551-560.

[9] Wang J, Lisanza S, Juergens D, et al. Scaffolding protein functional sites using deep learning. Science, 2022, 377: 387-394.

[10] Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, et al. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. PLoS Comput Biol, 2018, 14(4): e1006112.

[11] Li F, Yuan L, Lu H, et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nat Catal, 2022, 5: 662-672.

[12] Kintses B, Hein C, Mohamed M F, et al. Picoliter cell lysate assays in microfluidic droplet compartments for directed enzyme evolution. Chem Biol, 2012, 19(8): 1001-1009.

[13] Ma F, Chung M T, Yao Y, et al. Efficient molecular evolution to generate enantioselective enzymes using a dual-channel microfluidic droplet screening platform. Nat Commun, 2018, 9(1): 1030.

[14] Wang H H, Kim H, Cong L, et al. Genome-scale promoter engineering by coselection MAGE. Nat Methods, 2012, 9 (6): 591-593.

[15] Jakočiunas T, Pedersen L E, Lis A V, et al. CasPER, a method for directed evolution in genomic contexts using mutagenesis and CRISPR/Cas9. Metab Eng, 2018, 48: 288-296.

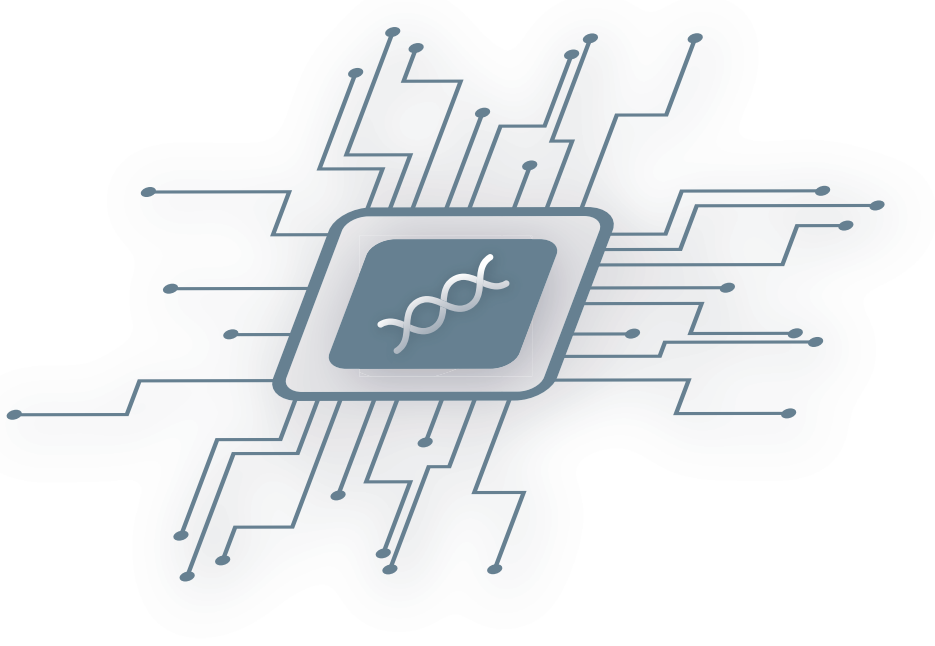
[16] Bao Z, HamediRad M, Xue P, et al. Genome-scale engineering of *Saccharomyces cerevisiae* with single-nucleotide precision. Nat Biotechnol, 2018, 36 (6): 505-508.

[17] Garst A D, Bassalo M C, Pines G, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nature Biotechnol, 2017, 35 (1): 48-55.

[18] Roy K R, Smith J D, Vonesch S C, et al. Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat Biotechnol, 2018, 36 (6): 512-520.

[19] Esvelt K M, Carlson J C, Liu D R, et al. A System for the continuous directed evolution of biomolecules. Nature, 2011, 472 (7344): 499-503.

Genetic Circuits



Authors Li Chun, Song Hao, Zhou Yong-Jin, Lian Jia-Zhang, Shi Shuo-Bo, Qin Lei

3.4 Genetic Circuits

3.4.1 Abstract

Genetic circuit engineering involves the design and construction of biological processes under engineering principles, assembling and programming fundamental bioparts to confer novel functions that are non-native or achieve critical, challenging objectives. Genetic circuits enable the design of complex synthetic biology pathways and physiological functions with dynamic control over synthetic regulatory systems, offering broad applicability. With advancements in artificial intelligence and the maturation of biological databases, the design of genetic circuits will become intelligent, precise and efficient in areas such as biosensors, logic gate circuits and natural/non-natural genetic circuit design. Furthermore, intelligent circuit design, coupled with precise, efficient, automated assembly and high adaptation towards chassis, will drive the development of robust genetic circuits.

3.4.2 Technical Overview

3.4.2.1 Genetic Circuit Design Principles

A genetic circuit refers to a genetic device in synthetic biology that is composed of various regulatory parts and regulated genes. It can express gene products in a tunable, timed, and quantitative manner under given conditions. Currently, the design of genetic circuits is based on the professional knowledge of organic chemistry, biochemistry, and electrical engineering, as well as literature reports and practical experience. Taking the interactions of bioparts and the characteristics of system signal transduction as the basic prerequisites, genetic circuits are designed by drawing on the reported natural or artificial biosynthetic pathways, the molecular interaction relationships of similar reaction types, and enzymatic biological reactions. Biopart refers to nucleotides with specific functions, which is the simplest and most fundamental BioBrick in the genetic system. With the enrichment of the database of biological reaction rules and standardized bioparts, by extracting and applying the templates of enzymatic biological reaction rules, selecting appropriate bioparts, and using bio-retrosynthetic computational tools to achieve the precise and automated design of functional genetic circuits will become the new criteria for genetic circuits design [1-3]. In the future, based on the large amount of data on

biological reactions and genetic circuits generated by automated design, artificial intelligence technology can be used to learn the essential characteristics and laws of biological reactions, to predict all potential biological reaction sites, reaction types, and corresponding probabilities, thus expanding the space and degree of freedom of genetic circuit design, and achieve efficient, novel, and optimized genetic circuit design.

3.4.2.2 Bioparts Development

Bioparts have various functions, including regulation, expression, response, etc., and are essential components of genetic circuits. However, the scarcity of types of bioparts, the unclear characterization and description, and the inadaptation between bioparts or between parts and chassis cells are among the main obstacles hindering their application in the design and development of genetic circuits. The mining, modification, and standardization of bioparts are crucial for achieving the efficient assembly of genetic circuits. Currently, there are more than 20,000 bioparts in the "Registry of Standard Biological Parts". Common bioparts include promoters, ribosome binding sites (RBS), terminators, aptamers, etc. The promoter is a key element that regulates gene expression at the transcriptional level and can be used to regulate the expression levels of each gene in the genetic circuit. By rationally designing and performing directed evolution on promoters, a set of core promoters and response parts can be developed and compiled to achieve predictable and adjustable precise expression control of multiple genes in organisms. The ribosome binding site (RBS) is a segment at the 5' end (upstream) of the mRNA molecule. It can bind to the ribosome and correctly position it at the translation initiation site, and it can also control the accuracy and efficiency of mRNA translation initiation. RBSs with different sequence compositions can be screened to affect the expression levels of downstream genes. The terminator is a control element that is independent of the gene coding sequence and promotes the termination of transcription. Terminators with different activities will directly affect the amount of synthesized mRNA and ultimately affect the degree of gene expression. Developing controllable, designable, and short terminators can avoid the interference caused by redundant sequences and help achieve fine regulation of metabolic pathways. An aptamer is a DNA or RNA sequence that can bind to a specific protein or small molecule ligand. It can bind to the target substance with high specificity and is widely used in the field of sensors. Through the efficient screening and design of aptamers, a high-throughput screening system for aptamer-based sensors can be established. At the same time, RNA aptamers can also be

used to construct molecular-inducible riboswitches to directly regulate gene expression at the translation level. In addition, transcription regulatory protein factors regulate gene expression at the transcriptional level by regulating the transcription of promoters. Through the molecular modification of transcription regulatory protein factors, they can regulate genetic circuits more efficiently and accurately.

3.4.2.3 Regulatory Genetic Circuit Design

Host specificity, influence of environment, modularity, and the tunability of components are all key factors in the design of genetic circuits. Functional genes and regulatory parts are the two foundations of biological genetic circuits, among which regulatory parts determine the controllable characteristics of genetic circuits. A regulatory genetic circuit refers to a dynamic genetic circuit in which the genetic circuit changes the output signal according to the change of the input signal, mainly including sensors, logic gate genetic circuits, and orthogonal expression systems. The existing regulatory genetic circuits have problems such as insufficient biosensing parts, unclear signal transduction mechanisms, and poor controllability and predictability in the design of genetic circuits. In the future, it is expected to make breakthroughs in aspects such as the construction of a library of molecular interaction pairs for biosensing, the efficient screening and design of sensors, the intelligent design of logic gates, and the intelligent response design of biological communities. This will enable the rapid and precise conversion of input and output signals of synthetic biological systems, as well as the intelligent and autonomous regulation of biological cells and systems.

3.4.2.4 Functional Genetic Circuit Design

The traditional design and construction methods of genetic circuits mainly imitate the design of natural functional genetic circuits, which involve a large number of trial-and-error processes. They have small throughput and low efficiency. The combination of the design approach and the synthetic regulatory parts is rather blind or random, and the designed circuits cannot be well adapted to the chassis cells. Therefore, automated design tools represented by bio-retrosynthetic tools replace the traditional design and construction methods that rely on experience and trial-and-error, achieving high-throughput, automated and precise design of functional genetic circuits. The concept of bio-retrosynthesis originates from chemical retrosynthesis and is divided into two stages: the prediction and screening of the bio-retrosynthetic pathway. In the stage of predicting the

retrosynthetic pathway, by using a set of biochemical reaction rules that describe the chemical transformation patterns between substrate and product molecules at the atomic level, the reactions for synthesizing the target compound and the enzymes catalyzing these reactions are inferred. This enables the transformation of the input compound (i.e., the target compound) into a series of intermediate compounds, and finally, the establishment of a functional genetic circuit that converts the substrate into the product.

3.4.2.5 High-efficiency Assembly for Genetic Circuit

To achieve the functional expression of genetic circuits in organisms, it is first necessary to construct parts such as promoters and genes into functional modules according to certain rules, and then assemble various functional modules in the organism into genetic circuits with specific functions. Currently, the main methods for *in vivo* assembly include homologous recombination, as well as the more efficient and modular GoldenGate and CRISPR-Cas technologies. However, their editing efficiency and scale still cannot meet the requirements for the efficient assembly of large-fragment multi-genetic circuits. Therefore, it is necessary to develop multifunctional, multi-site, high-throughput genome editing technologies and automated nucleic acid assembly platforms to further improve the assembly efficiency and throughput of genetic circuits.

3.4.2.6 Circuit-chassis Adaptation

The adaptation between the genetic circuit and the chassis cell determines whether the designed genetic circuit can function properly. Currently, the adaptation between the genetic circuit and the chassis cell is mainly achieved through trial-and-error experiments, and there is still a lack of certain theoretical basis and design principles in this process. Solving the following problems will help improve our understanding of the adaptation of the chassis cell, including: how to determine the integration method and the principle for selecting the integration site of multiple DNA fragments in the genome; how to screen the type of chassis cell that is most suitable for the target genetic circuit; how to screen or design standardized bioparts that meet the requirements of the genetic circuit; how to obtain a genetic circuit with universality and stable expression; how to efficiently evolve engineered cells to improve the adaptation between the genetic circuit and the chassis cell, so as to achieve a compatible, orthogonal and robust high-efficiency adaptation between the genetic circuit and the chassis cell, etc.

3.4.3 Roadmaps

Current Status		
Combining bio-retrosynthetic algorithms with artificial intelligence techniques allow for simple biosynthetic pathway design, where genetic circuit design relies heavily on experience.		
Objective 1: Achieve Efficient Design of Sensors and Logic Gate Circuits		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance the capacity of designing regulatory genetic circuits such as sensors and logic gates.	**Rational design of regulatory genetic circuits.** • Establish molecular interaction pair component database for signal transduction and sensing, analyze unknown signaling pathways in organisms, and enrich component database. • Incorporate gene expression interactions into the predictable design of cytogenetic systems to enhance the generality of genetic circuits. • Biosensing of environmental signals (light, electricity, physical forces, etc.) and metabolites in model organisms, and establishment of high-throughput screening methods for sensors. • Controllable and predictable design and modification of logic gate genetic circuits with multiple regulators. • Precise, predictable, and regulated dynamic gene expression for processes with multi-species composition.	**Intelligent design of regulatory genetic circuits.** • Further enrichment of the database of biomolecular interaction pairs enable precise regulation of the sensitivity, detection range, and thresholds of sensors. • Hierarchical models for simulating and predicting gene regulation, metabolism, and system level behavior. • Improvement of sensor screening throughput to enable biosensing of any environmental signal, metabolite, light, electricity, physical force, etc. in non-model organisms. • Application of artificial intelligence to the design of logic gate genetic circuits and development of online design tools. • Dynamic and controllable regulation of growth and production of various organisms among species biological community.

Objective 2: Realize Intelligent Automated Design of Genetic Circuits		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance biosynthetic pathway design capabilities and develop genetic circuit design software.	**Automated design of functional genetic circuits.** • Mine new bioparts, collecting resources, organizing, and expanding the bioparts, and improving data standards and testing standards for bioparts. • Use machine learning to summarize a large amount of literature related to circuit design and establish databases of biological metabolism and bioparts. • Use the existing biological reactions as templates and combine the resources of biological metabolic and component databases, utilize bio-retrosynthetic computational tools to realize the accurate and automated design of genetic circuits for biosynthetic pathways. • Develop genetic circuit design visualization software and simulate the selection and assembly process of bioparts in genetic circuit design software.	**Intelligent design of functional genetic circuits.** • Automatic real-time update of biological metabolic database and biopart database resources. • Establish intelligent algorithms through the characterization of known bioparts to realize the rational design of specific functional bioparts. • Based on biological reaction big data and artificial intelligence technology, extract the characteristic laws of biological reaction, realize the prediction of potential biological reaction types and probability of a given compound, and combine with enzyme function prediction and enzyme sequence design tools to realize efficient genetic circuit design of biosynthetic pathways. • Enrich the database of bioparts of genetic circuit design visualization software, increase the auxiliary functions of the software, and realize the interface with automated assembly equipment.

Figure 1 Roadmap for intelligent genetic circuit design

Current Status		
It can realize the assembly of DNA fragments of more than 10 kb and can realize the automated assembly of genetic circuits.		
Objective 1: Intelligent, Automated, Precise and Efficient Assembly of Genetic Circuits		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Improve the accuracy and automation of genetic circuit assembly.	**Efficient assembly of genetic circuits.** • One-step assembly of genetic circuits with more than 10 DNA fragments across the whole genome of a model host without off-target effects. • Integrate machine learning to identify problematic and difficult-to-integrate sequences with optimization recommendations. • Develop error correction methods for the genetic circuit assembly process to improve assembly accuracy.	**Automated assembly and error correction of genetic circuits.** • One-step assembly of genetic circuits with more than 20 DNA fragments across the whole genome of a model host without off-target effects. • Integrate machine learning to build an automated assembly platform to automate genetic circuit assembly based on optimized recommendations. • Realize automatic error correction in genetic circuit assembly process.

Objective 2: Achieve Efficient Adaptation of Genetic Circuits to Chassis Cells in a Compatible, Orthogonal and Robust Manner		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance the adaptation of genetic circuits to chassis cells.	**Rapid adaptation technologies for genetic circuits and chassis cells.** • Quantitative and specific integration of multiple loci in model organisms, monitoring and regulation of genetic and epigenetic mechanisms over time. • Develop generalized genetic circuits with stable and efficient expression in multiple species. • Rapid screening method for chassis cells adapting to genetic circuits. • Methods for rapid cell evolution of genetic circuits adapted to chassis cells.	**Intelligent adaptation technologies for genetic circuits and chassis cells.** • Quantitative, site-specific integration of multiple loci in non-model organisms. • Development of generalized genetic circuits for stable and efficient expression in multiple species. • Application of artificial intelligence to screen chassis cells for optimal adaptation to genetic circuits. • Intelligent cellular evolutionary methods and key mechanisms for the adaptation of genetic circuits to chassis cells.

Figure 2 roadmap for automated genetic circuit assembly and adaptation

3.4.4 Technical Pathways

3.4.4.1 Intelligence of Genetic Circuit Design

Current Technologies: Currently, the intelligence of genetic circuit design mainly relies on retrosynthetic pathway design tools such as Route Designer, PathPred, Sympheny, GEM-Path, RetroPath, SimZyme, Cellor, EcoSynther, RetroPath2.0, RetroPath RL and NovoPathFinder, etc [4-6]. Since retrosynthetic algorithms based on finite reaction rules for known biochemical reactions can generate a large number of predicted pathways, making most biological retrosynthetic algorithms complex and with a high false positive rate. To address this problem, researchers have begun to combine retrosynthesis algorithms with artificial intelligence techniques, such as the RetroPath RL tool [7]. Regulatory genetic circuits include sensors, logic gate genetic circuits, and orthogonal expression systems, which are used to sense and respond to changing internal and external conditions and translate them into the expression of genes in a cell or the activation of certain functions in a system. Researchers have developed a series of logic gates based on transcription factor interaction pairs and further utilize the developed genetic circuit design software Cello 2.0 to realize the dynamic regulatory process in any mode [8]. However, the controllable and precise design of regulatory genetic circuits still faces the problems of insufficient regulatory parts, function susceptibility to interference by chassis cell, and poor generalizability among different cells, therefore, new methods are needed to design regulatory genetic circuits and to resolve the mechanism of action behind them.

Objectives and Breakthroughs: Intelligent design of gene circuits; efficient design of sensor and logic gate circuits; intelligent automated design of genetic circuits.

Challenges: The transmission processes and mechanisms of a large number of biological and physical signals have not been analyzed yet. For example, during the construction of cell factories, it has been found that there are some promoters and responsive genes of plant-derived natural products in microorganisms, but their response mechanisms are still unclear [9]. It is only limited to sensors for some common metabolites, and there is a lack of the ability for orthogonal, programmable, and non-repetitive regulation of the required gene regulatory effects.

The interaction mechanism in multi-species systems is not clear. The molecular mechanism of sensors is not clear. It is necessary to figure out how to rationally modify

sensors to achieve the expected sensitivity, detection range, and threshold. In addition, there are also problems in the adaptation between sensors and chassis cells. The design of logic gate genetic circuits still depends on expert experience. The robustness of artificial biological communities is insufficient.

The types of bioparts are scarce, and their characterization and description are not clear. The technology of machine learning for reading literature is not perfect, and the metabolic databases and biopart databases supporting bio-retrosynthetic algorithms are incomplete. It is difficult to precisely design long-pathway biosynthetic circuits. There is a need for the construction and standardization of biopart database suitable for design software.

The collection, organization, and sharing mechanism of bioparts still need to be improved, and there is a lack of intelligent algorithm software. It is difficult to extract information such as biological reaction rules, reaction conditions, and enzyme sequences from the massive and real-time updated literature information and update the database in real time. The algorithms and software for the automated design of metabolic pathways, metabolic networks, and genetic circuits are still not perfect. There is a need to figure out how to improve the accuracy of the design software. It is also necessary to explore how to achieve the "one-step" full automation of genetic circuits from design to automatic construction.

Expected Progress Recently: Establish a biopart database of molecular interaction pair for signal transduction and sensing, analyze the important unknown signal transduction pathways in organisms, and enrich the biopart database of biological molecular interaction pair. Achieve biosensing of any environmental signals (such as light, electricity, physical forces, etc.) and metabolites in model organisms and establish a high-throughput screening method for sensors. Conduct controllable and predictable design and modification of logic gate genetic circuits with multiple regulators. Achieve precise, predictable, and controllable dynamic gene expression in processes composed of multiple species.

Mine, design, and construct new bioparts and build a biopart database. Use machine learning to summarize a large number of literatures related to circuit design and establish databases of biological metabolism and bioparts. Take the existing biological reactions as templates, combine the database resources of biological metabolism and bioparts, and use bio-retrosynthetic computational tools to achieve precise and automated design of the genetic circuits of biosynthetic pathways. Develop visualization software for genetic

circuit design and realize the selection of bioparts and the simulation of the assembly process in the genetic circuit design software.

Expected Progress by 2030: Further enrich the biopart database of biological molecular interaction pair. This will enable the precise regulation of sensitivity, detection range, and threshold of sensors. Increase the screening throughput of sensors and achieve biosensing of any environmental signals (such as light, electricity, physical forces, etc.) and metabolites in non-model organisms. Use artificial intelligence to design logic gate genetic circuits and develop online design tools. Achieve dynamic and controllable regulation of the growth and production of various organisms in biological communities.

Accelerate the sharing and distribution of biopart data and physical items. Develop decentralized data sharing technology, unify the sharing mechanism of biopart libraries, and establish the design principles of bioparts. Realize the automatic and real-time update of the resources in the biological metabolism database and the biopart database. Extract the characteristic laws of biological reactions based on big data of biological reactions and artificial intelligence technology and predict the potential biological reaction types and probabilities of given compounds. Combine enzyme function prediction and enzyme sequence design tools to achieve the design of genetic circuits for efficient biosynthetic pathways. Enrich the biopart database of the visualization software for genetic circuit design, add auxiliary functions to the software, and realize the connection with automated assembly equipment.

Potential Solutions

Analyze the unknown biological signal transduction processes through epigenetics, multi-omics technologies, and chip sequencing technologies, and discover new molecular interaction pairs. Collect the molecular interaction pairs for signal transduction and sensing and initially form a database of sensors. Establish efficient, rapid, and high-throughput screening and construction methods for sensors. Screen, construct, and test a large number of sensors that respond to environmental signals and specific compounds. Develop mutually orthogonal promoter-regulator pairs and gene manipulation systems orthogonal to the host. Develop mutually orthogonal CRISPR systems, recombinase systems, etc. Analyze the mechanism of material and energy exchange in multi-species systems. Reasonably design the regulatory genetic circuits in multi-species systems to achieve stable and controllable gene expression of genetic

circuits in multi-species systems.

Expand the biopart database of biological molecular interaction pair to increase the selectivity of signal sensing parts. Establish a digital description model of sensor, expanding from an empirical model to a theoretical model. Establish a sensor database. Use artificial intelligence to predict and design sensors and establish an ultra-high-throughput sensor screening method. Develop new methods for molecular manipulation technologies in non-model organisms. Establish methods for using artificial intelligence to predict and design logic gate genetic circuits. Develop algorithms, toolkits, software, websites, etc., for the design of logic gate genetic circuits. Establish a dynamic model of biological communities and develop algorithms and software for the design and prediction of biological communities. Use computer software to design and establish stable multi-species systems such as mutualism and commensalism.

Develop robust and high-throughput screening methods for bioparts; collect, organize and expand resources of bioparts, improve data standards and testing standards for bioparts, and establish a "Registry of standard biopart".

Integrate the resources of current open-source biological reaction, enzyme and metabolism databases. Supplement the database of metabolism and biopart information through crowdsourcing and other methods. Realize the automatic extraction of biological reactions, reaction conditions and enzyme sequences from PDFs, webpages and other sources of literature by means of natural language processing technology and use them for the automatic updating of metabolism databases.

Make full use of the current known data of metabolic and bioparts to construct a biological reaction pathway network through intelligent algorithms, which is coupled with the chassis cellular metabolic pathway network as a chassis network in order to reduce the length of the genetic circuit prediction. Evaluate the biological reaction rules and assign different weights to them and train the bio-retrosynthetic network by combining deep learning and neural network techniques for genetic circuit design. By considering the thermodynamic feasibility of the genetic circuits, compound toxicity, growth pressure on chassis cells and fitness with chassis cells and other constraints to assist the screening of long pathway genetic circuits.

Categorize bioparts according to function, origin and other properties, add standardized interfaces or design standardized modifications in the sequences of bioparts in a reasonable manner, help establish a standardized biopart assembly process, and publish it free of charge.

Develop intelligent biopart retrieval tools to accelerate the convergence of biopart data, physical objects and design tools, and provide big data learning for the design principles of bioparts. Take advantage of blockchain technology to decentralize the management of the biopart database. Summarize and optimize the existing software sharing mechanisms and formulate reasonable sharing protocols by combining the characteristics of bioparts.

Deep learning technology is utilized to extract features from biological reaction big data and genetic circuit big data, combined with genetic circuits manual design experience to provide a priori knowledge, and neural network structure search technology to retrieve suitable neural network models for the prediction of biological reaction types and reaction patterns. For the biochemical reactions in the metabolic database, clustering is performed based on the structural similarity of catalytic parts, reactants, and products, and the reaction cascade information is also extracted to construct a network model of biological reaction pathway diagrams, which can provide a basis for the design of novel genetic circuits.

Add factors affecting genetic circuits in the design software and be able to select and adjust these factors. Establish the docking interface between genetic circuit design software and automated assembly equipment.

3.4.4.2 Automation of Genetic Circuit Assembly and Adaptation

Current Technologies: Fragment assembly of more than 10 kb DNA can be achieved using *in vitro* (e.g., Gibson assembly) and *in vivo* (yeast-mediated homologous recombination) techniques [10]. The use of CRISPR technology enables rapid construction of target genetic circuits, simultaneous editing of 12 alleles or simultaneous activation, disruption and knockdown of multiple genes with 100% efficiency. Currently, the use of CRISPR technology in *Saccharomyces cerevisiae* is capable of simultaneous gene insertion at 5-8 loci [11].

Objectives and Breakthroughs: To realize intelligent and automated precise and efficient assembly of genetic circuits; efficient adaptation of genetic circuits to chassis cells in a compatible, orthogonal and robust manner.

Challenges: The performance limitations of current gene editing technologies and the poor fitness of exogenous genes to the host. The sequence laws that lead to difficulties in DNA assembly and their mechanisms of action are unclear. The basic principles of error correction in the genetic circuit assembly process are unclear. Lack of efficient

automated assembly platform technology.

The interrelationship between genetic circuit integration sites and gene expression levels and the regulatory mechanisms are unclear. The ability of genetic circuits to function in multiple species depends mainly on the generality of their regulatory parts. Currently there is a lack of guidelines for the selection of chassis cells. Genetic circuits have problems such as inadaptation with chassis cells and poor genetic stability. Genomic information and gene editing tools for non-model organisms are imperfect. The design and characterization of the versatility of regulatory parts are incomplete. The selection of chassis cells generally relies on the approach of expert experience, and the evolution of engineered cells is inefficient and labor-intensive.

Expected Progress Recently: One-step assembly of genetic circuits with more than 10 DNA fragments across the entire genome of a model host without off-target effects. Integration of machine learning to identify problem-prone, difficult-to-integrate sequences and give optimization suggestions. Development of error correction methods for the genetic circuit assembly process to improve assembly accuracy.

Perform quantitative, site-specific genetic circuit integration at multiple sites in model organisms. Develop universal genetic circuits that are stably and efficiently expressed in multiple species, as well as methods for rapid screening of chassis cells adapted to genetic circuits, and rapid cellular evolution of genetic circuits adapted to chassis cells.

Expected Progress by 2030: One-step assembly of genetic circuits with 20 and more DNA fragments across the whole genome of the model host without off-target effects; integration of machine learning to build an automated assembly platform to realize automatic assembly of genetic circuits based on optimization suggestions; realization of automated error correction in the assembly process of genetic circuits.

Quantitative and specific integration of multiple loci in non-model organisms; development of universal genetic circuits that are stably and efficiently expressed in any species; application of artificial intelligence to screen chassis cells with optimal fitness to genetic circuits; intelligent cellular evolutionary methods for adaptation of genetic circuits and chassis cells.

Potential Solutions

Improve base-editing enzymes to cover all possible PAM sequences or discover new

base-editing enzymes that do not require PAM sequences. Deeply understand the interactions among nucleic acid conversion, gene editing, exogenous gene insertion, and host DNA repair to improve the adaptation between exogenous genes and the host. Use machine-learning algorithms to predict and identify DNA sequences that lead to low DNA assembly efficiency and provide efficient interface sequences suitable for DNA assembly. Develop low-cost methods for reading ultra-long DNA fragment sequences. Utilize highly specific ribozymes such as genome-editing tools to develop new nucleic acid assembly methods and clarify the basic principles of error correction during genetic circuit assembly.

Discover highly efficient integration sites and guide sequences. Deeply understand the interactions among gene editing, exogenous gene insertion, and host DNA repair to improve the adaptation between exogenous genes and the host. Develop genome-editing tools with low or even no off-target effects. Integrate hardware modules with different functions to develop efficient automated nucleic acid assembly platform technologies. Clarify the basic principles of error correction during genetic circuit assembly. Construct new error-correction technologies for genetic circuit assembly.

Establish a database of genetic circuit integration sites and analyze the quantitative relationships among integration sites, genome structure, and gene expression levels. Establish monitoring and spatiotemporal regulation methods for genetic and epigenetic mechanisms. Develop universal and stable regulatory parts across multiple species and basically clarify the recognition and working mechanisms of bioparts in different species. Combine multi-omics data to establish a metabolic context library for chassis cells and quantify metabolic fluxes. Establish high-throughput methods for rapidly constructing target genetic circuits in different chassis cells. Establish rapid evolution methods for engineered cells or methods for coupling growth and production to improve the adaptation between genetic circuits and chassis cells.

Improve genomic information and gene-editing tools for non-model organisms and establish a database of integration sites for important non-model organisms. Develop universal and stable regulatory parts across species and clarify the recognition and working mechanisms of bioparts in different species. Based on the chassis cell database and artificial intelligence, establish algorithms and software for predicting and evaluating the most suitable chassis cells. Use tools such as gene editing to establish global or local accelerated evolution technologies. Establish autonomous and intelligent evolution methods and strategies for engineered cells to achieve the autonomous evolution of

engineered cells without artificial intervention.

3.4.5 Summary

In response to the current problems such as the lack of standardized bioparts in genetic circuit design and visualization design software, as well as the lack of efficient and automated methods for genetic circuit assembly, in the future, technical breakthroughs will be made in aspects such as the efficient design of sensor and logic gate circuits, the intelligent and automated design of genetic circuits, the intelligent, automated, precise and efficient assembly of genetic circuits, and the efficient adaptation of genetic circuits to chassis cells in a compatible, orthogonal and robust manner. These efforts aim to achieve technological development in the field of genetic circuit engineering.

References

[1] Segler M H, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555(7698): 604-610.

[2] Granda J M, Donina L, Dragone V, et al. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature, 2018, 559(7714): 377-381.

[3] Lian J, HamediRad M, Hu S, et al. Combinatorial metabolic engineering using an orthogonal tri-functional CRISPR system. Nature Communications, 2017, 8(1):1-9.

[4] Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of *de novo* synthetic pathways. Current Opinion in Chemical Biology, 2015, 28: 99-104.

[5] Yim H, Haselbeck R, Niu W, et al. Metabolic engineering of *Escherichia coli* for direct production of 1, 4-butanediol. Nature Chemical Biology, 2011,7(7): 445-452.

[6] Moriya Y, Shigemizu D, Hattori M, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Research, 2010, 38: W138-W143.

[7] Koch M, Duigou T, Faulon J L. Reinforcement learning for bioretrosynthesis. ACS Synthetic Biology, 2019, 9(1): 157-168.

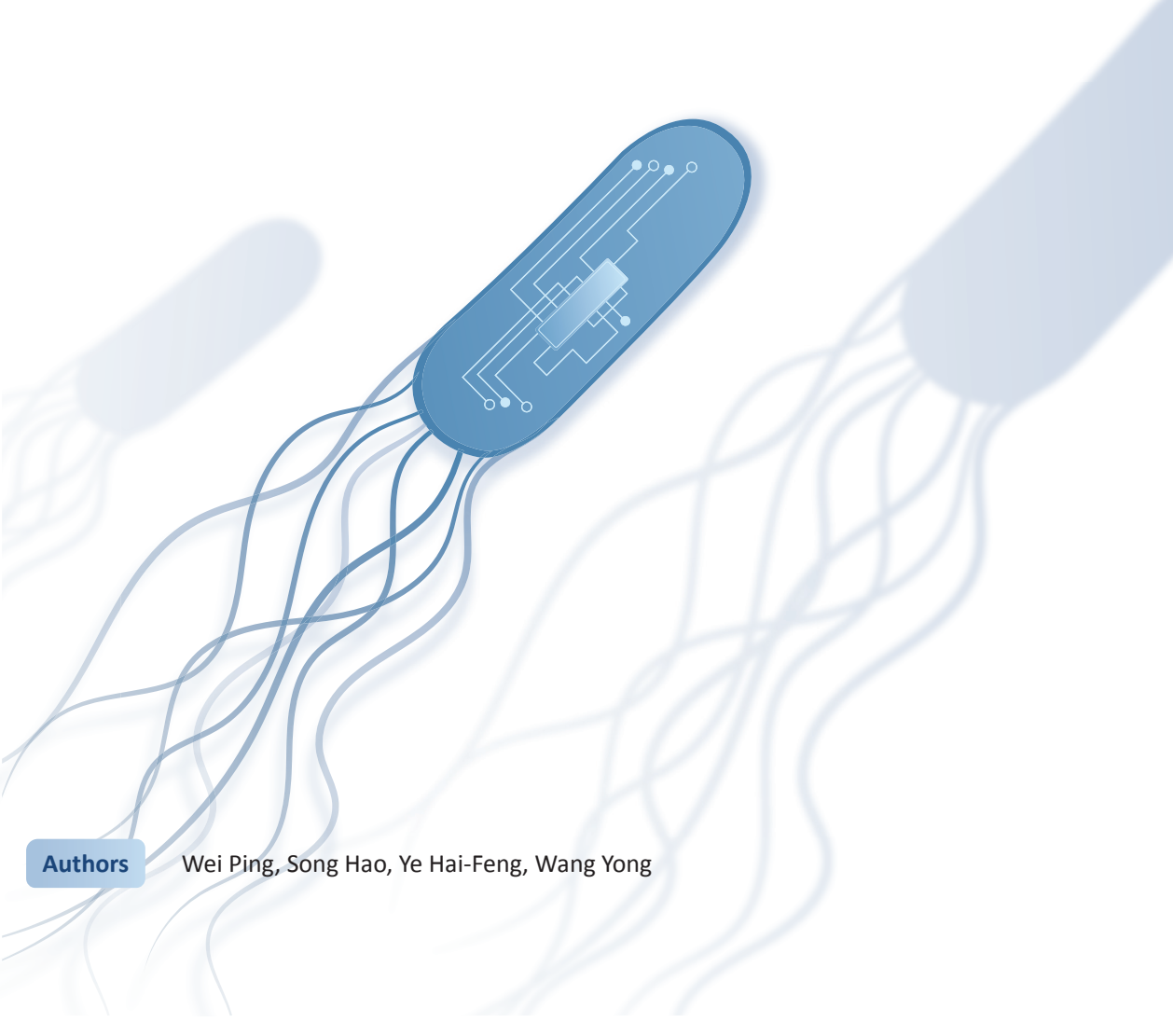
[8] Chen Y, Zhang S, Young E M, et al. Genetic circuit design automation for yeast. Nat Microbiol, 2020, 5(11):1349-1360.

[9] Li J, Kolberg K, Schlecht U, et al. A biosensor-based approach reveals links between efflux pump expression and cell cycle regulation in pleiotropic drug resistance of yeast. J Biol Chem, 2019, 294(4):1257-1266.

[10] Shao Z, Zhao H, Zhao H. DNA assembler, an *in vivo* genetic method for rapid construction of biochemical pathways. Nucleic Acids Research, 2008, 37(2): e16.

[11] Lian J, Bao Z, Hu S, et al. Engineered CRISPR/Cas9 system for multiplex genome engineering of polyploid industrial yeast strains. Biotechnology and Bioengineering, 2018, 115(6): 1630-1635.

Chassis Cells



Authors Wei Ping, Song Hao, Ye Hai-Feng, Wang Yong

3.5 Chassis Cells

3.5.1 Abstract

Chassis cells serve as the basic form of synthetic biology to realize the function, providing a self-replicating, inheritable, and evolvable active biochemical environment for artificial genetic circuits, functional molecules, organelles, genomes, and other hierarchical biological constructs. They also represent the primary engineering targets for synthetic biology applications in important economic production processes such as industry and healthcare. Chassis cell engineering includes three main types of microbial cells, mammalian cells, and plant cells, and is engineered according to the characteristics of different application scenarios in industrial production and healthcare. Despite established bioengineering foundations, current challenges include inefficient genetic manipulation techniques, limited genetic bioparts toolkits, and incomplete understanding of cellular molecular physiology. By 2030, breakthroughs in these areas are expected to significantly enhance industrialization capabilities across current industrial, environmental, energy, and healthcare sectors.

3.5.2 Technical Overview

3.5.2.1 Core Attributes of Chassis Cells

An ideal chassis cell must possess the following attributes. Ease of cultivation: Rapid and efficient proliferation under low-cost conditions. Ease of genetic modification: Compatibility with straightforward molecular techniques for gene delivery, genome reconstruction, or editing. Regulability: Responsiveness to external biological, physical, or chemical controls to enhance operational intelligence, safety, and reliability. High productivity: Optimized genetic and metabolic systems for synthesizing target molecules (e.g., small molecules, polysaccharides, lipids, proteins). Environmental adaptability: Controllable proliferation and functional deployment across diverse *in vivo* or industrial environments, with minimal ecological impact. Safety: Stable and reliable execution of engineered functions, avoiding genetic leakage, unintended mutations, or side effects such as mutations caused by genetic instability.

3.5.2.2 Industrial Microbial Chassis Cell Engineering

In response to the requirements for constructing microbial cell factories, industrial microbial chassis cells need to possess the characteristics of being able to utilize low-value renewable carbon sources, having high target productivity, high fermentation stability, being easy to scale up, and being amenable to rational design and optimization at the genomic level. Designing and constructing industrial microbial chassis cells using synthetic biology strategies involves organisms such as *Escherichia coli*, *Bacillus subtilis*, *Corynebacterium glutamicum*, yeasts (such as *Saccharomyces cerevisiae*, oleaginous yeasts, etc.), prokaryotic cyanobacteria, and eukaryotic microalgae. These cells are capable of efficiently utilizing first-generation organic carbon sources (such as glucose, etc.), second-generation sugar sources (such as lignocellulose, etc.), and third-generation one-carbon sources (such as one-carbon compounds like carbon dioxide, carbon monoxide, methane, etc.). In order to design and construct highly efficient industrial microbial chassis cells, it is necessary to address the metabolic regulation requirements of industrial microorganisms. Rational design and machine learning methods based on big data should be employed to design and construct bioparts and regulatory strategies that are compatible with the chassis cells. This enables the construction of highly efficient synthetic metabolic pathways and highly efficient microbial cell factories, achieving the high-yield synthesis of high-value products such as platform compounds and natural products [1-4].

3.5.2.3 Medical Microbial Chassis Cell Engineering

Compared with plant and animal cells, microorganisms are good chassis for synthetic biology design and construction of artificial biological cells due to their simple genetic composition, easy genetic manipulation, fast growth rate, and ease of mass culture. Among them, probiotics and living microalgae can be used as medical microbial chassis cells for disease diagnosis and precise and controlled drug delivery by incorporating functionalized biological system modules due to their natural promotion of human health [5,6]. In addition, some microorganisms have tumor tissue-tending colonization properties and natural tumor suppression effects, which have great potential for application in oncology diagnosis and treatment, and are good choices for chassis cells of tumor living cell drug factories [7,8]. However, the genetic stability and biosafety of chassis microbial cells are uncertain after their genomes have been optimized or new functionalized biological

system modules have been introduced. The creation of medical microbial chassis cells with low immunogenicity, tissue-targeting specificity, controllable fate, stable performance, and no impact on the ecological balance of the organism and the environment and on the development of biodiversity through synthetic biology is an important prerequisite for the development of microbial therapies and the integration of *in vitro* diagnosis and treatment of diseases.

3.5.2.4 Mammalian Chassis Cell Engineering

Mammalian cell engineering severely restricts the current development of the biomedical industry based on mammalian cells due to problems such as inefficient molecular genetic manipulation methods, lack of various tools, diverse cell types, stringent requirements for culture conditions, and high costs. Engineered mammalian cells are mainly applied in the medical and health fields related to biological macromolecules and *in vivo* drugs, and can be simply classified into two types. One type is the mammalian cell chassis used for producing drugs such as proteins and viruses. The other type is the living cell drug chassis used for delivering drugs in the body, eliminating or repairing lesions in the body [9].

For production-type mammalian cell chassis, including cell lines such as Chinese Hamster Ovary cell (CHO) and 293T, establishing simple and efficient cell molecular genetic manipulation techniques and tools, improving the efficiency of cell culture, reducing the cost of cell culture, and increasing the production of proteins and viruses are important goals for realizing the industrial value of the pharmaceutical industry. For living cell drug chassis, including human cells such as immune cells and stem cells, in order to achieve functions such as the delivery of protein drugs, the killing of target cells, and tissue repair, it is necessary to artificially enhance or externally regulate the behaviors of cell proliferation, differentiation, apoptosis, and movement, etc., on the premise of a deep understanding of the regulatory mechanisms of basic cell biology functions. Therefore, there is an urgent need to develop chassis cell technologies with low immunogenicity, the ability of allogeneic transplantation, and large-scale *in vitro* expansion [10].

3.5.2.5 Plant Chassis Cell Engineering

Plant cells including higher plant cells and lower algal cells have unique advantages

compared with other chassis cells. Plants can synthesize various complex metabolites through photosynthesis using only carbon dioxide and water as raw materials, without the need for a high-energy-consuming and high-oxygen-consuming fermentation process. The plant chassis can break through the limitations of heterotrophic microbial chassis, which have poor expression of cytochrome P450 enzymes and poor tolerance to active products. Plants are easy to be planted on a large scale indoors and in fields, and can continuously undergo asexual growth (non-flowering plants can produce new leaves infinitely) and reproduction, with a large biomass, making them suitable for large-scale production at low cost. Plants do not move, are easy to isolate, do not carry human pathogens, and have high safety. As a group of organisms with cellular diversity including single-celled organisms, multicellular colonies, simple or multicellular organisms with complex differentiation, the complex functional compartmentalization of the plant chassis itself, the fine division of labor and collaboration among different organs and tissues, and the special organs and tissues that have evolved to adapt to and utilize the environment (such as nitrogen-fixing root nodules and C4 leaf structures that efficiently carry out photosynthesis) all provide the possibility for the artificial design of complex functions. There are a large number of plant species, involving various types from lower plants to higher plants, which brings the possibility of diversified development of synthetic biological systems. However, currently, the genetic operation system has only been established in a small number of plants. Many plants grow relatively slowly, have complex ploidy, are difficult to regenerate, and contain diverse and complex metabolites, which pose some technical challenges for the engineering design and operation of these plant cells [11]. At present, research and development of plant chassis cells should be carried out on some model plants that are easy for genetic manipulation, have sufficient accumulation of basic research, grow rapidly, and have a simple genetic background (such as tobacco, rice, soybean, tomato, poplar, etc.).

3.5.3 Roadmaps

Current Status		
Core technologies and patents have been developed for key industrial microbial strains and mammalian cell lines, with expanded applications of microbial chassis cells in the medical field.		
Objective 1: Engineered Modification of Key Industrial Microbial Chassis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Engineer industrial microbial chassis cells to efficiently utilize diverse simple carbon sources.	• Modify cellular metabolic network according to the target demand, and construct different simplified chassis cells through genome reconstruction, editing and other technologies.	• Design chassis-compatible bioparts and regulatory strategies. • Establish microbial cell factories and the corresponding biomanufacturing systems (including next-generation bioreactors and process technologies, etc.).
Construct industrial microbial cell factories for high-efficiency synthesis of high-value compounds.	• Develop transcriptional/translational regulatory elements and genome-level editing tools for industrial microbial chassis. • Implement machine learning, evolutionary, and screening platforms for rate-limiting enzymes to optimize chassis cell metabolism and functional adaptation.	• Establish a computing and experimental platform for the design and construction of industrial microbial chassis cells, and achieve the whole-genome modification and artificial synthesis of chassis microbial cells. • Develop energy-metabolism coupling and redox balance technologies in microbial chassis cells to enhance directional synthesis of target metabolites. • Design and construct 15–20 distinct microbial chassis cells capable of pilot-and large-scale production in microbial cell factories.

Objective 2: Develop Universal Medical Microbial Chassis Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Establish a medical microbial chassis cell library.	• Explore human symbiotic and probiotic microbes with medical potential, tailored to diverse disease requirements, to build a foundational microbial chassis cell library.	• Engineer microbial chassis to develop simplified, low-immunogenicity, and biosafe living drug carriers with simple genomes.
Construct intelligent sensing microbial chassis cells.	• Develop safe, precise, high-sensitivity, and orthogonal genetic switches/biosensing systems, coupled with spatiotemporally controlled protein release mechanisms.	• Create engineered microbial chassis with tunable growth density, customizable protein expression/release profiles, and scenario-specific adaptability.

Objective 3: Develop Mammalian Cell Chassis for Industrial Fermentation and Cell Therapy		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Mammalian chassis cells for high-efficiency production.	• By integrating techniques such as genome engineering and promoter design, develop efficient cell lines for heterologous protein expression, significantly enhancing the efficiency and convenience of gene transfection, integration, and regulation, as well as improving the growth rate and anti-contamination capabilities of cells.	• Systematic optimize genetic and metabolic systems to create specialized fermentation chassis cells. • Implement serum-free culture and fermentation technologies to enhance safety and yield of protein-based drugs and viral vectors.
Living drug chassis cells for universal medical.	• Establish standardized genetic engineering platforms for cell modification. • Develop technologies to regulate fundamental cellular behaviors (e.g., motility, proliferation, differentiation, apoptosis). • Develop externally or autonomously controlled "smart" therapeutic cell chassis.	• Develop 2–3 categories of off-the-shelf living therapeutic cell chassis. • Achieve scalable production and amplification, *in vivo* precise targeting, and enhanced biosafety.

Objective 4: Development of High-Performance Plant Cell Chassis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop universal plant chassis.	• Develop genetic transformation and gene editing methods for high-efficiency plant chassis (including higher plants and algae cells). • Genetic enhance established plant chassis systems (e.g., soybean cotyledon cells, rice endosperm cells, tobacco trichome cells) to boost protein/metabolite synthesis and accumulation.	• Design and construct rapid-growth, scalable, and controllable universal plant chassis using next-generation genetic transformation and gene editing technologies. • Refine plant chassis systems for industrial-scale applications.
Plant chassis circuit engineering and cell factories	• Design and test plant genetic circuits. • Engineer protein expression, sorting, modification, and transport systems. • Precise metabolic pathway engineering.	• Develop application-specific genetic circuits for diverse scenarios. • Plant-based production of complex proteins and metabolites. • Design and construct 3–5 photosynthetic microalgal chassis cells and 8–10 high-value bioactive compounds, achieving pilot and large-scale production in microalgal cell factories.

Figure 1 Roadmap for chassis cell engineering

3.5.4 Technical Pathways

3.5.4.1 Engineering of Key Industrial Microbial Cell Chassis

Current Technologies: To construct efficient microbial cell factories and achieve the synthesis of high-value chemicals, high-throughput screening technologies should be developed to obtain efficient natural industrial microbial chassis cells. Moreover, through metabolic flux models at the whole-genome level, the metabolic pathways of microbial chassis cells should be rationally designed and constructed to obtain efficient industrial microbial cells that can utilize various carbon sources (such as glucose, xylose, lignocellulose, carbon dioxide, etc.). And cell factories for the high-yield synthesis of a variety of high-value products (such as platform compounds, energy products, natural products, protein drugs, etc.) should be constructed. In addition, different chassis cells should be developed, such as bacteria including *Escherichia coli*, *Bacillus subtilis*, *Corynebacterium glutamicum*, etc.; fungi including yeasts and actinomycetes, etc.; photosynthetic autotrophic microorganisms including eukaryotic microalgae, cyanobacteria, etc. Make full use of the treasure trove of species resources in nature and rationally transform industrial chassis microbial cells for different purposes.

Objectives and Breakthroughs: Engineer and transform industrial microbial chassis cells to efficiently utilize various carbon sources. Construct industrial microbial cell factories to efficiently synthesize various high-value compounds, including small-molecule products (natural products, chiral drugs, platform compounds, biofuels, etc.) and macromolecule products (protein drugs, enzymes, functional lipids, high-molecular polymers, etc.). Design and construct 15 to 20 different microbial chassis cells to achieve pilot-scale and large-scale production of microbial cell factories.

Challenges: Current microbial chassis cells lack metabolic pathways for the efficient utilization of various carbon sources. The light energy conversion efficiency of photosynthetic autotrophic microalgae and cyanobacteria, as well as the conversion efficiency of primary photosynthetic products into high-value-added compounds, are both relatively low. There is a lack of compatibility between the chassis cells and engineered bioparts, and there is a lack of disruptive biological manufacturing equipment and process technologies that are compatible with microbial cell factories.

There is a lack of abilities to assemble bioparts and dynamically regulate metabolic pathways in microbial chassis cells. The mismatch between the material and energy

metabolism in industrial microbial chassis cells limits the efficiency of product synthesis in microbial cell factories.

Expected Progress Recently: Reconfigure the cell metabolic network according to the target requirements, and construct different simplified chassis cells through technologies such as genome reconstruction and editing. Develop efficient genetic parts, and construct strategies for the design, assembly, and dynamic regulation of efficient and directional synthesis pathways of target products.

Expected Progress by 2030: Design bioparts and regulation strategies that are compatible with the chassis cells, and construct microbial cell factories. Strengthen the matching of material and energy metabolism and intracellular reducing power in the synthetic metabolic pathways of target products, and solve the compatibility problems between the exogenously introduced metabolic pathways and the chassis cells to achieve a reasonable distribution of intracellular resources, thereby achieving the efficient synthesis of target products.

Potential Solutions

Construct metabolic flux models at the whole-genome level and rationally design the metabolic pathways of microbial chassis cells for metabolic pathway design. Develop precise regulation technologies and strategies for metabolic pathways in microbial chassis cells. Study the perception and signal transduction of different wavelengths of light by microalgae and cyanobacteria chassis, and the influence of different light qualities and carbon sources on the intensity and efficiency of photosynthesis. Through the rational design of the central carbon metabolic pathway, promote the directional synthesis and accumulation of photosynthetic carbon assimilation into high-value-added target products.

Use rational design and machine learning methods based on big data to design and construct bioparts and regulation strategies that are compatible with the chassis cells to achieve efficient microbial cell factories, and establish an advanced biological manufacturing equipment and process system that is compatible with efficient microbial cell factories.

Develop corresponding regulation component libraries (at the promoter level, intermediate regulation level, and global regulation level) with intensity gradients and compatibility for chassis organisms, and achieve assembly and debugging. By establishing tools for the transcription and translation regulation elements of microbial chassis cells and genome-level editing tools, as well as tool platforms for machine

learning, evolution, and screening of key rate-limiting enzymes, achieve the physiological and metabolic transformation and functional adaptation of microbial chassis cells.

Build a computational platform and an experimental platform for the design and construction of industrial chassis microbial cells, complete the whole-genome modification and artificial synthesis of chassis microbial cells. Achieve the interaction simulation and two-way optimization between functional components and chassis cells. Develop technologies such as the coupling of energy and material metabolism and the balance of reducing power in microbial chassis cells to promote the efficient and directional synthesis of metabolites. Improve the design and modification technology system of biological systems based on synthetic biology.

3.5.4.2 Construction of Universal Medical Microbial Cell Chassis

Current Technologies: Currently, the application of microorganisms in the medical field is still concentrated on the fermentation of biological reactors to produce active functional substances. With the continuous deepening of the understanding of the relationship between the microbiome and human health, the use of microbial preparations to improve health conditions and treat diseases has become a research hotspot in recent years. The only microbial therapy currently approved by the FDA is fecal microbiota transplantation for the treatment of *Clostridioides difficile* infection, but it has not been widely promoted due to the lack of a clear mechanism of action. The rational modification and precise control of microbial cells using synthetic biotechnology is the key to use microorganisms for personalized disease treatment. Among them, the selection of medical microbial chassis cells is of vital importance. Most current studies use a limited number of model strains for engineering modification, and their safety and stability as living drugs are insufficient. In addition, in order to achieve tissue-targeted specificity of medical microorganisms and rapid and controllable drug release, technologies such as gene editing and metabolic engineering have been developed and applied to the directional modification of microbial cells. However, there is still a lack of highly universal, highly intelligent and controllable chassis cells and regulatory systems.

Objectives and Breakthroughs: Construct a library of medical microbial chassis cells. Construct intelligent sensing microbial chassis cells.

Challenges: Currently, most medical microbial cell chassis are limited to model microbial chassis, lacking universal and personalized medical microbial chassis. Microbial chassis have poor immunogenicity and unstable performance. Current gene expression

regulation is only limited to some common small molecule regulatory systems. Cell behavior has poor controllability and is easily interfered with by complex physiological conditions.

Expected Progress Recently: According to the needs of different types of diseases, excavate human symbiotic and probiotic microorganisms with medical application value, construct a library of chassis microbial cells, and develop a gene switch control system that is safe, precise, highly sensitive, and has good orthogonality, as well as a spatiotemporally controllable rapid protein release system.

Expected Progress by 2030: Engineer and modify the chassis microorganisms, and develop living drug carriers with a simple genome composition and low immunogenicity. Construct engineered microbial chassis cells with controllable growth density, adjustable protein expression and release, and personalized adaptation to usage scenarios.

Potential Solutions

Isolate and identify human symbiotic and probiotic microorganisms with medical application value and modification potential from human samples as chassis microorganisms.

Use gene editing and metabolic engineering technologies to modify the genome and metabolic network of chassis cells, simplify the endogenous network, and improve safety and stability.

Leverage bioinformatics screening and high-throughput screening technologies to excavate and construct an enabling technology component library, and develop biosensors for green and healthy small molecules, physiological markers, metabolites, and signals such as light, electricity, and magnetism.

Through the rational design and assembly of regulatory elements, introduce logic gate genetic circuits with multiple regulators. Combine computer simulation assistance with protein directed evolution technology to develop intelligent sensing microbial cells that integrate functions such as high stress resistance, high sensitivity in sensing physiological and metabolic indicators, the microenvironment of lesions, disease markers, and so on.

3.5.4.3 Development of Mammalian Cell Chassis for Industrial Fermentation and Cell Therapy

Current Technologies: Currently, the main application scenarios of mammalian cells are concentrated in the medical and health fields. Engineered chassis cells such as

CHO and 293F provide an important industrial foundation for the fermentation production of protein drugs such as antibodies and cytokines. In recent years, immune cells including mesenchymal stem cells, as well as CAR-T and CAR-NK, have become emerging technologies and directions for medical research and development against tumors, neurodegenerative diseases, and aging [12, 13].

Objectives and Breakthroughs: High-efficiency production-type mammalian cell chassis. Universal medical living drug chassis cells.

Challenges: There is a lack of various gene regulatory elements and an understanding of the molecular mechanisms of cell physiology, as well as a lack of efficient genetic manipulation techniques. The relationship and molecular mechanisms between the cell metabolic system and environmental nutrition are unclear. There is a lack of protein tool components, and the cellular biological mechanisms of chemotactic movement, proliferation, and differentiation of human primary cells are not yet clear. The understanding of the immunogenicity of cells is limited. There is a lack of the ability for large-scale cell culture and directional differentiation.

Expected Progress Recently: Combine technologies such as genome engineering and promoter design to develop cell lines with high-efficiency heterologous protein expression, significantly improve the efficiency and convenience of gene transfection, integration, and regulation, and enhance the cell's growth and anti-contamination capabilities. Form a standardized genetic manipulation technology platform for cell engineering modification, develop regulatory technologies for basic functions such as cell movement, proliferation, differentiation, and apoptosis, and develop intelligent cell drug chassis that can be externally and autonomously controlled.

Expected Progress by 2030: Systematically optimize the genetic and metabolic systems of cells, construct specific fermentation chassis cells for subdivided directions, develop serum-free cell culture and fermentation technologies, and greatly improve the production safety and yield of protein drugs, virus vectors, etc. Basically achieve large-scale production and expansion of 2-3 types of off-the-shelf living cell drug chassis, and achieve accurate *in vivo* targeting and safety and reliability.

Potential Solutions

Develop automated technologies for cell culture, design, and modification, design, construct, and identify promoters and terminators, as well as tool components for gene

recombination, integration, transcription, translation, and regulation. Identify and modify ideal genome integration sites. Develop a self-replicating plasmid system for mammalian cells.

Analyze the relationship between cell proliferation and nutrient metabolism. Combine high-throughput technologies and gene editing methods to systematically screen and identify cell lines that can proliferate and produce efficiently under different culture conditions, and artificially endow cells with controllable nutrient metabolism capabilities.

Develop key technologies for high-throughput and automated acquisition, culture, and modification of mammalian cells and human primary cells. Based on gene editing technology, design and identify the nutrient and cytokine conditions for cell culture on a large scale. Design artificial cytokines and artificial cytokine receptors to achieve orthogonal regulation of the cell proliferation and differentiation process.

Use high-throughput gene editing technology to systematically construct and identify major cell immunogens, analyze the laws of artificial cell *in vivo* localization, and design cell safety switches. Develop the theory and methods for designing gene circuits in mammalian cells, construct functional gene circuits, endow the cell chassis with the ability to recognize and adapt to the environment, and achieve intelligent, accurate, and artificially controllable anti-disease functions *in vivo*.

3.5.4.4 Development of High-performance Plant Cell Chassis

Current Technologies: There is a wide variety of plants, ranging from single-celled green algae to giant thousand-year-old redwoods. They possess tissues and organs with special functions to adapt to and utilize various natural environmental conditions, which provides numerous possibilities and potential for the development of plant chassis cells. However, compared with single-celled microbial systems, the development of genetic manipulation technologies in plant systems lags behind. The main problems are that the processes of genetic transformation and plant regeneration are relatively slow, and only a small fraction of plants can be transformed with exogenous genes and then regenerated. Many plants have problems such as long life cycles, complex ploidy, difficult transformation and regeneration, and diverse and complex developmental and metabolic traits. Therefore, there are many technical challenges in the engineering design and operation of these plant cells. In recent years, there have been many important advances in plant genetic modification and exogenous gene transformation research. For example, the efficient transient expression system in tobacco leaves or protoplasts has matured,

which can be used to quickly test a large number of different expression vectors, elements, and circuits. The stable introduction of exogenous genes via Agrobacterium is a well-established technique, which can be efficiently applied to various plants from lower plants (such as liverworts) to trees (such as poplars). More than 10 types of model plants (such as green algae, liverworts, tobacco, rice, soybeans, corn, rapeseed, poplars, etc.) have efficient genetic manipulation systems. Various gene-editing technologies based on the CRISPR method (including site-directed gene knockout, gene activation, and single-base editing) have been effectively applied in these model plants. However, some plants cannot be infected by *Agrobacterium* or cannot be regenerated, which limits the wider application of *Agrobacterium*-mediated gene introduction technology. High-efficiency single-base gene editing is quite difficult and is only applicable to a few plants. Currently, through interdisciplinary research, some new gene transformation methods have been developed, such as nanotube-mediated exogenous gene transformation and chloroplast engineering technologies [11, 14]. However, these methods have only been verified in a few plant systems, and the universality and efficiency of their application still need to be improved.

Objectives and Breakthroughs: Universal and efficient exogenous gene transformation, precise genetic manipulation, and plant regeneration technologies. Plant chassis circuit engineering and plant cell factories.

Challenges: Universal and efficient exogenous gene transformation and precise genetic manipulation technologies are the technical points that urgently need to be breakthroughs in this field. Basic research in plant cell biology, systems biology, biochemistry, etc., is still insufficient. For example, the mechanism by which individual proteins can be stably and abundantly accumulated in the cotyledons of leguminous plants is not yet clear. There is a lack of methods and platforms for large-scale standardized element testing. The diversity of plants themselves provides opportunities to solve these research problems, but more investment in basic research is still needed [15]. The potential problems of functional modification and assembly of complex heterologous drug proteins in plant cells have not been fully resolved. There is a lack of large-scale protein and small-molecule purification technologies and capabilities.

Expected Progress Recently: Establish a series of plant cell chassis systems suitable for the efficient production of heterologous drug proteins and plant-derived small-molecule drugs respectively. For example, use soybean cotyledon cells and tobacco leaf epidermal trichomes to produce proteins and small-molecule drugs respectively.

Construct a high-expression, correct-modification, and transport system for heterologous proteins, a high-expression and transport system for enzymes related to metabolic pathways, and establish an organelle system (such as plastids) and a transport system (such as vacuoles) for small-molecule drug synthesis.

Expected Progress by 2030: Basically achieve the production of proteins and small-molecule drugs using two types of off-the-shelf plant cell chassis respectively, enabling large-scale production and purification, and complete the testing of drugs in model systems and initial clinical trials.

Potential Solutions

Leverage the advantages of disciplines such as nanomaterials to develop a new generation of high-efficiency exogenous gene transformation methods. Conduct in-depth research on the mechanism of Agrobacterium infection of plants, and identify the factors that limit the general infection of plants by Agrobacterium to achieve the universal application of this gene transformation technology. Elucidating this mechanism will overcome the technical bottleneck of effective plant regeneration.

Use transient expression tools to test a series of plant cell systems with potential for the efficient production of heterologous drug proteins and plant-derived small-molecule drugs. Select two cell chassis respectively, and improve the production capacity and efficiency of these chassis cells through stable genetic modification. Combine transient and stable expression systems to strengthen the research on large-scale development, testing methods, and standards of elements in different plant chassis. Strengthen basic research in plant cell biology and secondary metabolism biochemistry.

Use the transient expression system to test the expression and function of various mammalian cell protein-modification and assembly elements in plant cells. Carry out large-scale purification research and platform construction.

3.5.5 Summary

Chassis cells are the fundamental part for exercising the functions of synthetic biology. Systematically carrying out the engineering transformation of chassis cells at different levels and for different application outlets, such as microorganisms, plants, and mammalian cells, is of great scientific and engineering significance. These studies not

only require the integration of new theories and technologies from interdisciplinary research but also a full understanding of elements such as cell growth, proliferation, division, and differentiation in order to effectively construct and optimize chassis cells. Technological breakthroughs in chassis cells can also provide research paths and technical means for exploring the design of artificial cells and understanding the connotations of life and non-life.

References

[1] Srinivasan P, Smolke C D. Biosynthesis of medicinal tropane alkaloids in yeast. Nature, 2020, 585(7826): 614-619.

[2] Keasling J, Garcia Martin H, Lee T S, et al. Microbial production of advanced biofuels . Nat Rev Microbiol, 2021, 19(11): 701-715.

[3] Courdavault V, O'connor S E, Jensen M K, et al. Metabolic engineering for plant natural products biosynthesis: new procedures, concrete achievements and remaining limits. Nat Prod Rep, 2021, 38(12): 2145-2153.

[4] Liew F E, Nogle R, Abdalla T, et al. Carbon-negative production of acetone and isopropanol by gas fermentation at industrial pilot scale. Nat Biotechnol, 2022, 40(3): 335-344.

[5] Steidler L, Hans W, Schotte L, et al. Treatment of murine colitis by *Lactococcus lactis* secreting interleukin-10. Science, 2000, 289(5483): 1352-1355.

[6] Riglar D T, Silver P A. Engineering bacteria for diagnostic and therapeutic applications. Nat Rev Microbiol, 2018, 16(4): 214-225.

[7] Zhou S, Gravekamp C, Bermudes D, et al. Tumour-targeting bacteria engineered to fight cancer. Nat Rev Cancer, 2018, 18(12): 727-743.

[8] Zhong D, Zhang D, Chen W, et al. Orally deliverable strategy basedon microalgal biomass for intestinal disease treatment . Sci Adv, 2021, 7(48): eabi9265.

[9] Cubillos-Ruiz A, Guo T, Sokolovska A, et al. Engineering living therapeutics with synthetic biology. Nat Rev Drug Discov, 2021, 20(12): 941-960.

[10] Mansouri M, Fussenegger M. Therapeutic cell engineering: designing programmable synthetic genetic circuits in mammalian cells. Protein Cell, 2022, 13(7): 476-489.

[11] Liu W, Stewart C N. Plant synthetic biology. Trends Plant Sci, 2015, 20(5): 309-317.

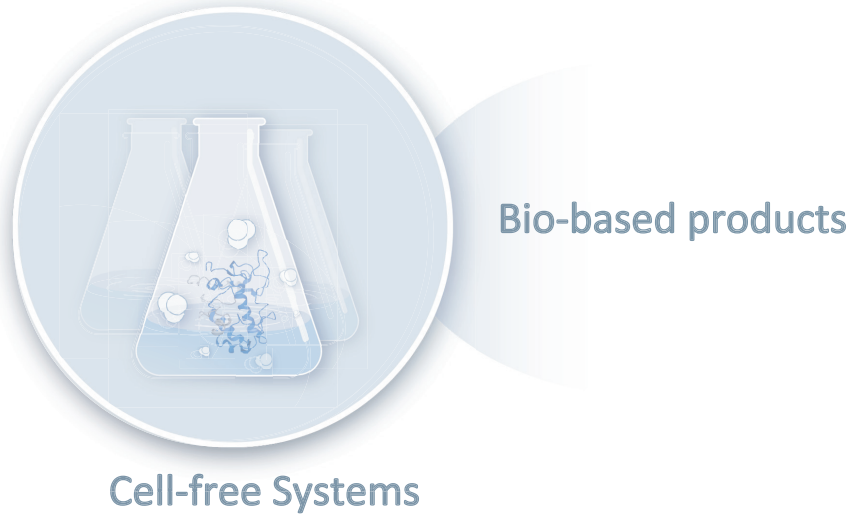
[12] Kitada T, Diandreth B, Teague B, et al. Programming gene and engineered-cell therapies with synthetic biology. Science, 2018, 359(6376): eaad1067.

[13] Saez-Ibanez A R, Upadhaya S, Partridge T, et al. Landscape of cancer cell therapies: trends and real-world data. Nat Rev Drug Discov, 2022, 21(9): 631-632.

[14] Wright R C, Nemhauser J. Plant synthetic biology: Quantifying the "known unknowns" and discovering the "unknown unknowns". Plant Physiol, 2019, 179(3): 885-893.

[15] Andres J, Blomeier T, Zurbriggen M D. Synthetic switches and regulatory circuits in plants. Plant Physiol, 2019, 179(3): 862-884.

Cell-free Biosystems



Authors Wang Qin-Hong, You Chun, Lu Yuan, Li Jian, Shi Jia-Fu

3.6 Cell-free Biosystems

3.6.1 Abstract

Cell-free biosystems are enabling technologies that utilize essential catalytic components or cell lysates to perform complex biotransformations. They hold significant potential for elucidating fundamental biological principles and advancing synthetic biomanufacturing, with promising applications in food, pharmaceuticals, sensing, and materials. However, challenges such as insufficient performance of key components, lack of standardized preparation methods for cell lysates, product-induced reaction inhibition, and limited large-scale datasets hinder their industrial scalability. To address these issues, advancements are needed in optimizing component performance, enhancing compatibility between system elements, and establishing standardized, engineered workflows.

3.6.2 Technical Overview

Cell-free biosystems reconstitute biological processes *in vitro* by assembling enzymes, cofactors, and other active components outside living cells and carrying out complex physiochemical reactions. These systems are emerging as powerful tools to understand, harness, and expand the capabilities of natural biological systems. They are broadly categorized into multi-enzyme cascade-based cell-free biosystems and cell lysate-based cell-free biosystems.

3.6.2.1 Multi-enzyme Cascade-based Cell-free Systems

Multi-enzyme cascade-based cell-free systems design artificial biocatalytic pathways and assemble enzymes and enzymatic modules *in vitro* to execute complex biochemical reactions for synthetic biomanufacturing. Compared to synthetic biomanufacturing systems based on living microbial cells, multi-enzyme cascade-based cell-free systems offer advantages such as fewer side reactions, higher product yields, faster reaction rates, simplified product purification, enhanced environmental tolerance, and greater operational flexibility. With the continuous development of biological big data, new enzyme catalytic pathways and enzyme catalytic functions have been identified, the cell-free systems have shown increasing competitiveness in synthetic biomanufacturing.

Nevertheless, several critical bottlenecks remain: leveraging big data and AI to predict novel enzymatic pathways and functions; rapid identification of optimal enzyme combinations and loadings for multi-enzyme pathways; improving enzyme stability, activity, and scalable production; enhancing coenzyme stability in energy-dependent reactions and mitigating substrate and toxic product inhibition to the system.

3.6.2.2 Cell Lysate-based Cell-free Systems

Cell lysate-based cell-free biosystems leverage cell extracts to perform transcription, translation, and other processes for *in vitro* protein synthesis, enabling efficient preparation of membrane proteins and toxic proteins. These systems are widely used to validate genetic circuits, prototype metabolic pathways, develop portable diagnostics, and manufacture biomacromolecules (e.g., antibodies). They have also contributed to foundational discoveries in the life sciences, such as deciphering triplet codon genetic coding. However, key challenges persist: establishing reproducible protocols for lysate preparation and improving lysate stability; building large-scale datasets and quantitative models to correlate cell-free and cellular system behaviors; designing genetically encoded biosensors for portable, on-demand synthesis applications to address food, water, and energy security; producing complex proteins such as glycoproteins and advancing high-efficiency production of composite formulations and vaccines.

3.6.3 Roadmaps

Current Status		
Literature-based enzyme databases are complete. Retrosynthetic analysis is mature. Several pathway design softwares have been developed, using experimental data and mathematical models for system adaptation of enzyme ratios and dosages and mainly using mathematical models. The stability of the enzyme is mainly improved by gene mining and engineering. Substrate inhibition is reduced and reaction efficiency is improved by means such as protein confinement and the construction of multienzyme complexes. Coenzyme regeneration is mainly achieved through substrate phosphorylation and the use of substances including glucose.		
Objective 1: Intelligent design of catalytic pathways for multi-enzyme cascades		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Promote intelligent design of enzymatic pathways.	• Develop natural language processing and machine learning approaches to acquire new enzymatic pathways. • Design novel and unique multi-enzyme catalytic pathways by integrating more novel enzymes.	• Use artificial intelligence to predict enzyme function from sequence. • Use artificial intelligence to guide enzyme engineering and design novel pathways for multi-enzymatic cascades.

Objective 2: Rapid adaptation of catalytic pathways for multi-enzyme cascades		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enable rapid pathway adaptation.	• Quantitative modeling of enzyme catalysis and rapid identification of kinetic parameters of enzymes. • Resolve quantitative relationships between enzyme and substrate promiscuity.	• Develop reaction pathway optimization tools to support multi-enzyme pathway adaptation processes. • For multiple-enzyme cascades, develop methods for rapid determination of enzyme combinations and rapid adaptation of enzyme loading amounts. The conversion yields of 2-3 products are over 90% in the industrial setting.
Objective 3: Novel hybrid materials and co-immobilization methods for multi-enzymes		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Extended running time.	• Build high-performance, low-cost, customizable immobilized multi-enzyme systems. • Precise immobilization and controlled intelligent assembly of enzyme molecules on the surface and in the pores of carrier materials.	• Reveal the assembly process-structure- activity evolution, and guide the optimization of multi-enzyme immobilization, and the immobilized multi-enzyme system can be reused more than 50 times. • Construct customized pore channels with integrated functions of substrate enrichment and efficient delivery.

Objective 4: Reduction of by-product formation and product inhibition of the system		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Improve operational efficiency.	• Develop multi-enzyme catalytic systems for phosphorus equilibrium. • Develop product displacement and dialysis strategies to remove inhibitors and products from the system. • Design novel enzymes that are not inhibited by by-products and intermediates.	• Resolve the mechanism of enzyme instability. • Design methods that spatially isolate catalysts from inhibitors. • Develop a reactor design and bioprocess for continuous catalyst replenishment, enabling stable operation of the system for more than 30 days.
Objective 5: Develop metabolic modules for continuous regeneration of energy and reduce power to enhance operational capacity		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance operational capacity.	• Realize regeneration of water-dissolved oxygen-coupled energy carriers and reducing power. • Construct heterojunction photoelectrocatalysts for enhanced interfacial electron transfer.	• Regenerate ATP using inorganic phosphate ions. • Use artificial cofactor to achieve the redox capacity to the levels of natural cofactor. • Hybrid enzyme enables the internal recycling of the coenzyme within its structure.

Figure 1 Roadmap for multi-enzyme cascade-based cell-free systems

Current Status
The production of cell lysates from several high-utilization model chassis has been achieved, including prokaryotic cells (*Escherichia coli*), yeast (*Saccharomyces cerevisiae*), plant cells (wheat germ), animal cells (rabbit reticulocytes, Chinese Hamster Ovary cells), and insect cells (*Spodoptera frugiperda*). Additionally, limited progress has been made in generating lysates from low-utilization non-model chassis. However, *E. coli*-based lysates remain the most widely used system. Effective operation of cell-free systems has been enabled through component selection/optimization and energy system optimization, while batch and continuous-operation bioreactors have been developed to mitigate system inhibition and achieve moderate scalability in cell-free protein synthesis. Post-modifications, such as precision glycosylation, have been explored to meet functional protein requirements.

Objective 1: Enable standardization and customized design of cellular extraction systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Standardize and customize the design of the system.	• Design a standardized cell preparation and extract preparation technology system that can prepare more than 500 ml of cell extracts in a single batch. • Expand library of transcription and translation elements for cell-free systems.	• Construct cell-free systems that allow for post-modification of proteins, such as definable glycosylation, and obtain to more than 10 post-modification reaction systems.

Objective 2: Enhancing the operational efficiency of cell extraction systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Improve stability of cell-free extracts.	• Systematically modify and optimize of enzyme components to extend the half-life of transcription and translation processes, and the system can be run stably for more than 5 days. • Construct highly efficient and stable circular or linear DNA templates.	• Develop new methods for efficient synthesis of complex protein. • Achieve hightly efficient synthesis of natural proteins and efficient orthogonal synthesis of non-natural proteins. • The system can be run stably for more than 10 days.

Objective 3: Develop and design different types of reaction devices and reaction environments to expand the scope of application of the system		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Expand the scope of application of the system.	• Design reactors for efficient gas-liquid mass transfer, component synergy, and batch reaction volumes up to 10-100 L. • Build portable on-demand, cell-free synthesis systems.	• Enhance efficient and correct synthesis of proteins, realizing industrial-scale efficient continuous protein synthesis mode. • Construct intelligent cell-free systems that respond to physical signals such as light, heat, electricity, and magnetism.
Objective 4: Develop the robustness of the applicable scale-up production system and improve the system operation capability		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Enhance system operation capacity.	• Develop large-volume high-pressure crushing technology to obtain large-volume cell extracts. • Prepare sufficient quantities of gene templates using industrialized or scaled-up plasmids to achieve gram-level yields.	• Construct *in vitro* scale-up reaction systems for gene templates. • Design continuous bioreactors. • Construct tunable and sustainable compartmentalized systems.
Objective 5: Improve protein post-translational modification		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Improve protein post-translational modification *in vitro*.	• Construct endotoxin-free cell lysates. • Integrate related glycosylase genes into the *E. coli* genome. • Flexibly select different strains from the strain library to prepare cell lysates.	• Screen and modify glycosylase function. • Control the spatiotemporal regulation of enzyme catalysis with oligosaccharide substrates and others. • Design modular cell-free systems to synthesize glycosylated proteins.

Figure 2 Roadmap for cell lysate-based cell-free systems

3.6.4 Technical Pathways

3.6.4.1 Multi-enzyme Cascade-based Cell-free Systems

Current Technologies: Multi-enzyme cascade-based cell-free systems involve the design of enzymatic pathways, functional exploration and screening of enzyme genes, compatibility and stability optimization of multi-enzyme systems, and stabilization/ regeneration of cofactors. These systems have enabled high-efficiency synthesis of numerous products, including vitamins [1], rare sugars [2], starch [3], and pharmaceuticals [4], with industrial-scale applications such as myo-inositol production from starch.

In terms of enzyme databases, there are literature-based databases such as KEGG, MetaCyC, BioCyC, Brenda, Uniprot, etc. Several design software programs are available for pathway retrosynthesis analysis, such as BlastKOALA, KAAS, GhostKOALA, and RAST. In terms of system adaptation, experimental data and mathematical models are used for system adaptation of enzyme ratios and dosages, and protein restriction domains and multi-enzyme complexes are used to reduce substrate inhibition and improve reaction efficiency. In terms of system stability, on the one hand, the stability of enzymes is improved through gene mining, protein design, directed evolution, and enzyme immobilization, and on the other hand, the stable encapsulation of enzyme molecules and the precise regulation of substance transport are achieved through the use of *in situ* encapsulation strategies based on novel materials with porous frameworks, porous networks, and other structures [5,6]. For cofactors, substrate phosphorylation and natural vesicles [7] could be used for coenzyme regeneration, and artificial coenzymes were utilized for orthogonal reactions [8,9]. In addition, regeneration ATP and NAD(P)H can be developed using polysaccharide substances such as starch in phosphate-balanced modes, and these cofactors can be regenerated using clean energy sources such as light and electricity [10].

Objectives and Breakthroughs: Build datasets of enzymatic reaction equations to enable intelligent pathway design. Establish quantitative enzymatic models for rapid adaptation of multi-enzyme cascade. Develop new hybrid materials and multi-enzyme co-immobilization methods to extend operational duration. Minimize byproduct formation and product inhibition to enhance efficiency. Engineer metabolic modules for regeneration of ATP and reducing power to enhance operating capacity.

Challenges: Current enzyme databases contain validated pathways, but it is difficult

to include enzymes with unvalidated functions. The performance of pathway design software is strictly dependent on the quality and quantity of enzymes available. It remains difficult to predict reaction types and kinetic constants from sequences. There is a lack of standardized parameters for enzymes and a common standard for enzyme databases.

It is difficult in precise immobilization and assembly of enzyme components during multi-enzyme catalysis. There is a lack of structure-activity interplay modeling during the assembly of enzyme and carrier; poor diffusion-reaction kinetics fit; accumulation of by-products (inorganic phosphates) leads to precipitation of metal ions and affects the efficiency of the system; enzyme instability is a major obstacle to the development of technology in this field; low regeneration efficiency of ATP and reducing coenzymes; and natural coenzymes are easily degraded.

Expected Progress Recently: Develop high-performance methods for precise immobilization of multi-enzyme systems. Rapidly identify enzyme cascade pathways for specific products. Establish quantitative enzymatic models for fast kinetic constant determination. Mitigate byproduct generation and product inhibition. Develop light/ electricity-driven systems for efficient ATP and coenzyme regeneration.

Expected Progress by 2030: Establish standardized datasets and repositories for enzymatic reaction equations. Achieve >90% conversion yields in industrial settings through rapid enzyme combinations or adaptation of enzyme ratios. Elucidate mechanisms of immobilized enzyme processes and reaction-transport coordination, enabling >50 reuse cycles. Develop ultra-stable catalysts and confinement technologies for systems operational for >5 days. Resolve the issues of coenzyme stability.

Potential Solutions

Apply natural language processing (NLP) and machine learning (ML) to predict unvalidated enzymatic pathways and incorporate predicted enzyme functions. Use artificial intelligence method to predict enzyme functions from sequences and guide enzyme engineering for multi-enzyme pathways that contain hyperthermostable enzymes. Characterize enzyme kinetics and substrate selectivity to define enzyme-substrate promiscuity rules. Develop pathway-specific optimization tools and accessible computational infrastructure. Achieve precise enzyme immobilization on carrier materials and controllable enzyme-carrier assembly. Leverage *in situ* analytical techniques to establish structure-activity correlations. Engineer customizable porous frameworks for

functional integration. Design inhibition-resistant enzymes and phosphorus-balanced multi-enzyme systems. Implement *in situ* product removal and dialysis strategies. Investigate enzyme instability mechanisms and develop reactors with continuous catalyst replenishment. Enhance water-splitting-coupled energy and reducing power regeneration. Utilize light and electricity for efficient ATP regeneration via proton gradients. Engineer enzymes to efficiently utilize artificial coenzymes at parity with natural counterparts. Develop hybrid enzymes for internal cofactor recycling.

3.6.4.2 Cell Lysate-based Cell-free Systems

Current Technologies: Cell lysate-based cell-free systems have demonstrated significant potential in genetic circuit research, protein engineering, artificial life system construction, and synthesis of complex natural products and sustainable chemicals. Current research primarily focuses on *E. coli* lysate systems, which remain the most widely used platform [11]. The introduction of precision glycosylation pathways into *E. coli* lysate systems enable precise modulation of therapeutic protein activity [12]. Diverse lysate systems are emerging, including those derived from *Streptomyces* spp., *Bacillus subtilis*, *Corynebacterium glutamicum*, and *Vibrio natriegens*. Beyond protein synthesis, these systems have been applied to construct *in vitro* metabolic pathways for producing bioactive natural products and pharmaceutical intermediates [13]. In process design, batch and continuous-exchange reaction systems have been optimized using spatial-temporal strategies (e.g., hydrogel confinement) and physical controls (e.g., light, temperature, magnetism) to precise regulate transcription/translation. Meanwhile, innovations like tube-in-tube microreactor systems further enhance protein synthesis capabilities [14].

Objectives and Breakthroughs: Standardize and customize systems using diverse model/non-model host cell-free lysates. Improve system efficiency by replacing unstable components and enhancing cell-free lysate stability. Expand application scope through novel reactor designs and environmental controls. Enhance system operational capabilities by developing robustness suitable for scalable production. Improve system applicability by engineering host cells via gene editing to enable advanced post-translational modifications (e.g., glycosylation).

Challenges: Variability in lysate preparation protocols and component formulations leads to inconsistent performance and comparability issues. Lack of standardized transcription/translation components. Challenges in precise synthesis and efficient post-translational modifications (PTMs) of proteins/unnatural proteins. Instability of

transcription/translation machinery and limited capacity to engineer core components (e.g., ribosomes, tRNAs). Poor mass transfer efficiency and controllability and limited portability for on-demand synthesis. Constraints in "one-pot" synthesis and intelligent spatiotemporal control. High costs and technical barriers in large-scale lysate preparation (e.g., gram-level DNA template production). Reduction of costs for large-scale production. Self-replication of gene templates. Continuous production for days or weeks. Endotoxin contamination in *E. coli* lysates affecting glycoprotein quality. Absence of endogenous glycosylation enzymes in *E. coli* and challenges in precise control of protein glycosylation and rare glycosylation pathway engineering.

Expected Progress Recently: Develop standardized prokaryotic/eukaryotic cell-free systems capable of single-batch lysate preparation exceeding 500 ml. Achieve stable cell-free transcription/translation systems operational for >5 days. Optimize momentum, heat, and mass transfer in reactors scalable to 10-100 L. Implement systems supporting diverse glycosylation modifications.

Expected Progress by 2030: Establish more than 10 personalized cell-free systems with customized PTM capabilities. Enable stable, high-efficiency protein synthesis systems operational for >10 days. Deploy industrial-scale, intelligent continuous protein synthesis platforms. Construct sustainable large-scale cell-free production systems at 1,000 L capacity. Achieve synthesis of proteins with arbitrary glycosylation patterns.

Potential Solutions

Establish standardized lysate preparation protocols with defined parameters. Expand libraries of transcription/translation components. Build synthetic platforms for precise protein post-translational modifications. Engineer protein translation machines (e.g., ribosomes and tRNAs) to specifically recognize unnatural amino acids and efficiently embed them into proteins. Construct efficient and stable DNA templates to reduce reaction system inhibitors or endotoxins. Modify and synthesize core translation machinery components such as ribosomes and tRNAs. Design batch reactors (10-100 L) and integrate modules that can be adapted to different requirements and scales. Develop simple and inexpensive technologies such as freeze-drying and paper-based carriers. Construct compartmentalized systems, and integrate genetic circuits with material components to enable the system to respond to physical signals (e.g., light, magnetism), achieving controllability and sustainability. High-cell-density fermentation of cells is

conducted using fermenters with a working volume of 10-15 L or larger. Large volume, high pressure disruption technology is employed to obtain a large amount of cell lysates.Use industrialized or large-scale plasmid preparation columns. Use low-cost materials for energy supply. Develop modules that can be integrated with fermentation tanks for real-time product separation/inhibitor removal. Develop DNA replicase-containing lysates for template amplification. Construct scaled-up reaction systems for gene templates. Design continuous bioreactors. Construct compartmentalization systems for regulation and sustainability. Knock out endotoxin synthesis genes or purify lysates to eliminate endotoxins. Integrate glycosylation enzyme genes into *E. coli* genomes and curate a library of 10+ glycosyltransferases for individualized protein expression and post-translational modification. Screen/optimize glycosylation enzymes, and screen/optimize rare glycosylation pathways and embed related genes into *E. coli.* Design modular systems to achieve the combination of more than 10 modular systems for enrichment followed by modification.

3.6.5 Summary

The development of cell-free systems can provide important platform technologies for genetic circuit design, biosensing, biomanufacturing, and the construction of artificial cells. By developing enzyme datasets and artificial intelligence technologies, standardizing the preparation of cell extracts, improving the performance of enzyme components and cofactors, and reducing the inhibitory effect of products, the efficiency of cell-free systems can be enhanced, making them standardized, and enabling them to become a platform for biomanufacturing on a par with cell-based synthetic systems.

References

[1] You C, Shi T, Li Y, et al. An *in vitro* synthetic biology platform for the industrial biomanufacturing of myo-inositol from starch. Biotechnol Bioeng, 2017, 14: 1855-1864.

[2] Li Y, Shi T, Han P, et al. Thermodynamics-driven production of value-added d-allulose from inexpensive starch by an *in vitro* enzymatic synthetic biosystem. ACS Catal, 2021, 11(9): 5088-5099.

[3] Cai T, Sun H, Qiao J, et al. Cell-free chemoenzymatic starch synthesis from carbon dioxide. Science, 2021, 373(6562): 1523-1527.

[4] Huffman M A, Fryszkowska A, Alvizo O, et al. Design of an *in vitro* biocatalytic cascade for the manufacture of islatravir. Science, 2019, 366(6470): 1255-1259.

[5] Liang K, Ricco R, Doherty C M, et al. Biomimetic mineralization of metal-organic frameworks as protective coatings for biomacromolecules. Nature Communications, 2015, 6(1): 7240.

[6] Shieh F K, Wang S C, Yen C I, et al. Imparting functionality to biocatalysts via embedding enzymes into nanoporous materials by a *de novo* approach: Size-selective sheltering of catalase in metal organic framework microcrystals. Journal of the American Chemical Society, 2015, 137(13): 4276-4279.

[7] Miller T E, Beneyton T, Schwander T, et al. Light-powered co2 fixation in a chloroplast mimic with natural and synthetic parts. Science, 2020, 368(6491): 649-654.

[8] Zachos I, Doring M, Tafertshofer G, et al. Carba nicotinamide adenine dinucleotide phosphate: Robust cofactor for redox biocatalysis. Angew Chem Int Ed Engl, 2021, 60(26): 14701-14706.

[9] Zhang L, King E, Black W B, et al. Directed evolution of phosphite dehydrogenase to cycle noncanonical redox cofactors via universal growth selection platform. Nat Commun, 2022, 13(1): 5021.

[10] Zhang S, Shi J, Sun Y, et al. Artificial thylakoid for the coordinated photoenzymatic reduction of carbon dioxide. ACS Catal, 2019, 9(5): 3913-3925.

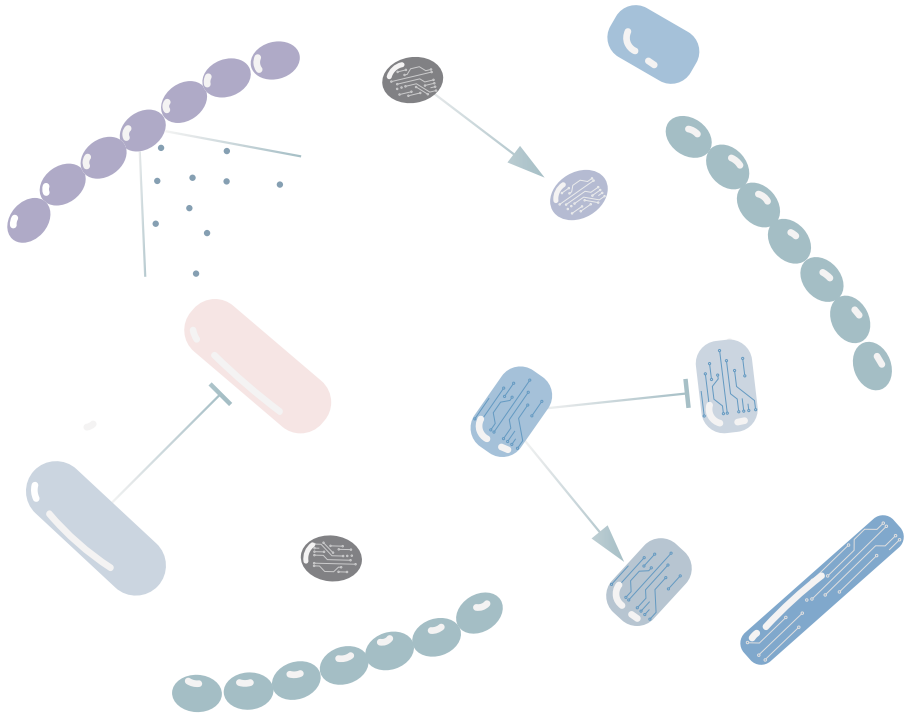
[11] Swartz J R. Expanding biological applications using cell-free metabolic engineering: An overview. Metabolic Engineering, 2018, 50: 156-172.

[12] Stark J C, Jaroentomeechai T, Moeller T D, et al. On-demand biomanufacturing of protective conjugate vaccines. Science Advances, 2021, 7(6): eabe9444.

[13] Tian X, Liu W Q, Xu H, et al. Cell-free expression of no synthase and p450 enzyme for the biosynthesis of an unnatural amino acid l-4-nitrotryptophan. Synthetic and Systems Biotechnology, 2022, 7(2): 775-783.

[14] Zhou C, Lin X, Lu Y, et al. Flexible on-demand cell-free protein synthesis platform based on a tube-in-tube reactor. Reaction Chemistry & Engineering, 2020, 5(2): 270-277.

Artificial Multicellular Systems



Authors Li Yin, Wu Xiao-Lei, Huang Jian-Dong, Dai Lei, Zhu Hua-Wei

3.7 Artificial Multicellular Systems

3.7.1 Abstract

Artificial multicellular systems represent critical technological systems in synthetic biology, enabling the transition from simplicity to complexity and from understanding nature to redesigning it. The primary objectives of this field are to establish theoretical foundations for artificial multicellular systems, develop technologies for design, construction, and functional testing, quantitatively characterize the relationships between individual and collective behaviors within these systems as well as their interactions with the environment, and predict novel phenomena and functionalities emerging from the progression of single-cell to multicellular architectures. These efforts aim to lay the technical groundwork for its applications in medicine, healthcare, industry, agriculture, energy, and environmental protection.

3.7.2 Technical Overview

Artificial multicellular systems are engineered biological systems made up of functionally distinct cells, which are broadly classified into two types. One type is derived from multiple species, which mainly exists the interactions between cells and compounds, cells and cells, and cells and the environment, playing an important role in areas like industry, agriculture, energy, environment and healthcare. The other refers to the heterogeneous multicellular systems derived from a single species, as well as the systems in which new functions emerge when the number of individual cells reaches a certain level. There mainly exist interactions between cells and cells, as well as between cells and tissues, playing crucial roles in biomedical applications.

The theories and enabling technologies of artificial multicellular systems consist of two main categories. The theoretical foundations involve understanding new biological phenomena, functionalities, and principles during the shift from artificial single-cell to multicellular systems, setting up design, construction, and operational principles for artificial multicellular systems, and clarifying the theoretical and control mechanisms that govern the spatial organization and spatiotemporal cell distribution within these systems. The enabling technologies include investigating signaling and communication mechanisms between cells and cells, cells and the environment, studying functional

specialization among cells of different roles or species to enable the construction, regulation, study and prediction of functional differentiation and collective behaviors, and developing technologies to ensure system stability and robustness across various environments, which includes rational design for multicellular systems, targeted addition or removal of specific cells, regulation of community structure formation, modulation of cell communication and metabolic capabilities, and modeling prediction.

From this, four objectives of the artificial multicellular systems are determined: the theory and design of the artificial multicellular system; the construction and testing of the artificial multicellular system; the functional research of the artificial multicellular system; and the environmental impact of the artificial multicellular system. Artificial multicellular systems are also classified into three types according to internal connectivity. Type I systems are mainly connected via metabolites (chemical compounds) and are designed for biodegradation or biosynthesis applications in industry, agriculture, energy, and environment. Type II systems are linked through metabolites and prokaryotic-eukaryotic cell interactions and are engineered to display specific physiological or immune functions for healthcare applications. Type III systems are governed by cell-cell and cell-tissue interactions, capitalizing on emergent biological functionalities for biomedical applications.

3.7.3 Roadmaps

Current Status
Based on the knowledge of biochemical pathways and the metabolic capabilities of species, metabolic division of labor communities have been constructed using natural strains, achieving the biodegradation of hardly degradable molecules such as aromatic hydrocarbons, brominated aromatic compounds, and cellulose, as well as the biosynthesis of high value compounds such as taxanes and anthocyanins. The high-throughput kChip technology and the eVOLVER parallel chemostat culture device have been developed to study the interactions between co-cultured microorganisms and the dynamics of synthetic communities. An in situ gene editing technology combining ET-seq and DART technologies has been developed, enabling the gene editing of specific microorganisms in model communities.

Objective 1: Theory and Design of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Rational design principles for multicellular systems with metabolic and synthetic functionalities.	• Modulize biochemical pathways to establish databases linking compounds, degradation or synthesis modules, and functional microbial strains. • Establish the basic theory and fundamental principles of metabolic division of labor, and realize the rational design of metabolic division of labor for specific compounds. • Propose the theory of species stability and realize rational design of genetically and functionally similar simple multicellular systems. • Simulation, prediction, and directed evolution of multicellular systems based on interactions, metabolic traits, and environmental conditions.	• Computerized screening and matching optimization of "compound-degradation or synthesis module-functional strains" correspondence according to the types and properties of degradation products and synthetic products. • Establish the theory of metabolic division of labor design that includes thermodynamic, kinetic and environmental impact parameters, and design artificial multicellular systems capable of degrading or synthesizing complex compounds. • Improve the theory of species stability, and realize the rational design of complex multicellular systems that are not easily compatible with genetics and functions. • Develop artificial intelligence technology to realize intelligent optimization of complex parameters and accurate prediction of aggregation of different artificial multicellular systems.

Objective 2: Construction and Testing of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Evaluation techniques for the construction of multicellular systems with substance metabolism and compound synthesis capabilities.	• Establish databases for strain growth conditions and metabolite distribution, develop cultomics, establish culture conditions and technologies that are compatible with different species, and achieve the rapid construction of artificial multicellular systems. • Develop high-throughput co-culture technology for artificial multicellular systems and parallel culture devices. • Develop the spatial assembly technology based on 3D printing and other methods. • Develop autonomous regulation technology based on single signaling molecule to establish close connection between species and enhance the population stability and function of simple multicellular systems.	• Establish a strain resource library based on degradation or synthesis modules. According to the degradation or synthesis ability of these modules, realize the modular and functionalized rapid construction and evaluation of artificial multicellular systems. • Develop parallel automatic culture devices for intelligent artificial multicellular systems at different scales. • Achieve the spatial self-assembly of artificial multicellular systems on special material carriers. • Develop autonomous regulation technology based on multiple signaling molecules. Strengthen the interactions through directed evolution and enhance the population stability and function of complex multicellular systems.

Objective 3: Functional Analysis of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
In situ regulation of multicellular systems with substance metabolism and compound synthesis capabilities.	• Develop orthogonal gene manipulation techniques applicable to *in situ* regulation of specific species and specific functions in complex communities. • Develop metabolite analysis techniques to realize the monitoring of metabolic division of labor.	• Improve the universality of *in situ* editing objects and realize broad-spectrum, molecular tools and phage-based *in situ* editing technologies for species and genes. • Develop metabolic flux analysis and regulation technologies, and realize *in situ*, real-time, dynamic monitoring and regulation of metabolic division of labor in complex multicellular systems.
Objective 4: Environmental Impact of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Biosafety management of multicellular systems with substance metabolism and synthesis capabilities.	• Conduct multi-omics joint evaluation and analysis in complex environments to reveal the impact on ecosystems in complex environments. • Establish implementable application methods. • Formulate biosafety management regulations for the application.	• Precision prediction of ecological impacts. • Design and establish biological escape prevention mechanisms and methods. • Scenario-specific system applications.

Figure 1 Roadmap for type I artificial multicellular systems

Current Status		
For complex microbial communities composed of multiple species, characterize and regulate their ecological networks, analyze their spatial structures, develop *in situ* editing techniques, and construct predictive models.		
Objective 1: Theory and Design of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Characterization and regulation of the spatiotemporal dynamics of the microbial communities.	• Non-destructive three-dimensional visualization of spatially structured microbial communities. • Design of three-dimensional structured microbial communities in a controlled environment.	• Manipulate the 3D structure of natural or engineered communities using environmental disturbance. • Be able to self-localize in complex environments and colonize in hard-to-reach locations.
Objective 2: Functional Research of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Characterize the functional guilds of the microbial communities.	• Describe and characterize the composition of functional guilds in the microbial communities and their interactions.	• *In situ* characterization of all species and their interactions in natural microbial communities.
Targeted regulation of the function of the microbial communities.	• Remove specific strains or entire functional guilds in natural microbial communities to analyze the causal relationship between host phenotypes and guilds and the interactions among microorganisms. • Design and introduce guilds with new functions or modify existing functions in a controlled microbial communities.	• Establish new regulatory methods to targeted inhibit or knockout specified functional guilds. • *In situ* edit specific species or guilds in natural microbial communities to introduce new functions or modify existing functions.

Objective 3: Environmental Impact of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Response of the microbial communities to environmental changes.	• Predict and engineer the interactions between the microbial communities and the environment.	• Predictive models of microbial community functions and responses to a wide range of environmental changes.

Figure 2 Roadmap for type II artificial multicellular systems

Current Status		
By controlling the differentiation of stem cells into various adult cells and encoding cell signaling pathways, the artificial synthesis of organoids, animal embryos, and organ chips has been achieved. By encoding cells to synthesize artificial biological structures and developing the interface of microbial-mammalian multicellular systems, the application of artificially encoded cells in biomedical fields such as vaccine development and cancer treatment has been realized.		
Objective 1: Theory and Design of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Rational design principles of biomedical artificial multicellular systems.	• The theory and design of generating artificial multicellular systems from single cells. • The theory and design of maintaining cell states. • The theory and design of cell proliferation and renewal. • The theory and design of the formation of low-level tissue structures. • The theory and design of the formation of simple functions.	• Establish the integration theory of artificial multicellular modules. • Establish the formation theory of the interface between machines and artificial multicellular systems. • Establish the theoretical basis for the interactive sensing between modules and between the interface of machines and artificial multicellular systems. • Establish the theoretical basis of adaptation relying on functional activities.

Objective 2: Construction and Testing of Artificial Multicellular Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Construction and evaluation of biomedical artificial multicellular systems.	• Use genetic and protein circuits to achieve the artificial design and control of cell fate. Understand and utilize the biological random decision-making mechanism. Construct simple modules of artificial multicellular patterns and structures through the self-organization and artificial control of artificial multicellular structures. Realize simple artificial multicellular modules with basic functions. • Design and construct the interface between machines and artificial multicellular systems.	• Construct higher-level tissue structures, including the integration of multiple artificial multicellular modules. Establish complex and coordinated functions. Construction of relatively mature phenotypes. • Realize the formation of the interface between machines/artificial multicellular systems. Realize the interactive sensing between modules and between the interface of machines and artificial multicellular systems. Start to design and construct bio-robots.

Figure 3 Roadmap for type III artificial multicellular systems

3.7.4 Technical Pathways

3.7.4.1 Type I Artificial Multicellular Systems

Current Technologies: Type I artificial multicellular systems achieve targeted biosynthesis or biodegradation through metabolite-mediated coupling [1]. In the context of biosynthesis, these systems have been able to utilize complex substrates such as lignocellulose [2], synthesize complex macromolecules like oxidized taxanes [3], produce biofuels such as butanol and isobutanol [4], and develop bioelectrochemical systems including bio-photovoltaic systems [5]. In terms of biodegradation, they have been used to degrade relatively simple pollutants such as benzene, toluene, and phenanthrene, yet applications for high-toxicity recalcitrant pollutants like polycyclic aromatic hydrocarbons and complex pollutant mixtures have not been reported [6–8].

Objectives and Breakthroughs: In terms of the theory and design of artificial multicellular systems, elucidate design principles. In terms of the construction and testing of artificial multicellular systems, develop the construction and evaluation technologies. In terms of the functional research of artificial multicellular systems, develop *in situ* regulation techniques. In terms of the environmental impact of artificial multicellular systems, strengthen the biosafety management.

Challenges: There are several challenges faced by these systems such as lack of modular division principles and theoretical research for synthesis and degradation pathways of various compounds; artificial multicellular systems involve numerous parameters and complex influencing factors, with insufficient understanding of their structural-stability relationship, along with deficiencies in metabolic modeling and computational simulation tools; incomplete comprehension of metabolic characteristics, functional modules, and interconnections between engineered microbial strains in artificial multicellular systems; divergent growth requirements and physiological demands among functional species, leading to instability in microbial community structure; scarcity of cultivation conditions and technologies supporting simultaneous spatiotemporal growth of diverse species; inter-species competition and environmental fluctuations compromising system stability; conflict between system capacity and cultivation throughput; poor stability in liquid environments exacerbated by interspecies competition and environmental variations; absence of efficient *in situ* genetic manipulation techniques, along with inadequate metabolic profiling and flux analysis technologies; interactions between growth processes

and metabolic activities within the system; insufficient preemptive knowledge regarding ecological impacts on native ecosystems; deficiency in multi-species containment technologies for artificial multicellular systems.

Expected Progress Recently: Establish compound-degradation/synthesis module-potential functional strain databases to realize rational design for metabolic division of labor; propose species stability theories and enable simulation/prediction/directed evolution of multicellular systems; develop universal and species-compatible co-cultivation conditions and technologies for rapid system construction; develop autonomous regulation technology based on single signaling molecule to strengthen the population stability and function of simple multicellular system; implement *in situ* genetic regulation and metabolite monitoring technologies; and conduct multi-omics integrated analysis to evaluate the impact of artificial multicellular systems on the ecosystem in complex environments.

Expected Progress by 2030: Achieve computational screening/optimization of compound-degradation and synthesis module-potential funtional strain matches for designing systems capable of degrading/synthesizing complex compounds; refine species stability theories and use AI-driven optimization of system parameters; deploye intelligent parallel cultivation devices and spatial self-assembly technologies; Develop autonomous regulation technology based on multiple signaling molecules to strengthen the population stability and function of simple multicellular system; develop *in situ* metabolic flux analysis and dynamic regulation tools, enabling *in situ* dynamic monitoring and regulation of metabolic division in complex multicellular systems; and establishing biocontainment technologies for multi-species systems.

Potential Solutions

Based on information such as biosynthetic pathways and strain genomes, establish the modular division principles for different compound synthesis routes, and create a database corresponding to compounds-synthesis modules-potential functional strains. Combine existing data with large-scale surveys, such as pollutant surveys, pollutant identification, and determination of soil physical and chemical properties, to establish databases of pollutant types and polluted environments in different regions. Develop cultivation techniques and culturomics, and combine high-throughput sequencing to establish databases of strain resources and gene resources.

Study the influencing factors and mechanisms of the model metabolic division of

labor system, and clarify the key factors for stable division of labor. Construct the nutritional complementary relationship between different strains to maintain the long-term stability of the artificial multicellular system. Optimize the existing metabolic models, integrate metabolic models with population dynamic models, study the interactions between microbe-microbe and microbe-environment, predict the spatiotemporal dynamic changes of the artificial multicellular system ecosystem caused by the intracellular metabolism of a single strain, and guide the adjustment of metabolic nodes. Develop methods for artificial selection and directed evolution of multicellular systems.

Based on genomics, metabolomics, and metabolic models, trace the metabolic flux and its dynamic changes of the artificial multicellular system under different environmental conditions, clarify the metabolic function division of different strains in the artificial multicellular system, optimize the modules of the artificial multicellular system, and improve the database corresponding to compounds-degradation or synthesis modules-potential functional strains. Establish a metabolic division of labor model for the artificial multicellular system, design the metabolic interactions between multiple species, and predict the metabolic flux and community structure under the condition of multi-metabolic pathway intersections.

According to the differences in the physiological characteristics of each species in the artificial multicellular system, establish models such as sequential inoculation models and spatial isolation models, and predict the bacterial community composition and dynamic changes of the artificial multicellular system under the spatiotemporal separation mode. Introduce environmental factors such as dissolved oxygen, temperature, and pH, predict the trend of bacterial community changes under intervention conditions, and find the cultivation conditions for the balance between species. Develop artificial intelligence technology to achieve the prediction and optimization of parameter sets. Develop a dynamic model of the artificial multicellular system under complex conditions, and study the environmental impact mechanism of artificial selection and directed evolution.

Establish a database of growth condition for different species, comprehensively analyze the common and individual growth conditions of different species, and achieve the rapid determination of the growth conditions for any combination of artificial multicellular systems. Develop spatially heterogeneous co-cultivation technologies such as compartmentalized membrane separation, immobilized embedding, and micro-lipid droplets to achieve the physical isolation culture of different species. Establish a quorum-sensing system to impose growth restrictions on strains with rapid substrate

utilization. Develop *in situ* fluorescence imaging technology to monitor the spatiotemporal distribution of species within the artificial multicellular system. Promote the construction of a stress-resistant functional gene element library, and explore and design stress-resistant functional element modules suitable for different scenarios.

Develop high-throughput parallel cultivation devices in different scales based on systems such as droplets, microplates, and small chemostats. Develop spatial assembly technologies for artificial multicellular systems based on 3D printing, etc., to improve the stability and robustness of the system. Develop new materials, explore the spatial self-assembly mechanisms and influencing factors of artificial multicellular systems, and achieve the spatial self-assembly of artificial multicellular systems on special material carriers. Establish a multi-species and multi-signal quorum-sensing system to control the growth balance between different functional species. Design and modify stress-resistant functional part circuits to be applicable to artificial multicellular systems. Design directed evolution technologies for artificial multicellular systems to strengthen the interactions and bacterial community stability within the artificial multicellular system.

Develop new DNA transformation/transduction technologies to improve the efficiency and host range of exogenous gene introduction. Develop isotope metabolic flux analysis technologies for artificial multicellular systems to improve the real-time monitoring ability of intermediate metabolites. Use multi-omics analysis methods to study the changes in gene expression and metabolic activity over time within the artificial multicellular systems.

Design multi-strain and multi-condition induction switch elements to regulate the expression of related genes during the cultivation process and achieve the dynamic regulation of the metabolic flux within the artificial multicellular system.

Improve the multi-omics joint assessment and analysis of artificial multicellular systems, use high-throughput technologies to periodically and standardly monitor the interactions between artificial multicellular systems and indigenous organisms in complex environments, and achieve a comprehensive evaluation of artificial multicellular systems on the local ecosystem. Through comprehensive evaluation, formulate the basic criteria and norms for the ecological application of artificial multicellular systems.

Use traditional biocontainment systems, such as auxotrophy, suicide switches, and gene flow barriers, to develop biocontainment technologies applied to artificial multicellular systems, ensuring that each species in the artificial multicellular system cannot survive in an open environment. Improve the biosafety management regulations for artificial multicellular systems.

3.7.4.2 Type II Artificial Multicellular Systems for Health Applications

Current Technologies: Microbial communities in nature are in dynamic change and exhibit obvious spatial distribution heterogeneity. Currently, the main methods for characterizing the spatial information of the microbial communities include: capturing the spatial composition of the host microbial communities through micrometre-scale metagenomic plot sampling sequencing (MaPS-seq) [9]; *in situ* fluorescent imaging of highly complex microbial communities to display the spatially structured host microbial community [10]; correlating metabolic phenotypes with the *in situ* microbial composition to map the spatial metabolic phenotypes of the microbial communities [11]. At present, the changes in the spatial structure of microbial communities after regulation are still unclear, and methods for characterizing and reshaping the three-dimensional structure of microbial communities need to be developed.

Engineered microbial communities need to function predictably in space and time, which requires new technical capabilities to control the composition, structure, and spread of the microbial communities at different scales and time scales. Complete spatiotemporal control of individual species within the microbial communities will take the lead in forming more advanced models. Currently, there are spatial patterns and turing patterns created by synthetic circuits in a single bacterial population [12]. Optogenetics has been used to control *Escherichia coli* to form templated two-dimensional patterns [13]. The population composition of two bacterial components is maintained through the feedback control of intercellular signaling molecules, and the community structure is precisely regulated using the strategy of spatial separation with engineered biomaterials to achieve the division of labor and communication within the community [14].

Genetically modifying the genomes of the microbial communities can endow the microbial communities with specific traits, including using engineered bacteria or performing *in situ* editing of the microbial communities. Currently, the genetic modification of microbial communities at the metagenomic level can be divided into non-targeted insertion of microbial genomes based on broad-spectrum insertion transposases and targeted editing based on the CRISPR-Cas system, including targeted insertion, deletion, and regulation of gene expression [15]. Although the CRISPR system already has some excellent tools at various levels such as genome engineering, transcriptional regulation, post-transcriptional regulation, and epigenetic editing, at the metagenomic and metatranscriptomic levels, there is still a lack of programmable

operation tools with good universality and specificity. Overall, the ability of current researchers to understand and manipulate systems with specific functions or to repair biomes and consortia that no longer operate as required is very limited.

Biological design faces special challenges because of the large diversity of systems and the still limited relevant control information. Various individual analysis tools, as well as more comprehensive data and analysis workbenches, have begun to emerge, but there are few widely used integrated computational design-build-test-learn support systems, and few can make full use of a large amount of diverse biological data and analysis resources. There have been some standardization efforts currently, but they are still rather isolated. In addition, the use of technologies that are not widely applicable for data representation and analysis execution often hinders the use and development of communities.

Objectives and Breakthroughs: In terms of the theory and design of artificial multicellular systems, achieve the characterization and regulation of the spatiotemporal dynamics of the microbial communities, and be able to study, manipulate, and program the three-dimensional structure of microbial communities. In terms of the functional research of artificial multicellular systems, characterize the composition of the functional guilds of the microbial communities, regulate the functional composition of the microbial communities, achieve new functions at the community and host levels, or solve the problem of ecological dysbiosis. In terms of the environmental impact of artificial multicellular systems, improve the response of microbial communities to environmental changes.

Challenges: New technologies are needed to report and visualize the three-dimensional structure and functions of communities. Self-assembly systems that use motility or chemotaxis to guide microbial movement require more advanced devices to limit the spatial diffusion of compounds.

There is a limited understanding of how communities dynamically respond to the environment, especially in the face of non-homogeneous environmental changes. It is difficult to add sensing and actuation functions to arbitrary cell types within the community environment. In hard-to-access environments, it is challenging to verify whether self-localization and three-dimensional construction occur.

There is a lack of knowledge about the functions and expression times of many genes, as well as their impacts on pathway/organism functions. Although genomic sequences are available, there is a shortage of the necessary information required for

predicting various functions such as cell morphology, metabolism, and proteomics.

Significant advancements in technologies such as microscopy, sample preparation, and image analysis are needed to achieve higher-resolution imaging in hard-to-reach environments like the gut microbial communities and plant microbial communities, and to use this for analyzing complex biological structures.

The ability to selectively add or remove species from biological communities; functional guilds may contain cross-kingdom species, and the integration of multiple technical operations is required; designing and introducing guilds with new functions or modifying existing functions in a controlled microbial communities; existing microbial communities may have more mechanisms to resist changes, especially those related to growth or replication.

There are still certain difficulties in targeting function-related species, especially cross-kingdom species, without off-target effects; effective *in situ* editing of communities is required, with good control over metabolic capabilities, the timing of editing, and the addition of target microorganisms.

Data from different ecosystems and environments are needed to construct predictive models; integrating the responses of microbial communities under various environments, including conditions with non-response or non-optimal responses.

It is necessary to model and compare different pathways to determine the most effective, prioritized, and resilient systems.

Expected Progress Recently: Conduct non-destructive three-dimensional visualization of microbial communities from various environments. Design microbial communities with three-dimensional structures in a controlled environment. Describe and characterize the composition of functional guilds in the microbial communities and their interactions. Remove specific strains or entire functional guilds from natural microbial communities to analyze the causal relationship between host phenotypes and guilds and the interactions among microorganisms. Predict and engineer the interactions between the microbial communities and the environment (such as temperature, oxygen content, pH, small molecules or drugs, and dietary components).

Expected Progress by 2030: Achieve the manipulation of the three-dimensional structure of natural or engineered communities under changing environmental conditions. Engineered microbial communities are able to self-localize in complex environments such as the human gut and colonize in hard-to-reach locations. *In situ* characterize all species and their interactions in natural microbial communities. Establish new regulatory

methods such as phages, small molecule inhibitors, and CRISPR to targeted inhibit or knockout specified functional guilds. *In situ* gene editing for specific species or guilds in natural microbial communities to introduce new functions or modify existing functions. Construct predictive models of microbial communities that functions and responses to different environmental changes.

Potential Solutions

Use imaging, sequencing, and omics technologies to study natural or engineered communities and dynamics. Develop new reporting systems such as cell-based sensors to evaluate and quantify functions and/or three-dimensional structures. 3D-print biodegradable materials into matrices containing chemical components to promote the growth of complex, structured microbial communities. Use optogenetics to more precisely control cell positions and determine bacterial adhesion and microbial communities' structures.

Investigate how natural and engineered communities respond to environmental changes and construct spatiotemporal models that incorporate genomes, functions, and environmental outcomes. Conduct targeted gene editing only on specific members of the community. Design compounds that are synthesized only when the microbial community reaches the designed spatial structure and are easy to detect.

Combine activity-based probes with proteomics to discover functions in complex communities. Design technologies capable of large-scale testing of the microbial communities to detect protein biochemistry, cell phenotypes, and gene functions, thereby increasing the amount of data available for annotation prediction programs.

Develop a library of specially-labeled probes to simultaneously track many different metabolisms and activities and define functionally similar species at multiple levels of abstraction. Develop technologies such as *in situ* fluorescence imaging and mass spectrometry imaging.

Develop targeted antibacterial agents and targeted phages. Expand the use of bacterial gene drive technologies or broad-host-range conjugative plasmids to achieve more consistent genetic manipulation in complex environments. Engineer systems that allow for rapid manipulation and testing and can continuously apply different methods to ensure the elimination of a certain function from the microbial communities. Selectively add community members, including engineered cells, and enable them to retain organisms through nutritional complementation and shared metabolism. Design the

microbial communities to use multiple mechanisms to achieve its functions.

Use high-throughput microbial communities model systems and gene or small-molecule inhibitor screening to enhance the gene function database in complex environments. Improve the technology for *in situ* manipulation of the microbial communities through spatiotemporal control. Introduce alternative metabolic pathways for carbon source utilization to selectively enrich species that can utilize specific nutrients.

Generate and incorporate data from many environments so that the model can be integrated with relevant information about the community, including key functions and the surrounding ecosystem. Through modeling or machine learning, determine the optimal response conditions of the microbial communities based on the individual responses of each functional group to complex environments (such as temperature, oxygen, pH, humidity).

Use controlled laboratory experiments or observational studies to define microbial communities' functions as a function of community composition and the environment. Use machine-learning methods to determine whether there are similar patterns among the microbial communities of interest.

3.7.4.3 Type III Artificial Multicellular Systems for Biomedical Applications

Current Technologies: There have been preliminary advancements in the biomedical research of artificial multicellular systems, such as controlling the differentiation of stem cells into various adult cells, encoding cell signaling pathways, controlling pattern formation, forming organoids from stem cells, and creating organ chips. These are mainly achieved by controlling cell gene expression, designing interactions between cells, and introducing mathematical modeling to make the design process more rational. However, currently, the formation of artificial multicellular systems mostly relies on the inherent natural functions of cells, high-throughput screening, and manual trial and error. There are few cases of artificial rational design being applied in practice. The construction of mammalian multicellular systems is mostly in the exploratory stage. There is a lack of biological devices for controlling cell behavior, controlling the microenvironment, and dividing multicellular regions, as well as artificial control tools. At the same time, there is a shortage of high-throughput cultivation devices of various scales and spatial assembly technologies.

Objectives and Breakthroughs: In terms of the theory and design of artificial multicellular systems, elucidate the rational design principles from a single cell type to

multiple cell types. In terms of the construction and testing of artificial multicellular systems, explore biological devices for regulating cell behavior, controlling the microenvironment, and dividing multicellular regions, develop tools for controlling cell behavior and spatial assembly technologies, construct various modular artificial multicellular systems from easy to difficult, and establish an evaluation system and technology for multicellular modules.

Challenges: There is a lack of reliable model construction and simulation calculation tools for differentiating multiple different cells from a single cell. The understanding of the principles of cell self-organization is not in-depth, and there is a lack of basis and means for artificially controlling cell behavior. There is a lack of understanding of the influence mechanisms of biological random decision-making on cell differentiation, the formation of biological structures, and the structure and stability. There is a lack of multicellular modules with various specific functions. There is a lack of non-invasive continuous monitoring methods for various characteristics of artificial multicellular systems. The conditions required for establishing modules of multicellular systems with different functions vary, and the integration of systems with higher functions is affected by multiple factors. There is a lack of materials and technical means for system integration to construct biotic-abiotic interfaces. The factors affecting the structural and functional stability of artificial multicellular systems are complex.

Expected Progress Recently: Elucidate the design theory of multi-level gene networks directionally guiding cell differentiation, as well as the theory of the formation of multicellular spatial structures by combining artificial control and cell self-organization. Through the exploration of the basic principles of biological random decision-making in artificial multicellular systems, achieve the utilization of the process of biological random decision-making, establish multicellular modules, and realize non-invasive continuous monitoring of artificial multicellular systems.

Expected Progress by 2030: Integration of multicellular modules. Integrate “living” and “non-living” systems to construct biotic-abiotic interfaces. Maintenance, regeneration, and adaptation of artificial multicellular systems.

Potential Solutions

By combining the existing data on the differentiation of stem cells into various cells,

establish a database of different cell differentiation pathways. Through artificially constructed genetic and protein circuits, efficiently and stably regulate the spatiotemporal process of cell differentiation and control the maintenance of cell states. Design automated model construction and simulation calculation tools, integrate various cell activities and behaviors, and predict the spatiotemporal dynamic changes of the three-dimensional structure of artificial multicellular systems through cell-cell and cell-environment interactions, so as to guide the artificial regulation of the spatiotemporal distribution of multicellular systems. Use mathematical models to calculate, predict, adjust and optimize the formation of the three-dimensional structure, quantify the growth, movement and other life activities of multicellular populations, coordinate the interactions between different cells, and derive the formation principles of the three-dimensional structure of artificial multicellular systems.

Study the influence and mechanism of randomness and noise on artificial multicellular systems, and clarify the key factors generating biological randomness. Develop methods to regulate biological noise.

Explore bioparts for regulating cell behavior and the microenvironment, and develop spatial assembly technologies for artificial multicellular systems such as cell self-assembly and 3D bio-printing. Develop multicellular modules of simple tissues, and attempt to establish the interface between the modules (living) of artificial multicellular systems and machines (non-living). Develop biological probes that can detect the functions of different cells, and establish instruments for non-invasive continuous monitoring.

Integrate different multicellular modules through co-culture. Induce the generation of different multicellular modules step by step. Integrate different multicellular modules through different spatial separations and microenvironment regulation. Establish a nutrient exchange and transportation system to make spatially separated integration possible. Develop materials for biological interfaces, and design and prepare microfluidic chips that can achieve the zonal culture of living cells. Develop dynamic models of the structure and function of artificial multicellular systems under complex conditions, and study the influence mechanisms of different internal and external environments on them.

3.7.5 Summary

Artificial multicellular systems composed of cells from multiple species, as well as

those formed by the development and differentiation of cells from a single species, have important application values in fields such as medicine, health, industry, agriculture, energy, and environment. Currently, there is a lack of in-depth understanding of the cell composition and its real-time changes in artificial multicellular systems, the functional diversity of each component cell, and the interactions between cells, metabolites, and the environment. The core issue is the lack of corresponding technical capabilities. Based on characterizing the distributed metabolic capabilities of artificial multicellular systems and the functional diversity of different cells, designing and constructing artificial multicellular systems with locatable distributions and predictable functions over time and space is an important task in synthetic biology research. Through the review of this research, the development directions of the theory and technology of artificial multicellular systems are clarified, and the milestones, breakthrough capabilities, implementation paths, and safety and regulatory considerations for the future up to 2030 are proposed, providing support for the rational design of artificial multicellular systems.

References

[1] Fritts R K, McCully A L, McKinlay J B. Extracellular metabolism sets the table for microbial cross-feeding. Microbiol Mol Biol Rev, 2021, 85: e00135-20.

[2] Shahab R L, Brethauer S, Davey M P, et al. A heterogeneous microbial consortium producing short-chain fatty acids from lignocellulose. Science, 2020, 369: eabb1214.

[3] Zhou K, Qiao K J, Edgar S, et al. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. Nat Biotechnol, 2015, 33: 377-383.

[4] Zhang H R, Wang X N. Modular co-culture engineering, a new approach for metabolic engineering. Metab Eng, 2016, 37: 114-121.

[5] Zhu H, Xu L, Luan G, et al. A miniaturized bionic ocean-battery mimicking the structure of marine microbial ecosystems. Nat Commun, 2022, 13: 5608.

[6] Zhang G B, Yang X H, Zhao Z H, et al. Artificial consortium of three *E. coli* BL21 strains with synergistic functional modules for complete phenanthrene degradation. ACS Synth Biol, 2022, 11: 162-175.

[7] Jia X Q, He Y, Jiang D W, et al. Construction and analysis of an engineered *Escherichia coli*-Pseudomonas aeruginosa co-culture consortium for phenanthrene bioremoval. Biochem Eng J, 2019, 148: 214-223.

[8] Sharma B, Shukla P. Designing synthetic microbial communities for effectual bioremediation: a review. Biocatal Biotransform, 2020, 38: 405-414.

[9] Sheth R U, Li M Q, Jiang W Q, et al. Spatial metagenomic characterization of microbial biogeography

in the gut. Nat Biotechnol, 2019, 37: 877-883.

[10] Shi H, Shi Q J, Grodner B, et al. Highly multiplexed spatial mapping of microbial communities. Nature, 2020, 588: 676-681.

[11] Geier B, Sogin E M, Michellod D, et al. Spatial metabolomics of *in situ* host-microbe interactions at the micrometre scale. Nat Microbiol, 2020, 5: 498-510.

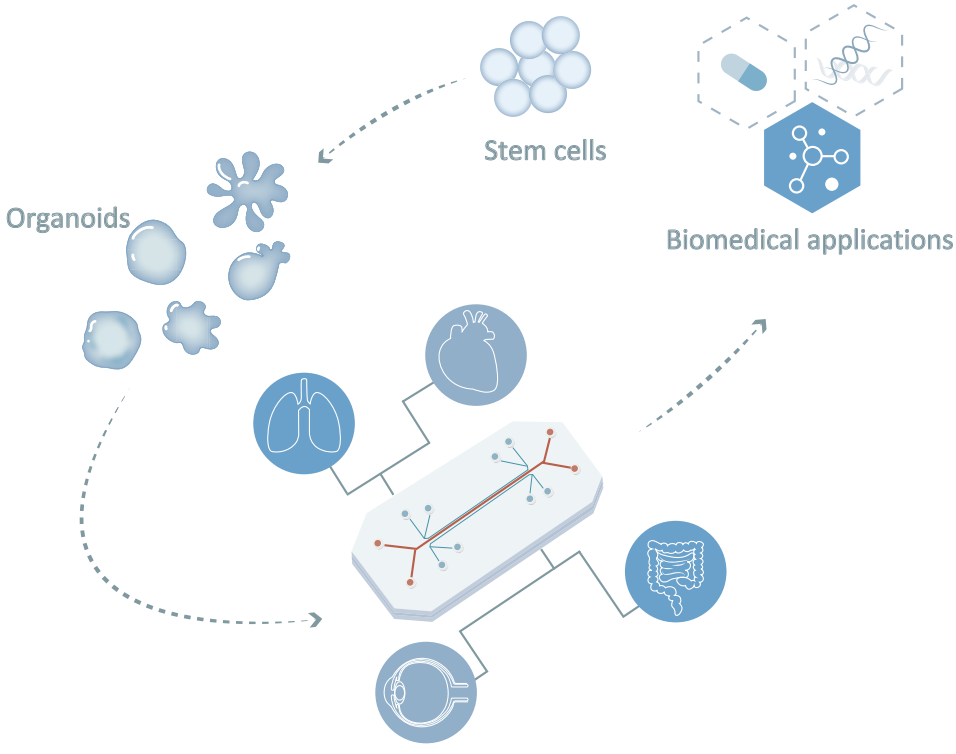
[12] Baumgart L, Mather W, Hasty J. Synchronized DNA cycling across a bacterial population. Nature Genet, 2017, 49: 1282-1285.

[13] Moser F, Tham E, Gonzalez L M, et al. Light-controlled, high-resolution patterning of living engineered bacteria onto textiles, ceramics, and plastic. Adv Funct Mater, 2019, 29: 1901788.

[14] Wang L, Zhang X, Tang C W, et al. Engineering consortia by polymeric microbial swarmbots. Nat Commun, 2022, 13: 3879.

[15] Rubin B E, Diamond S, Cress B F, et al. Species- and site-specific genome editing in complex bacterial communities. Nat Microbiol, 2022, 7: 34-47.

Organoid Engineering



Authors Qin Jian-Hua, Liu Hai-Tao, Wang Ya-Qing, Tao Ting-Ting, Zhang Xu

3.8 Organoid Engineering

3.8.1 Abstract

Organoids are three-dimensional (3D) microtissues typically derived from stem cells or organ-specific progenitors through *in vitro* proliferation, differentiation, and self-organization. They contain multiple cell types with specific spatial arrangements, partially recapitulating the key architecture and function of the counterpart tissues. Organoid engineering employs synergistic strategies that combining engineering and biology to controllably design the self-organizing process. It simulates complex tissue microenvironments to construct *in vitro* biomimetic 3D organ model systems with enhanced physiological relevance and higher fidelity. As a critical pathway for constructing complex multicellular systems, organoid engineering provides novel strategies and technologies for synthetic living systems. It holds significant promise across diverse fields including organ development, disease modeling, drug discovery, and regenerative medicine.

3.8.2 Technical Overview

3.8.2.1 Origins of Organoids

The term "organoid" first appeared in 1960s literature describing dissociated cell reorganization [1]. Advances in stem cell research revitalized organoid studies. In 2009, Hans Clevers et al. pioneered *in vitro* intestinal organoid construction using Lgr5+ intestinal stem cells in 3D culture, establishing the foundation for adult stem cell (ASC)-derived organoid research [2]. Recent developments have facilitated diverse organoid culture systems for developmental biology, disease modeling, and drug testing[2,3]. Tumor organoids are increasingly used for personalized drug evaluation [4,5].

3.8.2.2 Principles of Organoid Formation

Organoids are heterogeneous multicellular 3D structures through self-organization of stem cells that primarily include pluripotent stem cells (PSCs) like embryonic stem cells (ESCs)/induced pluripotent stem cells (iPSCs), or adult stem cells (ASCs). PSCs typically form embryoid bodies in 3D matrices (e.g., Matrigel), differentiating into three germ layers and ultimately tissue-specific organoids. These models recapitulate key

developmental events including self-renewal, lineage-specific differentiation, and self-organization [6,7]. Current culture methods include embedding in matrices, suspension culture, and air-liquid interface systems [2,3,8].

3.8.2.3 Design, Synthesis, and Construction Strategies of Organoids

Due to the complexity of the intrinsic functions of tissues and organs, current organoid systems still have many limitations, restricting their wide application. For example, there is a lack of defined extracellular matrices and controllable microenvironments; a lack of key cellular components (vascular endothelial cells, immune cells, etc.) and vasculature; low throughput and difficulty in reflecting multi-organ interactions. Based on developmental biology principles, the multigerm layer formation and spatiotemporal polarity structures in organoids can be designed; combined with technologies such as biomaterials, microfluidic organ chips, and bioprinting, it is beneficial to design, synthesize, and construct organoid models with higher fidelity, reflecting the physiological characteristics of organs *in vivo*. For example, the combination of new biomaterials with organoids can simulate the extracellular matrix and guide organoid morphogenesis; bioprinting technology is beneficial for constructing organoids with complex multi-layer structures and large scales [9]; microfluidic and organ chip technologies help achieve organoid microenvironment control, vascularization, and interaction research [10,11]; combined with gene editing, multi-omics analysis, and imaging, it helps in the design of organoids, in-depth analysis of structural functions, and acquisition of spatiotemporal information.

3.8.2.4 Biomedical Applications of Organoids

As a complex multicellular system, organoids have broad application prospects in life sciences, medical research, and drug development. Currently, organoids have been used in fields such as tissue/organ development, disease modeling and mechanism research, drug screening, and organ repair. For example, PSCs can derive cerebral organoids with different types of neurons and brain region structures under specific growth factor conditions, simulating the early development process of the brain[3]. Intestinal organoids derived from ASCs or PSCs can simulate the intestinal crypt structure *in vivo*, as well as intestinal absorption and secretion functions, and can be used for modeling diseases such as enteritis [12].

3.8.3 Roadmaps

Current Status		
Primary sources include ASCs, iPSCs, and ESCs. Due to the diversity of stem cell types and the tendency for mutations to occur during the culture process, the efficiency of organoid generation, biological functions, and translational applications are affected, which limits their wide application.		
Objective 1: Large-scale Standardized Stem Cell Production		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
By using advanced technologies such as gene editing, improve the stability of stem cells, optimize the types of stem cells, establish standardized stem cell operation standards, and provide a reliable cell source for constructing organoids.	**Establish an abundant and stable source of stem cells.** • ⩾20 genes modified for stemness. • ⩾10 therapeutic organoid types. • 10^9 cells/batch bioreactors.	**Standardization and regulation of stem cell preparation systems.** • ⩾60 genetic modifications for stemness • ⩾30 therapeutic types. • 10^{12} cells/batch bioreactors. Establish standardized procedures for the preparation of stem cell products, quality control methods, application methods, etc.

Objective 2: High-throughput Organoid Culture Systems and Quality Control		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
By adopting automated culture technology, a high-throughput organoid culture and real-time monitoring system is established, breaking through the technical bottlenecks of traditional manual operations, improving the throughput and stability of organoid preparation, and significantly reducing production costs.	**Establish of a high-throughput organoid culture system.** • Semi-automated platforms, which can partially accommodate operations such as cell fluid replacement and passage, and integrate monitoring system. • 10^6 cells/batch throughput of organoid culture and real-time monitoring system. • >5,000 organoid models from human stem cells, covering 3 types of organs and involving 5 major human diseases.	**Develop a large-scale, automated organoid culture system** • Full automation, which can accommodate operations such as cell fluid replacement and passage, and integrate monitoring system. • Establish a large-scale multi-dimensional tissue organoid library and integrate an intelligent analysis system. • By combining multiple detection and analysis methods, establish standards for organoid QC system.

Figure 1 Roadmap for seed cell sources and culture technologies of organoids

Current Status		
Currently, most organoid systems rely on animal-derived Matrigel for 3D culture, but the matrix components are unclear. Most organoids lack key cellular types such as vascular structures and immune cells. The functional maturity of organoids are low, and it is difficult to achieve multi-organoid functional interactions.		
Objective 1: Design and Development of New Organoid Cultivation Matrices		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Use de novo synthesis of polymers to obtain matrices with clear components needed for stem cell differentiation and organoid cultivation, guiding the controllable formation of different types of organoids.	**Reduce the batch component variation of synthetic materials (<10%).** • Use natural materials in combination with synthetic materials. • Increase purification and quality control steps for materials to reduce batch-to-batch variation. • Select the best matrix material components based on the characteristics of different organoids.	**Reduce the batch component variation of synthetic materials (<5%).** • Fully use synthetic materials. • Modify key functional groups from natural materials into synthetic materials to enhance material functionality. • Develop customized matrices based on the characteristics of different organoids.

Objective 2: Precise Control of Organoid Cultivation Microenvironment		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Combine engineering and materials science technologies, design and synthesize advanced functional materials with adjustable physical properties and consistent topological structures to achieve precise control over important parameters such as biochemical and physical aspects of the organoid microenvironment.	**Construct the microenvironment for organoid formation.** • Develop new responsive materials to precisely control the physical properties of matrices. • Increase material topological structures to guide organoid morphogenesis. • Establish controllable physical stimulation technologies such as light and electricity to simply control physical parameters.	**Precise control of the microenvironment for organoid formation.** • Analyze the correlation between important microenvironmental parameters and organoid structural functions. • Establish controllable physical stimulation technologies such as sound, light, electricity, magnetism, and force to achieve complex physical parameter control. • Combine microfluidics, concentration gradient generation, and other technologies to achieve biochemical parameter control.
Objective 3: Long-term Functional Maintenance and Large-scale Construction of Organoids		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Combine multidisciplinary methods to establish organoids with complex cell composition, structure, and function to address issues related to organoid vascularization, interaction, and the formation of large-scale organoids.	**Stable long-term culture of organoids.** • Establish functional vascular networks to promote internal substance exchange within organoids. • Stable culture organoids for 3-6 months, size >500 μm. • Establish complex organoids containing various cell types including immune cells. • Achieve co-culture and functional association of 3-5 types of organoids.	**Long-term stable culture and functional maintenance of organoids.** • Create hierarchical vascularized structures to form complex organoids. • Organoids cultured for more than 1 year, size increased to mm or cm scale. • Establish organoids with more complex structures, cell composition, and functions. • Develop a universal culture medium that can support the cultivation of 5 or more types of organoids.

Figure 2 Roadmap for organoid design, synthesis, and construction strategy

Current Status		
High-resolution and deep imaging equipment for organoids is scarce; reading, quantitative analysis, and digital modeling of complex biological information within organoids are challenging.		
Objective 1: Develop New Technologies and Methods for Organoid Information Collection and Analysis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Combine artificial intelligence and biosensing, imaging technology, etc., to develop multi-modal functional monitoring methods suitable for organoid systems to collect key biological information and establish organoid functional databases and intelligent analysis technology.	**Multi-modal and multi-dimensional functional monitoring methods for organoids.** • Develop microelectrode array integration technology for organoid culture system. • Establish organoid functional monitoring methods based on light, electricity, and other multi-modal sensing technologies. • Establish organoid data integration software and multi-modal databases, with a database sample size no less than 20,000. • Develop new transparent technology to increase the imaging depth of organoids to over 1 mm.	**Organoid intelligent analysis and data modeling methods.** • Implement real-time, online, and non-destructive monitoring of key organoid functions. • Combine artificial intelligence technology to establish organoid functional analysis systems, achieving intelligent analysis and prediction of biological data. • The sample size of organoid multi-modal databases reaches 100,000. • Achieve deep imaging of organoids up to 2 mm or more.

Objective 2: Digital Modeling of Organoids		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Based on organoid multi-modal databases, integrate organoid proliferation, differentiation, structure formation, and other information to construct digital models that reflect organoid development, structure, function, and physiological/ pathological transformation characteristics, and to achieve data visualization of some key organoid systems.	**Construct digital models of organoids.** • Based on research on organoid proliferation and differentiation, establish digital models that reflect organoid structure, function, and physiological/ pathological outcomes. • Initially establish standardized guidelines for stem cell and organoid data resources, supporting visualization of data from at least 5 sources.	**Assess the functionality of organoids digitally.** • Explore the relationship between organoid size, structure, cell types, physiological indicators, and image signals to establish precise organoid evaluation models. • Construct *in vitro* organ simulation data management systems, supporting visualization of data from at least 10 types of sources.

Figure 3 Roadmap for organoid information monitoring analysis and digital modeling

3.8.4 Technical Pathways

3.8.4.1 Seed Cell Source and Organoid Cultivation Technology

Current Technologies: Currently, the stem cells used for organoids mainly include ESCs, iPSCs and ASCs. Due to the limited long-term proliferation capacity, stemness maintenance ability, and genetic stability of stem cells, they cannot meet the demands of large-scale, long-term organoid research and applications. Traditional stem cell culture mainly uses 2D culture methods, and due to the limited surface space of culture dishes, cell culture and expansion efficiency are not high, which cannot meet the needs of large-scale applications. The self-assembly process of organoids is highly random, making it difficult to control the uniformity of organoid size and shape, resulting in high variability of organoids and low throughput, which cannot meet practical application needs.

Objectives and Breakthroughs: Large-scale production and standardization of stem cells, i.e., improving the stability of stem cells through advanced technologies such as gene editing, optimizing the types of stem cells, establishing stem cell operation standards, and providing a reliable cell source for constructing organoids; organoid culture system and quality control, i.e., using automated cultivation technology to establish a high-throughput organoid culture and real-time monitoring system, breaking through the bottleneck of traditional manual operation technology, improving the throughput and stability of organoid preparation, and significantly reducing production costs.

Challenges: Currently, the maintenance of stem cell phenotype is mainly achieved by optimizing culture conditions and combinations of cytokines, but fundamentally, the maintenance ability of cell genotype is the key to the long-term stability of stem cells, and related research and technological development are still lacking. The organoid formation process involves multiple steps and is cumbersome, which greatly affects the quality and high-throughput controllable production of organoids.

Genetically stable stem cells are of practical significance, but stem cells may show functional variation due to human and environmental differences during culture, storage, and use. Therefore, it is necessary to establish preparation processes, quality control methods, application methods, etc., to standardize the use of stem cells. In addition, how to improve the throughput of organoid production and reduce its cultivation cost has always been an important factor restricting the widespread application of organoids in the biopharmaceutical field. Although organoid culture methods have made significant progress and the corresponding equipment is relatively complete, the need for a large amount of

manual operation and the inability to systematically present data analysis are still difficult issues in this field. The key to solving this problem is how to automate and integrate existing technologies and equipment into the organoid culture process to obtain high-quality cell products while ensuring that all operational details are orderly and complete.

Expected Progress Recently: Establish a large and stable source of stem cells; establish a high-throughput organoid culture system.

Expected Progress by 2030: Standardize and regulate the stem cell preparation system; develop a large-scale, automated organoid culture system.

Potential Solutions

To obtain stem cell sources with stable stemness and long-term proliferation, gene editing technologies such as CRISPR-Cas9 can be used to perform gene editing and screening at the single-cell level of stem cells from different sources to establish stem cell operation standards and reduce deviations in the use of stem cells; construct bioreactor systems and biocompatible mesoporous materials suitable for organoid growth and development to solve the problem of low organoid culture throughput.

For already differentiated adult cells or ASCs, their stemness maintenance ability can be enhanced through reprogramming or gene editing; for totipotent or pluripotent stem cells, their ability to differentiate into a specific tissue or organ can be enhanced through gene editing to obtain stem cells with a clear differentiation direction and genetic stability, and further establish preparation processes, quality control methods, application methods, etc.

Integrating automated culture-related modules and equipment into the organoid culture process to meet the diverse needs of medium replacement, passaging, cell collection, etc., and provide a large number of usable cell products.

Based on current technologies, further improve some key technologies to meet the needs of large-scale production and application, and at the same time, use self-developed automated control programs for intelligent manipulation of the organoid culture process, saving labor costs and greatly improving production efficiency.

3.8.4.2 Organoid Design, Synthesis, and Construction Strategy

Current Technologies: Typically, the production of organoids involves embedding embryoids formed by stem cells in animal-derived extracellular matrices (such as Matrigel) for static culture, such as brain and optic cup-like organoids. However,

animal-derived extracellular matrices have complex components and batch-to-batch variations, which may lead to high variability of organoids. Existing organoid culture systems still lack controllable physical and chemical microenvironments, lack key cell types (vascular endothelial, immune cells, etc.), and vascular structures, which may lead to high variability and low maturity of organoids. In addition, the analytical throughput of the organoid system is low, making it difficult to study interactions between organoids.

Objectives and Breakthroughs: The design and development of new organoid culture matrices involve using *de novo* synthesis polymers to obtain materials with defined components needed for stem cell differentiation and organoid cultivation, guiding the controllable formation of different types of organoids.

Precise control of the organoid culture microenvironment involves combining engineering and materials science to design and synthesize advanced functional materials with adjustable physical properties and consistent topological structures to achieve precise control of important parameters such as biochemical and physical aspects of the organoid microenvironment.

Long-term functional maintenance of organoids and construction of large-scale organoids involve combining multidisciplinary methods to establish organoids with complex cell composition, structure, and function to address issues related to organoid vascularization, interaction, and the formation of large-scale organoids.

Challenges: The existing organoid culture matrices are mostly of natural origin, with batch-to-batch variations, reducing the controllability of the organoid matrix microenvironment and affecting the consistency of organoid formation.

Existing organoid culture systems cannot precisely control biophysical (such as mechanical force, fluid, and electrical stimulation) and biochemical (such as factor gradients) microenvironmental factors, leading to high variability and low maturity of organoids.

The microenvironment of tissues and organs *in vivo* is extremely complex, and existing organoid culture systems cannot simultaneously ensure the precise control of various microenvironmental factors, limiting the formation and development of complex organoids.

As the culture time of organoids is extended and their size continues to increase, cells within the organoids are prone to necrosis, affecting the long-term survival of organoids; the culture environments for different organoids are different, especially the differences in culture medium components, making it difficult to achieve co-culture and functional coupling of multiple organoids.

Existing organoid culture methods can simply co-culture vascular endothelial cells

and other cells with organoids, but they cannot achieve the development of complex and hierarchical structures, nor can they achieve co-culture and interaction research of multiple organoids, affecting the long-term survival and functional maturity of organoids.

Expected Progress Recently: Reduce the batch variation of component of synthetic materials (<10%); construct the microenvironment for organoid formation; stable long-term culture of organoids.

Expected Progress by 2030: Reduce the batch variation of component of synthetic materials (<5%); precisely control the microenvironment for organoid formation; long-term stable culture and functional maintenance of organoids.

Potential Solutions

Use a combination of natural and synthetic materials, increase material purification and quality control steps to reduce batch variations, and obtain matrices with defined components for organoid culture; optimize the best matrix components according to the characteristics of different organoids.

On the basis of fully synthetic artificial polymers, modify key functional groups (such as collagen and laminin) from natural matrices into them to increase material functionality and mechanical properties; develop customized matrices according to the characteristics of different organoids to improve the biomimicry of organoid models.

Develop new photoresponsive isomerized polymers to precisely control the physical properties of extracellular matrix; combine microprocessing technology to increase the topological structure of materials and guide organoid morphogenesis; establish controllable physical stimulation technologies such as light and electricity to simply control biological physical parameters in the organoid microenvironment.

Based on the correlation between important microenvironmental parameters and organoid functional differentiation and structural formation, establish controllable physical stimulation technologies such as sound, light, electricity, magnetism, and force to achieve complex physical parameter control; at the same time, combine microfluidic and factor gradient generation technologies to achieve precise control of biochemical parameters.

Establish functional vascular networks based on biological methods to promote internal substance exchange within organoids, enabling stable culture sizes of organoids more than 500μm, and improve the complexity of organoid cell composition; develop a universal culture medium compatible with the co-culture of 3-5 types of organoids to

initially achieve co-culture and functional association of organoids.

Combine multidisciplinary methods (such as bioprinting) to create hierarchical vascularized structures to form complex organoids and enable long-term culture of more than 12 months; at the same time, increase the size of organoids to the millimeter or centimeter level, further establish organoids with more complex structures, cell composition, and functions; develop a more compatible universal culture medium to meet the co-culture needs of 5 or more types of organoids.

3.8.4.3 Information Monitoring Analysis and Digital Modeling of Organoids

Current Technologies: The size of organoids under static culture is mostly in the micrometer to millimeter range. If dynamic culture systems and vascularization are introduced to extend the culture time of organoids, their size can be further increased, which poses great difficulties for deep imaging and high-content analysis of organoids. In addition, the development process of organs *in vivo* is complex, involving numerous biochemical and biophysical signal changes inside and outside cells. Existing organoid systems lack effective *in situ* information collection and high-throughput analysis technologies. Using a combination of biosensing technology and artificial intelligence for multi-modal, real-time detection, intelligent analysis, and theoretical model establishment of organoid functions is one of the key development directions in this field.

Objectives and Breakthroughs: Develop new technologies and methods for information collection and analysis of organoids. Combine artificial intelligence with biosensing and imaging technology to develop new multi-modal functional monitoring methods suitable for organoid systems, collect key biological information from organoid models, and establish organoid functional databases and intelligent analysis technology.

Digital modeling of organoids involves integrating information on organoid proliferation, differentiation, function, and structure formation based on organoid multi-modal databases to construct digital models that reflect the developmental, structural, functional, and physiological/pathological transformation characteristics of organoids, achieving data visualization of some key organoid systems.

Challenges: In existing organoid culture systems, optical detection is predominantly relied upon, making it difficult to achieve online collection of multi-modal and multi-dimensional biological information, and it is even more impossible to achieve real-time dynamic information feedback and regulation. The development process, morphological changes, and functional expression of organoids are all very complex, and it is difficult to

discover their internal mechanism through the recording and analysis of a single parameter. There is a lack of summary and predictive theoretical models. For complex biological systems, especially organoid systems, there are no targeted software and algorithms.

Expected Progress Recently: Multi-modal and multi-dimensional functional monitoring methods for organoids; construction of digital models of organoids.

Expected Progress by 2030: Intelligent analysis and data modeling methods for organoids; digital assessment of organoid functionality.

Potential Solutions

By integrating microelectrode arrays into organoid culture devices and establishing multi-modal functional monitoring methods based on light and electricity, real-time online monitoring of organoid metabolism and electrophysiological functions can be achieved. The development of new transparentization techniques can reduce experimental time and costs. Based on this, organoid data integration software and multi-modal databases can be established, with a database sample size of no less than 20,000.

Artificial intelligence technology is suitable for deep analysis and trend prediction of big data. Its use in the intelligent analysis of organoid multi-modal data can achieve intelligent analysis and prediction of key biological data of organoids. Combined with deep learning, organoid image intelligent analysis software can be developed to further expand the sample size of organoid multi-modal databases to 100,000.

Based on the data of organoid proliferation and differentiation obtained from the construction and culture process of organoids, digital models that reflect the structure, function, and physiological/pathological outcomes of organoids can be established, thus initially establishing standardized guidelines for stem cell and organoid data resources and supporting the visualization of data from at least 5 sources.

Based on existing image reconstruction and finite element calculation analysis technology, explore the relationship between organoid size, structure, cell types, physiological indicators, and image signals. Construct *in vitro* organ simulation data management systems to support the visualization of data from at least 10 sources.

3.8.5 Summary

Organoids have shown broad applications in fields of tissue and organ development,

disease modeling, drug screening, and organ repair. However, the organoid model system is not yet standardized in terms of cell sources and extracellular matrices, and there is still great room for development in organoid long-term culture, functional maturity, multi-organ interconnection, functional information output, and theoretical modeling. This study combines multidisciplinary methods to apply technologies such as gene editing, biomaterials, organ chips, bioprinting, and multi-modal information collection and analysis to organoid engineering research. It is expected to create a 3D tissue organ model system with higher physiological relevance and fully realize the application value of organoids from both theoretical and practical perspectives, providing new strategies and platforms for the *in vitro* construction of complex human life systems and animal alternatives in biomedicine applications.

References

[1] Weiss P, Taylor A C. Reconstitution of complete organs from single-cell suspensions of chick embryos in advanced stages of differentiation. Proc Natl Acad Sci USA, 1960, 46(9): 1177-1185.

[2] Sato T, Vries R G, Snippert H J, et al. Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. Nature, 2009, 459(7244): 262-265.

[3] Lancaster M A, Renner M, Martin C A, et al. Cerebral organoids model human brain development and microcephaly. Nature, 2013, 501(7467): 373-379.

[4] Li L, Knutsdottir H, Hui K, et al. Human primary liver cancer organoids reveal intratumor and interpatient drug response heterogeneity. JCI Insight, 2019, 4(2), e121490.

[5] Lesavage B L, Suhar R A, Broguiere N, et al. Next-generation cancer organoids. Nat Mater, 2022, 21(2):143-159.

[6] Kim J, Koo B K, Knoblich J A. Human organoids: model systems for human biology and medicine. Nature Reviews Molecular Cell Biology, 2020, 21(10): 571-584.

[7] Lancaster M A, Knoblich J A. Organogenesis in a dish: modeling development and disease using organoid technologies. Science, 2014, 345(6194),1247125.

[8] Mccracken K W, Cata E M, Crawford C M, et al. Modelling human development and disease in pluripotent stem-cell-derived gastric organoids. Nature, 2014, 516(7531): 400-404.

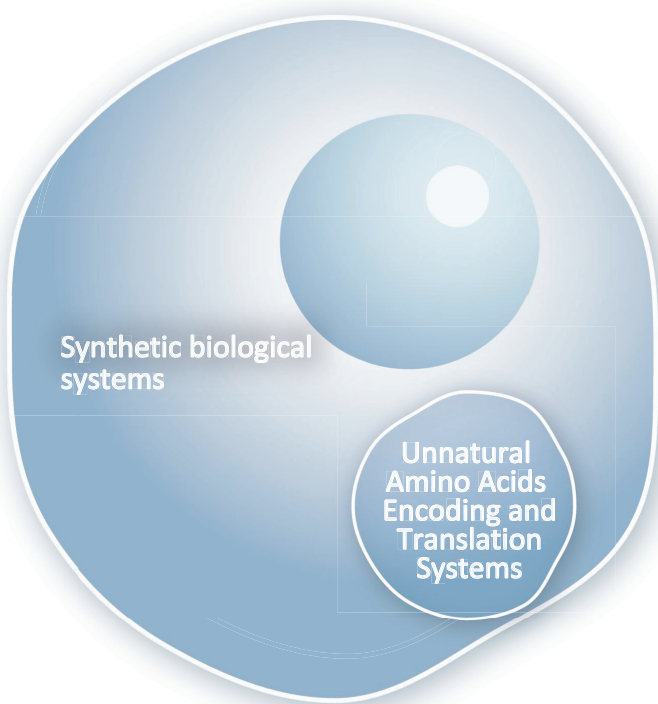
[9] Zhang Y S, Pi Q M, Van Genderen A M. Microfluidic bioprinting for engineering vascularized tissues and organoids. JoVE-Journal of Visualized Experiments, 2017, (126), e55957.

[10] Yin F, Zhang X, Wang L, et al. HiPSC-derived multi-organoids-on-chip system for safety assessment of antidepressant drugs. Lab Chip, 2021, 21(3): 571-581.

[11] Tao T, Deng P, Wang Y, et al. Microengineered multi-organoid system from hiPSCs to recapitulate human liver-islet axis in normal and type 2 diabetes. Adv Sci (Weinh), 2022, 9(5): e2103495.

[12] Wang X, Yamamoto Y, Wilson L H, et al. Cloning and variation of ground state intestinal stem cells. Nature, 2015, 522(7555): 173-178.

Unnatural Amino Acids Encoding and Synthetic Biosystems



Authors Chen Peng, Wang Jie, Ge Yun, Hao Zi-Yang

3.9 Unnatural Amino Acids Encoding and Synthetic Biosystems

3.9.1 Abstract

The genetic encoding system is highly conserved. Both lower and higher organisms use the same 20 amino acids, the same set of codons, and unified encoding rules for protein synthesis. To achieve genetic coding of unnatural amino acids (UAA), it is necessary to construct a completely new and orthogonal set of bioparts, including new tRNAs, aminoacyl tRNA synthetases (aaRS), codons, and even new ribosomes[1,2]. Over nearly 30 years of research, scientists have continuously developed and updated various components, successfully achieving genetic coding for over 300 types of unnatural amino acids [3], and can simultaneously introduce up to 4 unnatural amino acids at specific sites within a protein[4]. The development of unnatural encoding systems has greatly expanded the chemical space and functional scope of proteins. In this section, we clarify the current bottlenecks of unnatural encoding systems and the main future development directions, including: the expansion of genetic codons and the expansion and establishment of unnatural amino acids encoding; further development of genetic information system expansion technology based on artificial nucleobases; and applications of unnatural amino acids in protein engineering.

3.9.2 Technical Overview

3.9.2.1 Genetic Codon Expansion and Unnatural Amino Acids Encoding

In the classic protein translation process, free amino acids are catalyzed by aminoacyl tRNA synthetases to form aminoacyl tRNA with the corresponding tRNA, which is then inserted into the peptide chain with the help of free ribosomes by forming a peptide bond with the carbon end of the peptide chain. By utilizing the orthogonality between aminoacyl tRNA synthetase/tRNA pairs from different species and using the amber terminator (TAG) as the encoding codon, the encoding of unnatural amino acids can be achieved. Since the introduction of unnatural amino acids into proteins through genetic encoding *in vitro* in 1989, followed by the discovery of the pyrrolysine aminoacyl tRNA synthetase system (PylRS-tRNA), and the subsequent expansion of genetic codons

in transgenic mice, researchers have made groundbreaking progress in the expansion of genetic codons and the encoding of unnatural amino acids[2].

Firstly, by mining, modifying, and evolving aminoacyl tRNA synthetases and corresponding tRNAs, more unnatural amino acids can be recognized and encoded. This includes the development and evolution of biologically orthogonal components from multiple species, the discovery and modification of shuttle systems such as PylRS-tRNA, and the design of chimeric synthetases, which have laid the foundation for introducing chemical diversity into proteins[5].

Secondly, to encode multiple unnatural amino acids in the same protein, having only one encoding codon TAG is far from enough, so developing new codons for encoding unnatural amino acids is also an important part of genetic codon expansion. By using quadruple codon technology or replacing and degenerate redundant codons at the genomic level (for example, reducing six serine codons to four), the number of codons for encoding unnatural amino acids can be increased to achieve simultaneous encoding of multiple unnatural amino acids[6].

Finally, to achieve the application of genetic codon expansion technology in different organisms, this technology needs to be extended to various organisms, from prokaryotes such as *Escherichia coli* and *Shigella*, to eukaryotic systems such as yeast, and then to multicellular model organisms such as nematodes, fruit flies, and zebrafish, and finally to mammals such as mice[7]. Researchers have continuously broken through the limits and boundaries of unnatural amino acid encoding technology, achieving another dimension of genetic codon expansion.

3.9.2.2 Transcription and Translation of Unnatural Base Nucleic Acids

In 1989, Benner's research group first successfully achieved *in vitro* replication and transcription of DNA based on hydrogen bonding with artificial base pairs isoG-isoC[8], opening the prelude to artificially expanding the genetic alphabet. In 2006, the Hirao research group first achieved efficient PCR amplification and *in vitro* transcription containing Ds-Pa artificial bases[9]. It took nearly 20 years to move from *in vitro* to *in vivo*, until the Floyd E. Romesberg research group successfully constructed a six-nucleotide artificial synthetic bacterium containing a pair of artificial bases (dNaM-d5SICS) in 2014, and for the first time successfully achieved *in vivo* replication of artificial base pairs[10]. This milestone achievement marked the official transition of artificial base research from *in vitro* to *in vivo*. Artificial bases not only simulate the characteristics of natural bases in

structure but also are designed and optimized from a functional perspective. Currently, artificial bases including s-y, Ds-Pa/Ds-Px, NaM-5SICS/NaM-TPT3, etc., have achieved replication and transcription *in vitro* and *in vivo*. In 2017, the Floyd E. Romesberg research group modified tRNA to recognize artificial DNA bases X and Y (dNaM and dTPT3), and transported two unnatural amino acids PrK and pAzF to the ribosome, ultimately successfully achieving *in vivo* replication, transcription, and translation of artificial bases[11]. This work, starting from the artificial base system, for the first time reproduced the central dogma in an unnatural form, providing a highly promising new path for codon expansion research and applications. The introduction of X and Y artificial bases increased the genetic code from 4 to 6, potentially adding up to 152(6^3-4^3) different amino acid encodings. This greatly enriched the insertion of unnatural amino acids in functional proteins, further expanding the unnatural coding system and even new forms of life.

3.9.2.3 Protein Engineering Based on Unnatural Amino Acids

Unnatural amino acids, with the diversity of their chemical side chains, greatly expand the diversity of protein structure and function[2]. As biophysical probes, unnatural amino acids can assist in characterizing proteins using various biophysical imaging methods such as NMR and X-ray crystallography; by introducing unnatural amino acids containing bioorthogonal functional groups[12], site-specific labeling and tracing of proteins can be achieved, and protein activity regulation can also be achieved through cleavage reactions[13]; introducing photocrosslinking groups can capture interacting protein complexes[14]; and site-specific insertion of unnatural amino acids can simulate post-translational modifications of proteins[15]. With this expanded set of unnatural amino acids, the physicochemical properties of proteins can be improved, and new types of active proteins can be designed and evolved[16].

In the application of disease treatment[17], in addition to assisting in the design and development of protein prodrugs, protein conjugate drugs, and covalent protein drugs, the orthogonality of the genetic codon expansion system can be used to construct unnatural amino acid-dependent defective organisms, and to develop small molecule-regulated CAR-T[18], attenuated vaccines[19], and bio-safe engineered bacteria[20] and pathogens.

3.9.3 Roadmaps

Currently, genetic encoding of various unnatural amino acids has been achieved, but there are still two major scientific issues that need to be broken through: first, the number of codons for encoding unnatural amino acids is still limited, and encoding multiple unnatural amino acids in a protein at the same time is still quite challenging; second, the efficiency of introducing unnatural amino acids is insufficient, and the artificially evolved aaRS-tRNA UAA system is still difficult to compare with the components of the natural translation system, especially when introducing multiple unnatural amino acids, the efficiency will significantly decrease. The main technical difficulties in applying genetic codon expansion technology in eukaryotes are: the lack of specificity in the translation process of mRNA, which may lead to erroneous translation; the limited number of codons that can be reallocated without changing the host function; and the lack of orthogonal aaRS/tRNA pairs. To overcome these technical difficulties, the following three aspects are the key breakthroughs in this field.

3.9.3 Roadmaps

Current Status		
Various codons encoding UAA have been developed, and biorthogonal components of multiple UAA translation systems have been developed, but the translation efficiency is relatively insufficient and cannot compare with the components of the natural translation system.		
Objective 1: Design, Generate, and Evolve Proteins Containing Multiple Unnatural Amino Acids on Demand		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Expand codons for unnatural amino acids, including the release of redundant codons in the genome, and the development of new codons.	**Develop 5 new codons for encoding unnatural amino acids.** • Large-scale genome editing in model organisms to release redundant codons. • Develop new codons (quadruplex codons). • Improve the orthogonality between new codons.	**Develop 10 codons for unnatural amino acids with efficiency close to the natural translation system.** • Continue to release redundant codons and develop new codons. • Develop more than 10 tRNA components matching new codons. • Significantly improve the translation efficiency and broad orthogonality of new codons and corresponding aaRS/tRNA components.

Objective 2: Systematic and Large-Scale Development of Orthogonal Genetic Encoding Elements		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Significantly expand broad orthogonal elements in the unnatural encoding system, identification and site-specific introduction of various structurally different unnatural amino acids.	**Construct more than 20 broadly orthogonal tRNA and corresponding synthetase components.** • Development and directed evolution of various broad orthogonal tRNA and aaRS. • Cross-docking between orthogonal tRNA and aaRS components and new codons. • Characterization of orthogonality of different translation systems in common model organisms.	**"Plug and play" of various broad orthogonal elements.** • Achieve *in vivo* synthesis of multiple unnatural amino acids and complete docking with genetic codon expansion technology. • Construct various UAA on-demand encoding tools to achieve "plug and play" of UAA encoding systems in different model organisms.

Figure 1 Roadmap for genetic codon expansion and encoding of unnatural amino acids

Current Status		
Various unnatural base pairs based on hydrogen bonding and hydrophobic interactions have been developed, semi-synthetic systems have been constructed in *E. coli*, and one pair of artificial bases has achieved *in vivo* replication, transcription, and translation of unnatural amino acids.		
Objective 1: Expand the Codons of Unnatural Bases to Achieve Genetic Encoding of Multiple Unnatural Amino Acids		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Optimize artificial base pairs to achieve stable and efficient encoding of multiple unnatural amino acids *in vivo*.	**Expand 2-3 new artificial base pairs for efficient encoding of unnatural amino acids.** • Simultaneous efficient encoding of 2-3 unnatural bases in the same gene in prokaryotes. • Develop an artificial base encoding system maintenance system and reduce base mismatching and DNA repair.	**Develop 10 codons for unnatural amino acids with efficiency close to the natural translation system.** • Introduce artificial bases into 4-base codons, optimize the translation system, and achieve efficient insertion of more than 10 unnatural amino acids. • Develop a high stability and high fidelity artificial base encoding system close to the natural translation system.
Objective 2: Integration of Unnatural Bases in Model Life Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Expand the chassis microorganisms encoding artificial bases and develop eukaryotic *in vivo* artificial base codon expansion systems.	**Adapt chassis microorganisms based on existing unnatural base pairs for codon expansion.** • Establish efficient artificial base codon expansion systems in 2-4 types of chassis bacteria.	**Establish semi-synthetic biological systems in eukaryotes.** • Establish *in vivo* artificial base codon expansion systems in eukaryotes to achieve efficient insertion of unnatural amino acids.

Figure 2 Roadmap for transcription and translation of unnatural base nucleic acids

Current Status		
Modification and derivatization of biological macromolecules have been achieved through UAA, and a variety of new protein drugs have been developed; UAA-dependent CAR-T cells and viral vaccines have been successfully prepared; however, the functionality of UAA is relatively singular, and the application scenarios are limited.		
Objective 1: Obtain New Therapeutic Strategies through Unnatural Amino Acids		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Design new protein drugs containing unnatural amino acids; production of drugs dependent on unnatural amino acids; construction of intelligent entities with multiple insertions of unnatural amino acids and routes.	**Expand functional proteins containing multiple unnatural amino acids.** • Achieve site-specific multi-point insertion of post-translational modifications such as phosphorylation and glycosylation. • Design and develop more than 5 new types of unnatural amino acid catalytic centers and reaction types. • Design and implement various new modes of action for protein drugs using unnatural amino acids for disease treatment. • With the aid of optimized unnatural amino acid translation system, optimize the industrial synthesis, insertion number, efficiency, and corresponding protein expression of UAA, so that functional proteins containing UAA meet industrial production requirements.	**Construct intelligent bacteria or cells using unnatural amino acids.** • Construct and design platforms for multi-orthogonal synthetic systems and new functional protein elements containing unnatural amino acids, and collect and improve data. • Prepare intelligent cells or engineered bacteria integrating multiple signal pathways containing unnatural amino acids with more than three functions. • Develop intelligent bacteria/cells dependent on unnatural amino acids for disease treatment and complete clinical trial research.

Figure 3 Roadmap for protein engineering based on unnatural amino acids

3.9.4 Technical Pathways

3.9.4.1 Genetic Codon Expansion and Encoding of Unnatural Amino Acids

Current Technologies: Although various unnatural amino acids can be genetically encoded, there are still significant challenges in achieving on-demand design and synthesis of proteins containing unnatural amino acids. First, the codons for encoding unnatural amino acids are still very limited, and it is difficult to achieve simultaneous encoding of multiple unnatural amino acids and efficiently obtain proteins composed of multiple unnatural units. Second, the aminoacyl tRNA synthetase systems for recognizing unnatural amino acids are relatively limited, which means that the range of unnatural chemical structures that can be introduced still has a large expansion space. Finally, the efficiency of introducing unnatural amino acids that have been genetically encoded is still behind that of the natural amino acid encoding system, and the translation efficiency will significantly decrease when introducing multiple unnatural amino acids.

There are already various unnatural amino acid encoding systems such as MjTyrRS, PylRS, LeuRS, ChPheRS, etc., but they are still not enough for simultaneous encoding of multiple amino acids. Although Chin et al. have developed various broad orthogonal tRNAs through design and evolution, their translation efficiency is still far from that of the natural system.

Objectives and Breakthroughs: Design, generate, and evolve proteins containing multiple unnatural amino acids on demand, translating into proteins with at least 5 different unnatural amino acid building blocks; significantly expand the unnatural encoding system and construct more than 20 broad orthogonal unnatural encoding elements.

Challenges: There are insufficient orthogonal codons for encoding unnatural amino acids; there are insufficient mutually orthogonal aminoacyl tRNA synthetases (aaRS)/tRNA pairs for the synthesis of unnatural structures. Unnatural amino acids are mostly chemically synthesized, which may limit their use in large-scale applications. There are hundreds of unnatural amino acids that can be genetically encoded, but the limited number of encoding system components restricts the introduction of chemical diversity. The number of orthogonal systems for encoding unnatural amino acids is limited; the encoding efficiency of most unnatural amino acid encoding components is not high.

Expected Progress Recently: Synthesize proteins containing three or more different unnatural amino acids; construct more than 20 broad orthogonal elements based on

orthogonal component mining, chimeric technology, and directed evolution.

Expected Progress by 2030: Integrate the unnatural amino acid biosynthesis system and the genetic codon expansion system; develop new biorthogonal elements based on protein/tRNA *de novo* design and large-scale genome mining to achieve an unnatural amino acid introduction efficiency similar to "wild type".

Potential Solutions

Develop a 4-base codon system, release redundant codons in the genome, develop strains with orthogonal ribosome genome encoding, and use special tRNA for special information genetic encoding; design and modify orthogonal ribosomes, or expand new, mutually orthogonal aaRS/tRNA pairs; use organelle genetic codes and related aaRS/tRNA pairs designed for biosynthesis pathways that can produce unnatural amino acids *in vivo*; develop new protein improvement strategies and tools (such as chimeric technology, continuous directed evolution technology), construct various broad orthogonal encoding system components; establish *de novo* design models based on protein-tRNA complexes, and develop co-evolution methods for synthetases and tRNA.

3.9.4.2 Transcription and Translation of Unnatural Base Nucleic Acids

Current Technologies: Unnatural bases X and Y (dNaM and dTPT3) have been inserted into codons to form new codons (such as AXC); by introducing tRNA with anticodon GYT, preliminary encoding and site-specific insertion of unnatural amino acids have been achieved. The genes of unnatural bases can initially achieve transcription, translation, and encoding of unnatural amino acids in *E. coli*.

Objectives and Breakthroughs: Use the codons of unnatural bases to achieve genetic encoding of multiple unnatural amino acids, optimize artificial base pairs, and achieve encoding of multiple unnatural amino acids *in vivo*. Adapt chassis microorganisms based on existing unnatural base pairs for codon expansion to expand semi-synthetic organisms (SSO) systems and achieve integration of unnatural bases in model life systems.

Challenges: The 4-base codon system of natural bases has been applied to the insertion of unnatural amino acids, and artificial bases have only been incorporated into 3-base codons, not yet expanded to 4-base codons; maintaining the stability of genes containing artificial bases and reducing DNA repair caused by base mismatching with artificial bases is a current

major challenge; the replication fidelity of artificial base pairs is still significantly different from that of natural base pairs. Artificial bases can only encode unnatural amino acids in *E. coli* and currently cannot encode unnatural amino acids in eukaryotes.

Expected Progress Recently: Simultaneously and efficiently encode 2-3 unnatural bases in the same gene in prokaryotes, achieving efficient insertion of 2-3 unnatural amino acids; develop an artificial base encoding system maintenance system and reduce base mismatching and DNA repair; expand 2-4 chassis microorganisms suitable for replication, transcription, and translation of artificial base pairs.

Expected Progress by 2030: Introduce artificial bases into 4-base codons, optimize the corresponding translation system, achieve efficient insertion of more than 10 unnatural amino acids; establish a high stability and high fidelity artificial base codon expansion system based on artificial bases; establish an artificial base codon expansion system in eukaryotes to achieve efficient insertion of unnatural amino acids for drug proteins, vaccines, targeted protein site-specific modification, and functional regulation.

Potential Solutions

Explore alternative (previously explored) unnatural base pairs under optimized conditions, especially those that do not disrupt the double helix structure of DNA and can be incorporated into any sequence environment; improve the existing Z-PandDs-Px system or develop new artificial base systems to achieve *in vivo* replication and transcription; expand tRNA recognizing unnatural bases to simultaneously and efficiently encode 2-3 artificial bases and efficiently insert multiple unnatural amino acids in the same gene in prokaryotes; optimize corresponding tRNA and ribosomes for recognition of 4-base codons containing artificial bases; modify the transport synthesis, DNA replication enzymes, and functional proteins related to DNA repair of unnatural bases/nucleotides, such as using the CRISPR/Cas system to remove mutated artificial base pairs, to make artificial bases stably exist *in vivo* DNA for a long time; optimize various enzymes and working components related to replication, transcription, and translation to improve the fidelity of artificial base replication; based on the existing *E. coli* artificial base encoding system, find related translation machinery in other microorganisms, optimize replication, transcription, translation, and other encoding systems, and expand artificial base pairs to 2-4 types of chassis microorganisms. In addition to considering the properties of the unnatural base pairs themselves, modify

related functional proteins such as artificial base transport synthesis-related proteins, DNA replication enzymes, RNA polymerases, tRNA, ribosomes, etc., in eukaryotic cells to achieve encoding of unnatural amino acids based on artificial bases in eukaryotic cells.

3.9.4.3 Protein Engineering Based on Non-natural Amino Acids

Current Technologies: Currently, site-specific insertion of unnatural amino acids can achieve modification and derivatization of biological macromolecules such as cytokines, growth factors, and antibodies, improving drug efficacy, uniformity, targeting, and safety. In recent years, covalent protein drugs with unnatural amino acid insertion have also begun to develop with the help of proximity effects. In addition, with the orthogonality of the genetic codon expansion system, UAA-dependent CAR-T, attenuated or weakened HIV and influenza vaccines can be prepared. Moreover, genetic encoding of unnatural amino acids with active structures has shown initial success in the design of new artificial enzymes and the development of molecular recognition proteins, and has been applied to fields such as efficient synthesis, accurate recognition, and inert bond activation. Currently, the application of genetic encoding of unnatural amino acids is mainly focused on using the reactivity and orthogonality of a single unnatural amino acid, and the functional mode is relatively singular, making it difficult to achieve on-demand use of multiple unnatural amino acids in the same protein.

Objectives and Breakthroughs: Obtain new therapeutic strategies through unnatural amino acids, break through the bottlenecks of efficiency and orthogonality of genetic codon expansion technology, design new modes of action and application scenarios for unnatural amino acids in therapeutic strategies, and achieve the development of integrated intelligent drugs.

Challenges: The types of unnatural amino acids and the expression efficiency and scale of the genetic codon system in proteins still need to be expanded; it is still difficult to construct multi-path biological pathways or signal routes containing multiple unnatural amino acids; the safety and efficacy of intelligent bacteria within biological organisms are not high; it is difficult to achieve quantitative control of multiple unnatural amino acid introduction pathways in intelligent bacteria or cells.

Expected Progress Recently: Expand functional proteins containing multiple unnatural amino acids.

Expected Progress by 2030: Construct intelligent bacteria or cells using unnatural amino acids.

Potential Solutions

Develop specific strains, aminoacyl tRNA synthetase/tRNA pair systems, ribosomes, and bacteria with genome re-encoding to achieve insertion expression of special unnatural amino acids; improve the orthogonality of the unnatural system, optimize the industrial synthesis, insertion number, efficiency, and corresponding protein expression of unnatural amino acids; develop new modes of protein functionalization mediated by unnatural amino acids.

Construct multi-orthogonal synthesis systems. Construct and improve the data collection and functionality of platforms for the design of combinational routes of new functional protein components containing unnatural amino acids; precisely quantify and uniformly design and balance the activity of protein functional elements corresponding to different unnatural amino acid insertion efficiencies.

3.9.5 Summary

The introduction of unnatural amino acids into proteins has greatly expanded the chemical and functional space of proteins. Although various unnatural amino acids have been genetically encoded, the advantages of the unnatural amino acid translation system in protein function modification and improvement have not yet been fully developed. The future development direction of this field includes the following two points: first, significantly expand the codons encoding unnatural amino acids to achieve simultaneous encoding of multiple unnatural amino acids in proteins; second, improve the efficiency of introducing unnatural amino acids so that the artificially evolved aaRS-tRNA UAA system can compare with the components of the natural translation system in encoding multiple unnatural amino acids. Therefore, systematically developing codons, elements, and translation systems for encoding unnatural amino acids will provide strong support for the improvement of existing functional proteins and the development of new functional proteins, and will also bring revolutionary changes to the industrial systems of functional proteins in related fields such as pharmaceuticals, chemical industry, food, enzyme industry, and agriculture.

References

[1] Liu C C, Schultz P G. Adding new chemistries to the genetic code. Annual Review of Biochemistry,

2010, 79:413-444.

[2] Chin J W. Expanding and reprogramming the genetic code. Nature, 2017, 550:53.

[3] Dumas A, Lercher L, Spicer C D, et al. Designing logical codon reassignment-Expanding the chemistry in biology. Chemical Science. 2015, 6(1): 50-69.

[4] Dunkelmann D L, Oehm S B, Beattie A T, et al. A 68-codon genetic code to incorporate four distinct non-canonical amino acids enabled by automated orthogonal mRNA design. Nature Chemistry, 2021, 13(11):1110-1117.

[5] De La Torre D, Chin J W. Reprogramming the genetic code. Nature Reviews Genetics, 2021, 22(3):169-184.

[6] Fredens J,Wang K, De La Torre D,et al. Total synthesis of *Escherichia coli* with a recoded genome. Nature, 2019, 569(7757):514-518.

[7] Brown W, Liu J, Deiters A. Genetic code expansion in animals. ACS Chemical Biology, 2018, 13(9): 2375-2386.

[8] Switzer C, Moroney S E, Benner S A. Enzymatic incorporation ofanew base pair into DNA and RNA. Journal of the American Chemical Society, 1989, 111(21): 8322-8323.

[9] Hirao I, Kimoto M, Mitsui T, et al. An unnatural hydrophobic basepair system: Site-specific incorporation of nucleotide analogs into DNA and RNA. Nature Methods, 2006, 3:729-735.

[10] Malyshev D A, Dhami K, Lavergne T, et al. A semi-synthetic organism with an expanded genetic alphabet. Nature, 2014, 509(7500):385-388.

[11] Zhang Y, Ptacin J L, Fischer E C, et al. A semi-synthetic organism that stores and retrieves increased genetic information. Nature, 2017, 551(7682): 644-647.

[12] Lang K, Chin J W. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. Chem Rev, 2014, 114(9): 4764-4806.

[13] Wang J, Wang X, Fan X, et al. Unleashing the power of bond cleavage chemistry in living systems. ACS Cent Sci, 2021, 7(6): 929-943.

[14] Nguyen T A, Cigler M, Lang K. Expanding the genetic code to study protein-protein interactions. Angew Chem Int Ed, 2018, 57(44):14350-14361.

[15] Conibear A C. Deciphering protein post-translational modifications using chemical biology tools. Nature Reviews Chemistry, 2020, 4(12):674-695.

[16] Drienovska I, Roelfes G. Expanding the enzyme universe with genetically encoded unnatural amino acids. Nature Catalysis, 2020, 3(3):193-202.

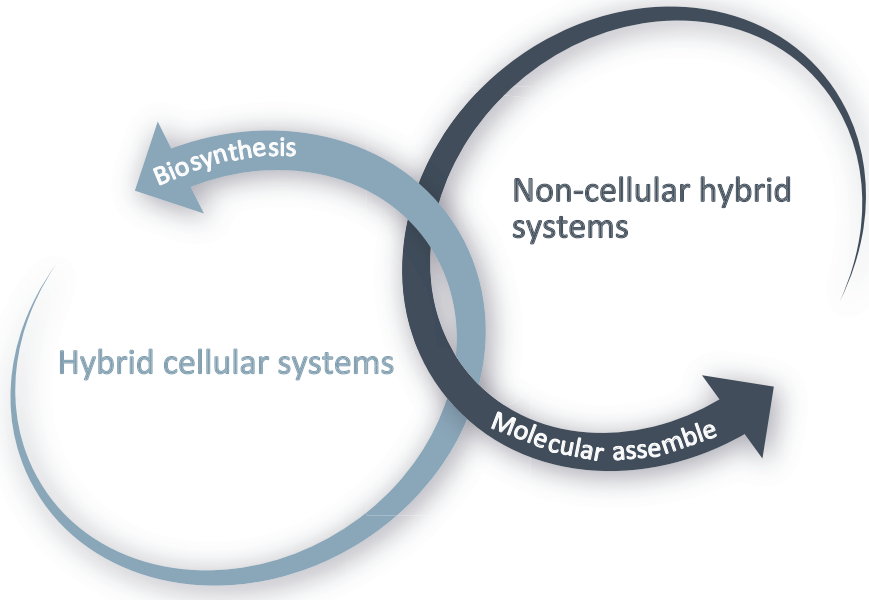
[17] Huang Y, Liu T. Therapeutic applications of genetic code expansion. Synth Syst Biotechnol, 2018, 3(3): 150-158.

[18] Ma J S, Kim J Y, Kazane S A, et al. Versatile strategy forcontrolling the specificity and activity of engineered T cells. Proc Natl Acad Sci USA, 2016, 113(4): E450-458.

[19] Si L, Xu H, Zhou X, et al. Generation of influenza A viruses as live but replication-incompetent virus vaccines. Science, 2016, 354(6316):1170-1173.

[20] Mandell D J, Lajoie M J, Mee M T, et al. Biocontainment of genetically modified organisms by synthetic protein design. Nature, 2015, 518(7537): 55-60.

Biotic-abiobic Hybrid Systems



Authors Li Feng, Cui Zong-Qiang, Pang Dai-Wen, Zhang Xian-En

3.10 Biotic-abiotic Hybrid Systems

3.10.1 Abstract

Biotic-abiotic hybrid systems refer to elements or systems composed of bioparts (such as nucleic acids, proteins, viruses, etc.) and abiotic components (such as inorganic micronanomaterials). These components can be orderly combined through molecular assembly, biomineralization, and other means, leading to enhanced properties or emergent functions. In recent years, the construction technology of biotic-abiotic hybrid systems has begun to form and develop through the integration of different disciplines and technologies such as nanoscience, materials science, and synthetic biology, showing great application potential in various fields including medicine, diagnostics, sensing, and energy. However, due to the limited biotechnological synthesis of abiotic materials and weak control over the orderly hybridization of biological and abiotic components, there is a lack of systematic theoretical guidance, limiting the construction efficiency and application scope of hybrid systems, and there are difficulties in stable, large-scale preparation. It is expected that by 2030, biotic-abiotic hybrid system technology will achieve breakthroughs in the biosynthesis of abiotic components, hybrid regulation of biotic-abiotic components, construction and application adaptation of hybrid chassis cells, and provide key technology platforms for areas such as biological imaging and sensing, drug delivery, new vaccine development, and artificial photosynthesis.

3.10.2 Technical Overview

The molecules that perform functions in organisms are mainly organic, including proteins, nucleic acids, lipids, etc. Introducing inorganic materials and other abiotic molecules into biological systems through synthetic biology can significantly expand their functions. In recent years, people have been able to load inorganic nanoparticles directly onto cells or synthesize inorganic nanomaterials directly within living cells through metabolic regulation, showing the unique advantages of biotic-abiotic hybrid systems, which have become a new growth point in the development of synthetic biology [1-3]. Unlike traditional cell engineering and modification, the introduction of abiotic components such as inorganic materials faces greater challenges in predictability, compatibility, and standardization. Biotic-abiotic hybrid systems can be divided into

non-cellular hybrid systems and hybrid cellular systems, involving multiple levels including the design of biological synthesis of abiotic components, hybridization of biological and abiotic components, adaptation of hybrid systems to living cells, and coordination between hybrid cells and natural cells.

3.10.2.1 Non-cellular Hybrid Systems

Non-cellular hybrid systems are composed of biological and abiotic components that are synthesized separately and then combined. Their combination is generally achieved through chemical cross-linking, adsorption, mineralization, assembly, and other means. Biological components are synthesized through engineered living cells; abiotic components are mainly synthesized through chemical methods. For biotic-abiotic hybrid systems, abiotic components synthesized by chemical methods usually refer to inorganic nanomaterials (such as gold nanoparticles, quantum dots, iron oxide nanoparticles, etc.) and polymeric materials (such as polymeric nanoparticles) that are comparable in size to biomacromolecules (such as proteins, nucleic acids, viruses, etc.). In 1998, Douglas and others used the pH-dependent gating mechanism of Cowpea chlorotic mottle virus (CCMV) to mineralize and synthesize inorganic nanoparticles within its empty capsid[4]. In the following 20 years, inorganic nanoparticles of different components have been synthesized within various virus-like particles (VLPs) or cage proteins, and have been applied in different fields such as biomedical imaging, disease diagnosis and treatment, and microelectronic devices[5-6]. Almost at the same time, the co-assembly technology of chemically synthesized nanomaterials and biomacromolecules has also been developed. For example, Dragnea and others have since 2003 achieved the encapsulation of inorganic materials such as gold nanoparticles, quantum dots, and magnetic particles within VLPs[7]; Zhang and others encapsulated quantum dots within VLPs and achieved long-term, single-particle tracking of viral infection processes[8-9]. Currently, scientists can mineralize or package inorganic nanomaterials in some cage-shaped proteins with self-assembly properties *in vitro*, but there are still multi-faceted challenges in achieving precise and orderly control of the assembly.

3.10.2.2 Hybrid Cellular Systems

Hybrid cellular systems are engineered cells that carry or contain abiotic components such as inorganic nanomaterials and use them as products of biological synthesis or

functional units. The unique optical, electrical, thermal, catalytic, and other properties of abiotic components provide a unique manipulation space for enhancing or creating cellular functions. In fact, nature itself contains systems that construct high-performance or unique functional structures through biological-inorganic hybridization. For example, bone is a tissue formed by the highly ordered arrangement of organic matter and inorganic minerals, which has properties such as toughness, hardness, and lightweight; magnetotactic bacteria synthesize magnetosomes within their cells, using the Earth's magnetic field to navigate and find suitable habitats. However, the types of materials involved in such natural systems are very limited. In 2009, Pang's team proposed a "spatial-temporal coupling regulation strategy for nanoparticle synthesis in living cells" and synthesized CdSe quantum dots with different colors of fluorescence within yeast, opening up a new era for the biological synthesis of quantum dots[10]. In 2016, Yang and others used *Thermus thermophilus* as a chassis to synthesize CdS quantum dots, leveraging the excellent light absorption properties of quantum dots to transform non-photosynthetic bacteria into photosynthetic ones, directly reducing CO_2 into chemical raw materials[11]. Therefore, hybrid cellular systems hold great innovative potential and can provide artificially enhanced chassis cells. Hybrid cellular systems can be constructed by loading chemically synthesized abiotic components onto or into living cells or by directly synthesizing abiotic components within living cells. Depending on the application purpose, the loading form is divided into two situations: cell surface loading and transmembrane transport, with the latter requiring the delivery of abiotic components to specific subcellular regions, which is more technically challenging. Biological synthesis of abiotic components can generally be achieved through metabolic engineering, spatial-temporal coupling, and other strategies. The main types of inorganic nanomaterials that can currently be synthesized by living cells are quantum dots and magnetic particles, and the types are still relatively limited. Control over the physical and chemical properties of biologically synthesized inorganic nanomaterials also remains challenging[12].

3.10.3 Roadmaps

Current Status		
It has been possible to achieve hybridization of bioparts (proteins, nucleic acids, etc.) with abiotic components (quantum dots, gold nanoparticles, magnetic particles, and other inorganic nanomaterials) through chemical cross-linking, adsorption, mineralization, assembly, and other means, obtaining complementary properties of biological and abiotic components, and a large number of new functional concept validation studies have been conducted. However, at present, hybridization can only be achieved with a few types of components (mainly 2-3 types).		
Objective 1: Ordered and Controllable Hybridization of Multiple Components		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Increase the complexity of hybridization of multiple biological and abiotic components, endowing materials with functional diversity.	**Orderly and controllable hybridization of four and more components.** • Integrate two or more inorganic nanomaterials and two or more biological components in the same hybrid system. • Analyze of interfacial interactions between biological and abiotic components.	**Orderly and controllable hybridization of multiple components and dynamic control.** • Design hybrid systems to respond to external conditions such as temperature, environmental pH, and light. • Form universal principles for the design of biological-inorganic nanomaterial hybrid systems.

Objective 2: Precise Control of Hybridization between Biological and Abiotic Components		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Improve the degree of adaptation and precision of hybridization between biological and abiotic components.	**Precise quantitative control of hybridization between biological and abiotic components.** • Enrich the surfaces and interfaces of abiotic and biological components, enhance the specificity and adaptability of active reaction groups. • Develop precise characterization and measurement methods for non-cellular hybrid systems.	**Precise site-specific control of hybridization between biological and abiotic components** • Controllable synthesize abiotic materials on specific biological components to achieve labeling of specific biological components. • Establish plug-and-play assembly interfaces between abiotic components such as quantum dots, magnetic nanoparticles, and gold nanoparticles and biological components such as proteins and nucleic acids. • Achieve tunable and optimized function enhancement.

Figure 1 Roadmap for non-cellular hybrid system technology

Current Status		
Molecular-level regulation of cellular hybrid systems, achieving the construction of multiple functional cellular hybrid systems through the coordinated regulation of multiple metabolic pathways within cells, such as the synthesis of various quantum dots in yeast, bacteria, and mammalian cells.		
Objective 1: Biosynthesis of Abiotic Components by Living Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
For specific inorganic materials, rapidly select chassis cells and formulate biosynthesis routes.	**Inorganic material types and chassis cell adaptation.** • Achieve biosynthesis of more than 3 inorganic material systems. • Form more than 5 types of chassis cells for biosynthesis of different inorganic materials. • Clarify the principles and routes of inorganic material biosynthesis.	**Theoretical system for *in vivo* biosynthesis of abiotic components by living cells.** • Regulate the physical and chemical properties of biosynthetic inorganic materials. • Establish a theoretical system for the biosynthesis of abiotic materials, providing predictable and customizable solutions for the biosynthesis of abiotic materials from multiple dimensions such as chassis adaptation, material basis, energy drive, and regulatory mechanisms.

Objective 2: *In situ* Hybridization of Biological and Abiotic Components in Living Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Specifically control the hybridization of biological and abiotic components within living cells to avoid interference from non-specific biological components and achieve enhanced biological functions.	***In situ* hybridization of biological components with exogenously introduced inorganic materials in living cells.** • Develop 2-3 interfaces for controlling *in situ* hybridization. • Establish 2-3 genetic circuits for *in situ* hybridization. • Achieve efficient assembly of quantum dots, magnetic particles, etc., introduced into living cells with biological molecules.	**Biosynthesis and *in situ* hybridization of biological and abiotic components in living cells.** • Develop 4-6 interfaces for controlling *in situ* hybridization. • Establish 4-6 genetic circuits for *in situ* synthesis and hybridization. • Achieve efficient *in situ* assembly of quantum dots, magnetic particles, and other materials synthesized within living cells with biological molecules, as well as high-yield preparation of hybrid products.
Objective 3: Function Enhancement and Expansion of Artificial Hybrid Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Ability to enhance and expand cell functions using hybrid systems.	**Demonstration of hybrid cell applications.** • Construct artificial cells that can stably and efficiently produce hybrid viruses, virus-like particles, exosomes, and other carriers. • Develop new technologies such as bioimaging, disease diagnosis and treatment, vaccines, and artificial photosynthesis based on hybrid systems. • Form 2-3 biological and medical application demonstrations based on hybrid cell systems.	**Standardization of hybrid cell applications.** • Establish standardized interfaces for hybrid cells to work with natural cells. • Achieve functional coordination and emergence of new functions between hybrid cells and natural cells.

Figure 2 Roadmap for cellular hybrid system technology

3.10.4 Technical Pathways

3.10.4.1 Non-cellular Hybrid System Technology

Current Technologies: It has been possible to achieve hybridization of biological components (proteins, nucleic acids, etc.) with abiotic components (quantum dots, gold nanoparticles, magnetic particles, and other inorganic nanomaterials) through chemical cross-linking, adsorption, mineralization, assembly, and other means, obtaining complementary properties of biological and abiotic components, and a large number of proof of concept studies have been conducted. However, at present, hybridization can only be achieved with a few types of components (mainly 2-3 types).

To meet the needs of various application fields for new components, new systems, and enhanced and expanded synthetic biological systems, further development of non-cellular hybrid system technology will focus on breakthroughs in diversification of hybrid components, site-specific, precise and quantitative control of hybridization, and standardization of hybrid systems.

Objectives and Breakthroughs: Ordered and controllable hybridization of multiple components, enhancing the complexity of hybridization of multiple biological and abiotic components, and endowing materials with functional diversity; precise control of hybridization between biological and abiotic components, improving the degree of adaptation and precision of hybridization between biological and abiotic components.

Challenges: The hybridization process of biomacromolecules with inorganic materials involves various intermolecular forces and numerous interaction sites, making it difficult to control accurately. Existing non-cellular hybrid systems generally only achieve orderly control in simple systems (one type of biological component and one type of abiotic component); as the number of biological and abiotic components increases, the number of factors involved in the hybridization process increases sharply, increasing the difficulty of control. Existing non-cellular hybrid systems are mainly static structures, limiting the potential for properties and functions of hybrid systems, and unable to match specific application scenarios.

Due to the numerous surface active groups on biological components and the difficulty in precisely controlling the surface functionalization sites and quantities of abiotic components such as inorganic nanomaterials, the interface between biological and abiotic components in hybrid systems is complex, and the stoichiometric ratio of the two

in the hybrid system is difficult to control accurately.

Since biological components typically have numerous surface-active groups with scattered locations, while surface functionalization sites of non-biological components such as inorganic nanomaterials are often isotropic, the connection or binding sites between biological and non-biological components during hybridization are random and lack spatial specificity.

Expected Progress Recently: Achieve orderly and controllable hybridization of four or more components; precise quantitative control of hybridization between biological and abiotic components.

Expected Progress by 2030: Orderly and controllable hybridization of multiple components and dynamic control; precise site-specific control of hybridization between biological and abiotic components.

Potential Solutions

Use structural biology, molecular dynamics simulation, and other means to screen several model biological components (such as multi-component viral capsids, exosomes) and abiotic component systems (quantum dots, magnetic particles with clear surface properties) to understand their multi-component assembly mechanisms and develop multi-level assembly strategies; based on model assembly systems, combined with cryo-electron microscopy, solid-state nuclear magnetic resonance, molecular dynamics simulation, and other technologies, analyze the assembly interface of hybrid systems, study their molecular interaction patterns, and guide the controllable construction of biotic-abiotic hybrid systems.

Design hybrid systems to respond to external conditions such as temperature, environmental pH, and light, and achieve coordinated operation of environmental response characteristics between biological and abiotic components to dynamically regulate the properties and functions of hybrid systems; based on research of model hybrid systems, establish universal principles for the design of biological-inorganic nanomaterial hybrid systems.

Optimize the modification of surfaces of biological and abiotic components, starting from the reaction interface, design and modify surfaces that are compatible with each other, enrich the surfaces and interfaces of abiotic and biological components, enhance the specificity and adaptability of active reaction groups; develop precise characterization and measurement methods for non-cellular hybrid systems to serve the controllable construction of hybrid systems.

Use high-resolution structural information of biological components to introduce functional sites on specific biological components to mediate the synthesis of abiotic components, achieving site-specific hybridization; develop precise functionalization strategies for nanoparticles and biological molecule linkers to establish plug-and-play assembly interfaces between abiotic components such as quantum dots, magnetic nanoparticles, and gold nanoparticles and biological components such as proteins and nucleic acids.

3.10.4.2 Cellular Hybrid System Technology

Current Technologies: Currently, it is possible to regulate cellular hybrid systems at the molecular level, using only a single metabolic pathway within cells or intracellular oxidizing/reducing species for redox reactions to synthesize inorganic materials. Scientists can construct multiple functional cellular hybrid systems through the coordinated regulation of multiple metabolic pathways within cells, such as the synthesis of various quantum dots in yeast, bacteria, and mammalian cells.

To meet the needs of various application fields for new components, new systems, new chassis, and enhanced and expanded synthetic biological systems, further development of hybrid cellular system technology will focus on breakthroughs in the biosynthesis of abiotic components, live-cell *in situ* hybridization, standardization of hybrid cells, and functional coordination between hybrid cells and natural cells.

Objectives and Breakthroughs: For live-cell *in situ* biosynthesis of abiotic components, for specific inorganic materials, the ability to quickly select chassis cells and formulate biosynthesis routes. For live-cell *in situ* hybridization of biological and abiotic components, the ability to specifically control the hybridization of biological and abiotic components within living cells to avoid interference from non-specific biological components and achieve enhanced biological functions. For function enhancement and expansion of artificial hybrid cells, the ability to use biotic-abiotic hybrid systems to enhance and expand cell functions.

Challenges: The processes involved in the uptake and incorporation of source materials required for the synthesis of inorganic materials by living cells, the intracellular transport and transformation of source materials, and the impact of inorganic materials on the viability of cells themselves are not well understood, making it difficult to rationally select chassis cells and design synthesis schemes. The live-cell *in situ* synthesis of abiotic components (mainly inorganic materials) is an emerging field, and the existing synthesis systems can only sporadically achieve material synthesis, and have not yet formed a

theoretical system that can guide design.

After the introduction of inorganic materials into living cells, their subcellular delivery is often limited by endocytosis pathways; the mutual recognition and assembly of biological components with exogenously introduced inorganic materials within living cells are subject to interference from widely existing biological molecules within living cells, making hybridization more difficult to control than in extracellular solutions. After the synthesis of inorganic materials within living cells, their surface properties and associated biological molecules are unknown, making it impossible to control their orderly assembly with specific biological components. Due to the compatibility issues between inorganic materials and living cells, the robustness of hybrid cellular systems needs to be improved, and the unique advantages of hybrid cellular systems remain to be explored.

Whether hybrid cells are used for the production of hybrid structures or the entire cell serves as a functional unit, there is currently a lack of unified principles in terms of source materials, biosynthesis, and hybridization regulation methods; regarding how hybrid cells work with natural cells, there have not been predictable pathways or interface systems.

Expected Progress Recently: Inorganic material types and chassis cell adaptation; *in situ* hybridization of biological components with exogenously introduced inorganic materials; demonstration of hybrid cell applications.

Expected Progress by 2030: Theoretical system for live-cell *in situ* biosynthesis of abiotic components; live-cell *in situ* synthesis and hybridization of biological and abiotic components; standardization of hybrid cell applications.

Potential Solutions

Use proteomics, mass spectrometry, molecular imaging, and other analytical means to explore the uptake, transport, transformation patterns of source materials required for the synthesis of inorganic materials, as well as their impact on cells; achieve the biosynthesis of more than three inorganic material systems; form more than five types of chassis cells for the biosynthesis of different inorganic materials; clarify the principles and routes of inorganic material biosynthesis.

On the basis of establishing the synthesis systems and chassis cell adaptation for inorganic materials such as quantum dots, magnetic particles, and gold nanoparticles, regulate the physical and chemical properties of the biosynthetic inorganic materials, extract patterns; establish a theoretical system for the biosynthesis of abiotic materials to provide predictable

and customizable solutions for the biosynthesis of abiotic materials from multiple dimensions such as chassis adaptation, material basis, energy drive, and regulatory mechanisms.

Develop 2-3 control interfaces suitable for live-cell *in situ* hybridization; establish 2-3 genetic circuits for regulating the intracellular delivery of inorganic materials and their *in situ* hybridization with biological molecules, achieving efficient assembly of quantum dots, magnetic particles, etc., after introduced into living cells with biological molecules.

Use structural biology, omics, mass spectrometry, and other technologies to study the surface properties of inorganic materials synthesized within living cells in depth, and establish surface *in situ* functionalization strategies for different types of inorganic materials, develope 4-6 control interfaces for *in situ* hybridization; establish 4-6 genetic circuits for *in situ* synthesis and hybridization according to the needs of hybrid interfaces; achieve efficient *in situ* assembly of quantum dots, magnetic particles, and other inorganic materials synthesized within living cells with biological molecules, as well as high-yield preparation of hybrid products.

Optimize chassis cells to construct artificial cells that can stably and efficiently produce hybrid virus-like particles and exosomes and other carriers; develop new technologies such as bioimaging, disease diagnosis and treatment, vaccines, and artificial photosynthesis based on biotic-abiotic hybrids; form 2-3 biological and medical application demonstrations based on hybrid cell systems.

Further improve the robustness of hybrid cells through genetic editing and circuit design, and optimize and formulate clear guidelines for source material selection, biosynthesis, and hybrid regulation methods (such as metabolic pathway selection, spatial-temporal coupling strategies, etc.) for specific inorganic materials such as quantum dots, magnetic particles, and gold nanoparticles; select several pairs of hybrid cell and natural cell systems, and establish standardized interfaces for hybrid cells to work with natural cells through research and artificial modification of their intercellular interaction patterns; explore the functional coordination and emergence of new functions between hybrid cells and natural cells on standardized hybrid cell systems.

3.10.5 Summary

The biotic-abiotic hybrid system is a product of the integration of synthetic biology and nanobiology. The hybrid system combines abiotic components with biological components to form hybrid structures, or introduces abiotic components into living cells, enabling living cells to obtain functional units beyond nature. Obviously, this type of hybrid

system has opened a door for synthetic biology, making some functions that were originally impossible in natural life systems possible, thus becoming a new type of enabling technology platform. The development towards 2030 will focus on the continuous upgrading and improvement of the biotic-abiotic hybrid itself and its adaptation to application scenarios. The hybrid system mainly addresses issues such as controllable biosynthesis of abiotic components, precise and controllable hybridization of abiotic components with biological components, construction of stable hybrid chassis cells, and functional coordination between hybrid cells and natural cells. Directions with a good research foundation, such as bioimaging, drug delivery, multimodal diagnosis and treatment, new vaccines, and artificial photosynthesis, will be the main application outlets.

References

[1] Li F, Zhang X-E. Nanosynthetic biology: a new dimension of fusion innovation. Syn Biol J, 2022, 3(2): 253-255.

[2] Zhang H, Wu Q, Li H, et al. Cross-integration of synthetic biology and nanobiology and its application in biomedicine. Syn Biol J, 2022, 3(2): 279-301.

[3] Feng Q, Zhang T, Zhao X, et al. Synthetic nanobiology: the intersection of synthetic biology and nanobiology. Syn Biol J, 2022, 3(2): 260-278.

[4] Douglas T, Young M. Host-guest encapsulation of materials by assembled virus protein cages. Nature, 1998, 393(6681): 152-155.

[5] Edwardson T G W, Levasseur M D, Tetter S, et al. Protein cages: from fundamentals to advanced applications. Chem Rev, 2022, 122(9): 9145-9197.

[6] Li F, Wang Q B. Fabrication of nanoarchitectures templated by virus-based nanoparticles: Strategies and Applications. Small, 2014, 10(2): 230-245.

[7] Aniagyei S E, Dufort C, Kao C C, et al. Self-assembly approaches to nanomaterial encapsulation in viral protein cages. J Mater Chem, 2008, 18(32): 3763-3774.

[8] Li F, Zhang Z P, Peng J, et al. Imaging viral behavior in mammalian cells with self-assembled capsid-quantum-dot hybrid particles. Small, 2009, 5(6): 718-726.

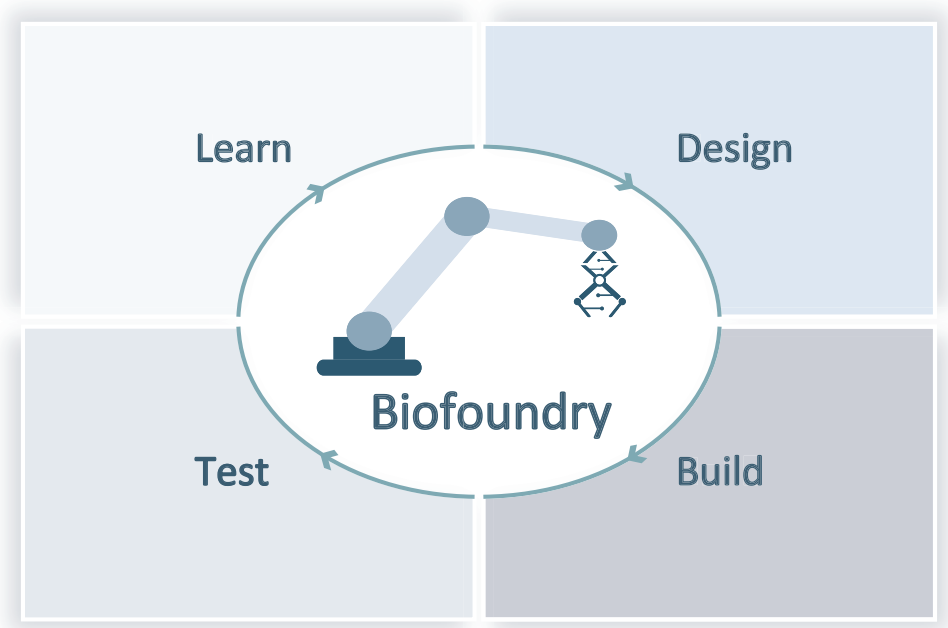
[9] Li Q, Li W, Yin W, et al. Single-particle tracking of human immunodeficiency virus type 1 productive entry into human primary macrophages. ACS Nano, 2017, 11(4): 3890-3903.

[10] Cui R, Liu H H, Xie H Y, et al. Living yeast cells as a controllable biosynthesizer for fluorescent quantum dots. Adv Funct Mater, 2009, 19(15): 2359-2364.

[11] Sakimoto K K, Wong A B, Yang P D. Self-photosensitization of nonphotosynthetic bacteria for solar-to-chemical production. Science, 2016, 351(6268): 74-77.

[12] Jia J, Yang L, Liu A, et al. “Time-space coupling” synthesis of quantum dots in living cells. Syn Biol J, 2022, 3(2): 385-398.

Biofoundry



Authors Si Tong, Jin Fan, Wang Meng, Zhang Chong

3.11 Biofoundry

3.11.1 Abstract

Due to the lack of rational design principles, synthetic biology currently requires extensive trial-and-error experimentation to gradually approach engineering goals. Biofoundries address this challenge by introducing standardized and automated experimental methods to accelerate design-build-test-learn (DBTL) cycles, transforming synthetic biology research from manual, low-throughput processes to automated, high-throughput workflows[1-5], and therefore substantially shortening the experimental cycle and improving research and development efficiency. The efficient operation of biofoundries requires research and innovation in high-throughput synthetic biology processes, automated instrumentation and integration, and information management systems. Currently, dozens of biofoundries have been established or are under construction worldwide. In 2019, the Global Biofoundry Alliance (GBA) was established to strengthen international collaboration and develop unified standards[6].

3.11.2 Technical Overview

Automated synthetic biology addresses the fundamental challenge of achieving predictable design and construction of synthetic biosystems, by enhancing standardization and modularization of experimental subjects, methods, workflows, and technologies in synthetic biology. It seeks to achieve automated closed-loop operation of massive engineering trials while continuously develops the capability to rationally design synthetic biosystems. Research in automated synthetic biology can not only rapidly accumulate large numbers of high-quality gene functional modules and establish standardized workflows, but also obtains high-quality, massive experimental data, thus using data-driven approaches to develop and optimize computational models for the system design and functional prediction of synthetic organisms[3-5].

Biofoundries aim at automating each stage of the DBTL cycles. Computer-aided biological design automation is a critical step in introducing engineering principles into synthetic biology research. During the experimental phase, synthetic biology processes compatible with automated instrumentation must be developed, including the construction and quality control of engineered DNA, genetic manipulation of chassis

systems, and functional testing of synthetic organisms. Among them, the construction process of DNA mainly includes gene synthesis, amplification, restriction enzyme digestion, assembly, and extraction, while the quality control process mainly includes concentration analysis, fragment size analysis, qPCR, and sequencing analysis. The construction process for engineered DNA chassis systems includes cell transformation (heat shock, electroporation), colony plating, colony picking, cell lysis, etc., while the functional testing process includes transcriptomics/proteomics/metabolomics analysis, optical analysis (imaging, flow cytometry, UV/visible/infrared/Fluorescence/Raman spectroscopy), Mass spectrometry, sequencing, fermentation process evaluation, etc [3-5].

The automated operation of synthetic biology workflows requires biofoundry infrastructure integrating both hardware and software systems. On the hardware side, physical automation requires standardized experimental containers (SBS-format microtiter plates), along with compatible instruments including liquid handling stations, microplate centrifuges, sealers and peelers, automated shaking incubators, and colony pickers. These instruments need matching physical interfaces to connect with robotic arms, conveyor belts, and other automated transfer devices to transport samples, reagents, and consumables. On the hardware side, an integrated software system is required to automate the control of experimental operations for instruments, equipment, and transport devices. Additionally, a material and information management system is required to record and coordinate operational processes, experimental results, and materials [3-5] .

3.11.3 Roadmaps

Current Status		
Achieved automated assembly of thousands of cloned DNA fragments per week on a single production line; achieved automated operations for model microbial chassis such as *E. coli* and yeast based on microplate containers.		
Objective 1: Achieve Automated Operation for Large DNA Fragments and Multiple Chassis Systems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop automated workflow and equipment to achieve high-throughput, standardized construction of large DNA fragments.	**Microliter-scale system achieves automated assembly of cloned DNA fragments with single production line throughput of up to 10 Mb per day, reducing costs by an order of magnitude compared to manual operations.** • Optimize enzyme- and cell-based systems for automated DNA assembly in microplates. • Develop computer-aided design software supporting the entire automated workflows from DNA design to script generation of machine execution.	**Nanoliter-scale system achieves automated assembly of cloned DNA fragments with single production line throughput of up to 100 Mb per day, reducing costs by two orders of magnitude compared to manual operations.** • Intelligently orchestrate microplate, microfluidic, and microarray systems, as well as enzyme- and cell-based assembly methods to formulate production plans for large DNA fragments. • Develop hardware equipment compatible with large DNA fragment operations.
Develop automated workflows and equipment to achieve automated, high-throughput, and standardized operations of multiple chassis systems.	**Achieve automated operations for 10+ microbial chassis, reducing single-clone operation costs by an order of magnitude compared to manual operations.** • Develop whole-genome-level genetic manipulation tools. • Establish automated genetic manipulation workflows for single-cell microorganisms. • Establish an automated equipment system by combining multi-device connection and single-device integration.	**Achieve automated operations of 30+ microbial chassis and 5+ plant/animal cell chassis, reducing single-clone operation costs by two orders of magnitude compared to manual operations.** • Develop 5-10 loci parallel genome manipulation tools. • Establish automated manipulation workflows for multinucleate and multicellular chassis. • Improve efficiency and expand functionality of automated equipment to be compatible with special physicochemical and environmental control.

Objective 2: Achieve Multi-Modal, Cross-Scale, Automated, and Quantitative Testing of Synthetic Biosystems		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop automated workflows and equipment, expand automation-compatible testing methods, establish automated calibration frameworks, and achieve multi-modal, cross-scale, and quantitative functional characterization of synthetic organisms.	**Achieve 20+ types of automated functional testing for biomolecules and microorganisms.** • Develop high-throughput sample preparation methods for sequencing, spectroscopy, mass spectrometry, and other analytical techniques. • Establish automated testing workflows for analyzing biomolecules such as nucleic acids, proteins, and metabolites, as well as microbial physiological. • Develop standard physical interfaces from automated building to automated testing.	**Achieve automated, cross-scale, multi-modal testing for complex synthetic biosystems, establishing standards for quantitative test data generation, sharing, and integration.** • Establish cross-scale, multi-modal and automated functional testing workflows for complex synthetic biological systems such as multicellular organisms, compatible with different scales from microdroplets to microplates and to fermenters. • Establish automated calibration frameworks to output quantitative data. • Perform automated processing and collection of operation metadata.

Figure 1 Roadmap for process flow and hardware equipment of biofoundries

Current Status		
Completed construction and operation of biofoundry concept prototypes, initially established laboratory information management system prototypes, relying on commercial integrated systems with deep customization for synthetic biology automation research.		
Objective 1: Establish Hardware Equipment, Integrated Software, and Logistics Systems and Standards for Automated Biofoundry Operation		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve automated integration of synthetic biology equipment, develop equipment form factors, interfaces, and data encoding standards, and establish intelligent logistics systems for automated operation.	**Achieve automated hardware integration for 20+ synthetic biology workflows, develop prototype information logistics systems for biofoundries.** • Develop equipment function integration middleware modules to unify equipment interfaces. • Develop self-avoiding AGVs and collaborative robots equipped with vision systems to achieve automated material handling and transport, as well as intelligent target recognition and positioning. • Establish automated storage system for automated material management and replenishment.	**Establish open, advanced standards for instrument form factor, interface, and data encoding; establish an intelligent and information-based logistics system.** • Develop standardized integration technology for synthetic biology laboratories, publish SCADA protocols and physical parameter standards, and develop equipment compliant with new standards and protocols. • Develop intelligent logistics system software. • Achieve port docking and information sharing for across-device, production lines, and factory logistics.

Objective 2: Establish Self-Driving Cloud Laboratories to Achieve Full-Process Informatization, Digitalization, and Intelligence of Synthetic Biology Research		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve information-based generation and execution of experimental plans, conduct comprehensive real-time recording of personnel, equipment, materials, methods, and environment, achieve intelligent handling of operational errors, and connect standard interfaces across all stages of the Design-Build-Test-Learn engineering cycle and biofoundry operations.	**Establish information system prototypes, implement "Science as a Service" concept, achieve structured generation, sharing, and aggregation of synthetic biology big data.** • Achieve fully automated software planning of experimental protocols and optimized scheduling. • Achieve real-time tracking and management of material information. • Automatically conduct target function analysis, customized design, and machine-readable instruction generation, achieving deep integration and interaction between software and hardware.	**Build unattended self-driving biofoundries, achieving large-scale transcendence of "data-information-knowledge-wisdom" in synthetic biology.** • Distributed sensor networks acquire and feed back information in real-time. • Establish standard interfaces for "data flow and control flow interaction". • Conduct comprehensive statistics and visualization of all biofoundry elements and achieve intelligent handling of operational errors.

Figure 2 Roadmap for automated integration and information operation of biofoundries

3.11.4 Technical Pathways

3.11.4.1 Biofoundry Automation Workflows and Hardware

Current Technologies: Automated construction of large DNA fragments requires the development of standardized and modular DNA assembly methods (including Golden Gate, Gibson, TAR, LCR, etc.) [7] and compatible robotic platforms. Existing automated DNA assembly is mainly based on microtiter plate systems, which have issues of high costs and large reagent consumption. The latest trend is to build miniaturized assembly systems based on microfluidics, nanoliter pipetting, and other technologies. However, some molecular cloning operations are difficult to automate— agarose gel electrophoresis involves visual judgment and semi-solid gel cutting, necessitating customized instrument development. Furthermore, DNA sequencing is an important means to verify successful DNA assembly. Combining pooling indexing and high-throughput sequencing technology can greatly reduce the sequencing cost per DNA construct [8].

Automated genetic manipulation of chassis systems is primarily limited to model microorganisms (*E. coli* and *Saccharomyces cerevisiae*), with emerging capabilities for industrial strains including *Corynebacterium glutamicum*, *Bacillus subtilis*, filamentous fungi, and *Streptomyces*[9,10]. To achieve automated chassis operation, it is necessary to develop automation-compatible workflows and hardware equipment for experimental steps including cell culture, genetic transformation, monoclonalization, colony picking, and physiological characterization. For non-model chassis organisms, large-scale customized development of equipment, consumables, and processes is required to address specific environmental conditions (e.g., anaerobic, light-dependent, thermophilic) and special physical-chemical properties (e.g., multinucleate microorganisms, multicellular chassis, irregular colonies, viscous extracellular matrix, etc.).

For functional testing of synthetic organisms, different biological processes such as RNA transcription, protein translation, metabolism, and signal transduction need to be analyzed using multi-modal methods such as optics, chromatography, and mass spectrometry at different scales from microplates to shake flasks to fermenters. To bridge the increasing throughput of automated synthetic biology construction, rapid, parallel, and standardized sample pretreatment needs to be developed. For example, the combination of liquid handling stations, microfluidics and barcode tagging permits thousands of samples to be analyzed for genomic and transcriptomic sequencing with single-cell

resolution. Based on principles such as biosensing and chemical coupling, the functions and performance of synthetic biology systems can be converted into signals such as fluorescence, growth, and ion strength, enabling the efficient, specific, dynamic, and quantitative acquisition of genotype-phenotype correlation data. In addition, advanced imaging methods permit *in situ*, *in vivo* and dynamic observation of engineered organisms. For example high-content imaging screening coupled with high-throughput synthetic biology allows for systemic understanding of complex processes. Furthermore, online optical probes plays an increasingly important role in the parameter control and optimization of fermentation processes.

Objectives and Breakthroughs: For automated and standardized operations of large DNA fragments and multiple chassis systems, develop automated processes and equipment to achieve high-throughput, standardized construction of DNA large fragments and automated, high-throughput, standardized operation of multiple chassis systems.

For multi-modal, cross-scale, automated, and quantitative testing of synthetic biosystems, develop automated processes and equipment to expand automated compatible functional testing methods, establish automated calibration processes, and achieve multi-modal, cross-scale, and quantitative functional characterization of synthetic life systems .

Challenges: Lack of unified standard DNA assembly workflows; design and experimental phases for different DNA constructs still require manual intervention; low automation compatibility of workflows, equipment, reagents, and consumables; lack of hardware equipment and standard procedures for automated operation of megabase (Mb)-scale large DNA fragments; microplate-based DNA assembly faces cost and throughput limitations; lack of design software, genetic tools, and workflows for automated microbial chassis operations; lack of design software, genetic tools, and workflows for automated operation of non-model microorganisms, multicellular plants and animals, and other chassis systems; lack of automated equipment and supporting reagents and consumables compatible with special physicochemical properties and environments; lack of high-throughput sample pretreatment methods and supporting hardware equipment to achieve deep integration of automated construction and testing processes; lack of standardized measurement systems and unified standards for synthetic biology data generation, sharing, and integration; lack of equipment systems supporting structured, cross-scale, multi-modal testing.

Expected Progress Recently: Microliter-scale systems achieve fully automated

assembly of cloned DNA fragments with single production line throughput of up to 10 Mb per day, reducing operation costs by one order of magnitude compared to manual operations; achieve fully automated operation of 10+ microbial chassis with single-clone operation costs reduced by one order of magnitude compared to manual operations; achieve 20+ types of automated functional testing for biomolecules and microorganisms.

Expected Progress by 2030: Nanoliter-scale systems achieve fully automated assembly of cloned DNA fragments with single production line throughput of up to 100 Mb per day, reducing operation costs by two orders of magnitude compared to manual operations; achieve automated operation of 30+ microbial chassis and 5+ plant/animal chassis with single clone operation costs reduced by two orders of magnitude compared to manual operations; achieve automated, cross-scale, multi-modal testing for complex synthetic biological systems and establish standards for quantitative test data generation, sharing, and integration.

Potential Solutions

For automated DNA assembly, improve workflows using 384-well, 1536-well and other microplates as experimental containers, efficiently integrate required instruments using robotic platforms to improve operational efficiency. For enzyme-based systems such as Golden Gate and Gibson Assembly methods, use methods like directed evolution to improve tool enzyme performance and use active learning methods to optimize reaction systems. For cell-based assembly systems of *E. coli* and *S. cerevisiae*, establish automated operation workflows, improve assembly accuracy and efficiency through chassis engineering, and develop modular, multi-round assembly technologies for large DNA fragments. Develop and iteratively optimize unified DNA assembly design standards, develop supporting computer-aided design software, integrate machine learning algorithms to identify DNA sequences and experimental conditions that easily lead to errors and reduced efficiency, and intelligently recommend sequence splitting, conversion, and process improvement solutions.

Improve computer-aided design algorithms, comprehensively apply enzyme and cell assembly systems, and generate high-throughput DNA synthesis and assembly solutions for 100 kb-scale fragments; research automated conjugation transfer and protoplast fusion technologies, develop supporting high-throughput equipment platforms to achieve large fragment DNA delivery via bacterial, yeast, and other intermediate hosts to plant and

animal cells, and further hierarchically assemble into ⩾1 Mb cloned fragments in target cells based on microcell-mediated chromosome transfer. Based on microarray, microfluidic, and other technologies, miniaturize DNA assembly systems to nanoliter scale, develop corresponding workflows and hardware equipment to significantly reduce DNA assembly costs and improve throughput, accuracy, and efficiency. Based on second- and third-generation high-throughput sequencing technologies, develop automated sequencing library construction workflows to achieve rapid verification of 100 kb-scale sequences.

Develop computer design algorithms to generate high-throughput experimental plans for multi-round engineering through optimization and selection of gene modification sequences and gene editing technologies, and automatically design required DNA sequences (such as sgRNAs, homology arms, restriction sites, primers, etc.). By researching and modifying microbial DNA repair mechanisms, transiently changing membrane structures of non-model microorganisms, and developing more effective exogenous DNA transmembrane delivery technologies, expand the application range of CRISPR and other technologies, improve genome editing efficiency by 1-2 orders of magnitude while achieving efficient editing of 3 loci simultaneously; based on microplate and droplet microfluidic systems, develop hardware equipment compatible with automated experiments through multi-device serial connection and single-device integration, add equipment such as high-throughput microbial electroporation, achieve fully automated workflows for microbial chassis culture, transformation, clonal isolation, and genotype verification, and establish standardized operation procedures and reporting specifications.

Develop next-generation endonucleases (new CRISPR systems, Argonaute, etc.), artificial DNA repair machines, and other next-generation gene editing technologies to achieve 5-10 loci parallel genome operations in multiple chassis microorganisms, apply machine learning algorithms to intelligently generate DNA sequences required for multi-round, automated chassis engineering; for special environmental requirements for cell chassis growth such as light, anaerobic, high temperature, and high pressure conditions, develop customized high-throughput photobioreactors and instruments compatible with anaerobic/high salt/high temperature/high pressure conditions and integrated robotic platforms; for chassis systems with irregular morphology and special physicochemical properties such as multinucleate mycelia, spores, protoplasts, plant embryos, adherent cells, and organoids, develop customized supporting equipment such as clone picking instruments and robotic grippers, workflows, reagents, and consumables.

For different testing methods including spectroscopy, mass spectrometry, imaging, and sequencing, develop parallel sample preparation methods based on microplates, microarrays, and microdroplets to support high-throughput, automated biophysical and biochemical characterization of nucleic acids, proteins, metabolites, and their interactions, and automatically conduct multi-omics analysis and physiological characterization at genome, transcriptome, proteome, metabolome, and metabolic flux levels for microbial systems. Based on biosensing, chemical coupling, and other principles, convert target functions and performance into signals such as fluorescence intensity and growth rate to achieve efficient, specific, dynamic, and quantitative acquisition of genotype-phenotype relationship data; develop standardized physical interfaces for "build" and "test" connection, and develop automation-compatible workflows, supporting equipment, reagents, and consumables.

Based on new methods and technologies such as microfluidics, ultrasonic pipetting, optical tweezers, and online sensing, develop new equipment to support cross-scale automated integration from microdroplets and microarrays to microplates and fermenters; establish calibration workflows corresponding to various automated experimental workflows to convert relative measurements to absolute measurements; systematically explore how process parameters (experimental conditions and testing methods, etc.) introduce experimental errors and data noise based on automated platforms, and on this basis, develop standard operating procedures, reporting specifications, and quality control methods, and establish methods for different experiments to generate, share, and integrate synthetic biology big data; establish standard interfaces with synthetic biology databases and knowledge bases for automated collection and structured processing of experimental data and process metadata.

3.11.4.2 Biofoundry Automation Integration and Information Operations

Current Technologies: Intelligent logistics systems for biofoundries still need further exploration, testing, and optimization. Biofoundry hardware facilities can be divided into different levels including workshops, production lines, functional areas, and equipment. Material transfer between functional areas within production lines and between equipment within functional areas can be achieved through collaborative robots to handle sample containers in a fast, repeatable, and position-accurate manner. However, logistics requirements at the workshop and production line levels still lack good solutions. Similar solutions already exist in manufacturing, such as Enterprise Resource Planning

(ERP) and Manufacturing Execution Systems (MES) for controlling material flow. Intelligent logistics systems establish automated synthetic biology laboratory material management and transfer systems to achieve high-throughput equipment, sample, resource, and information management, as well as automated transfer solutions for consumables and samples [3,4].

Laboratory information management systems (LIMS) provide various platform-based practical functions, configuration tools, and workflow engines, enabling laboratories to control their unique workflows and working modes and continuously expand their functional scope as needed. In biofoundries, researchers will be freed to easily set parameters on computer terminals, with robots completing a series of basic operation steps and finally receiving experimental data. The goal of material and information management systems is to achieve comprehensive informatization of synthetic biology research, including target requirement design, expert system decision-making, real-time information acquisition and status detection, and recording and analysis of experimental results. To further support the informatization and digitalization of synthetic biology research processes, biofoundries integrate with computer-aided design, machine learning, and other modules through standardized interfaces, i. e, build "cloud laboratories" that integrating automation, informatization, and biotechnology, using internet sharing to provide high-throughput, standardized synthetic biology research and development capabilities to national and even global synthetic biology needs [3,4].

Objectives and Breakthroughs: For automated biofoundry operation, establish hardware equipment, integrated software, and logistics systems; achieve automated integration of synthetic biology equipment; develop equipment configurations, interfaces, and data encoding standards; establish intelligent logistics systems for automated operation.

Establish self-driving cloud laboratories to achieve comprehensive informatization and digitization of intelligent synthetic biology research; achieve information-based generation and execution of experimental plans; conduct comprehensive real-time recording of personnel, equipment, materials, methods, and environment; achieve intelligent handling of operational errors; connect standard interfaces across all stages of the DBTL engineering cycles and biofoundry operations.

Challenges: Currently limited instruments and equipment are compatible with automated integration, equipment configurations and interfaces from different manufacturers are not unified, functions are incomplete, and data encoding formats

particularly lack unification; some manufacturers' software is closed-source and difficult to integrate; material transfer in foundry operations requires significant manual intervention, with low informatization in the transfer process, relying on manual entry and management.

Communication protocols and physical parameters of automated equipment lack unified standards; biofoundry logistics lack cross-scale integrated operating systems for equipment, production lines, and factories; lack of standard interfaces for "human-machine interaction" and "software-hardware interaction," unable to automatically generate and execute experimental plans or comprehensively record experimental elements including personnel, equipment, materials, methods, and environment in real-time; lack of information systems that can fully integrate with experts, hardware, and software; lack of standard interfaces connecting all stages of the entire process.

Expected Progress Recently: Achieve automated hardware integration for 20+ synthetic biology workflows and develop prototype information logistics systems for biofoundries; establish information system prototypes to implement the "Science as a Service" concept and achieve structured generation, sharing, and aggregation of synthetic biology big data.

Expected Progress by 2030: Establish open, advanced standards for instrument form factors, interfaces, and data encoding; establish intelligent, information-based logistics systems; build unattended self-driving biofoundries, achieving large-scale transcendence of "data-information-knowledge-wisdom" in synthetic biology.

Potential Solutions

Design and develop equipment function integration middleware modules for each type of key equipment, unify equipment interfaces through secondary development, and supplement essential functions; actively incubate and develop suppliers of high-end instruments, connecting with actual needs for customized development; develop autonomous obstacle-avoiding AGV mobile robots to achieve mobility functions, combined with collaborative robots to achieve material grasping and transport, and equip vision systems to achieve target recognition and positioning; establish automated storage modules to achieve automated sample storage design and information management, as well as information interaction with mobile transport robots; provide methods for safely

identifying samples at each step of the workflow and methods for real-time querying of each sample's location.

Leverage biofoundry facility construction opportunities to publish advanced, open-source laboratory equipment SCADA protocols and physical parameter standards; form research ecosystems through academic activities to promote concepts and demand recognition for standardized laboratory equipment; develop and support excellent instrument suppliers to cooperatively develop laboratory automation-specific equipment meeting new standards and protocols; design intelligent logistics system software for synthetic biology research needs, including Warehouse Management Systems (WMS), Warehouse Control Systems (WCS), Automated Guided Vehicle Systems (AGVS), and docking software systems with upper-level systems.

Develop computer algorithms to achieve software-assisted or fully automated experimental planning, including required materials, equipment operation parameters, and testing status, encode these into experimental plans and executable instruction sets to guide automated research equipment operation; apply advanced digital technologies such as visualization, digital twins, and design simulation to online collect and display experimental processes, test results, and equipment status in biofoundries, map various attributes of physical layer equipment to virtual space, stimulate design activities such as simulation, batch replication, and virtual synthesis, significantly reducing the number of experiments, time, and cost in iterative processes; for actual biofoundry operational processes, automatically analyze target functions of synthetic biological systems, customize designs and convert them into laboratory process procedures, generate experimental plans and machine-readable instructions, and interact with physical laboratories through software modules including Advanced Planning and Scheduling (APS), Warehouse Management Systems (WMS), Manufacturing Execution Systems (MES), and Quality Management Systems (QMS) to complete high-throughput construction and testing experiments.

Based on distributed multi-sensor network real-time information acquisition, monitor real-time status of automated hardware platforms and handle exceptions to ensure collaborative and efficient operation of hardware platforms, achieving systematic and structured recording of experimental data and process metadata; based on real-time collected experimental data and process metadata, achieve rolling planning and reupdating of operation instructions through deep learning, forming a closed loop of "design-build-test-learn"; develop expert decision systems where researchers provide

upper-level decisions for highly automated operation and testing process control, complete online optimization of design solutions, and improve foundry success rates; establish standard interfaces for "data flow and control flow interaction" to achieve fully automated control and management of business processes in cloud laboratories, supporting product lifecycle management, physical layer field data collection, and big data machine learning feedback.

3.11.5 Summary

To address the problems of traditional biological experiments being tedious, time-consuming, error-prone, and difficult to scale, the design and construction of biofoundries aim to achieve automated operation of the DBTL cycles by improving standardization and modularization of experimental subjects, methods, and techniques in synthetic biology. However, current limitations include high costs of large DNA fragment manufacturing, limited chassis systems, and few high-throughput functional testing methods. Future key technologies requiring breakthroughs include synthetic biology workflows for automated DNA and chassis operations, automation-compatible instruments, intelligent material transfer robots and control software, and "cloud laboratory" information operation architectures to achieve equipment interconnection, intelligent scheduling, dynamic monitoring, and information integration. Combined with artificial intelligence methods to achieve dynamic optimization of algorithms and models, this will improve synthetic biology research efficiency and expand the scope and scale of biofoundry research subjects. Breakthroughs in these key technologies require collaboration among researchers and engineers from multiple fields including synthetic biology, automation, analytical chemistry, and information technology to conduct multiple rounds of engineering iterative optimization of biofoundries themselves, providing support for basic and applied synthetic biology research and bringing about revolutionary impact.

References

[1] Zhao G P. Synthetic biology: opening a new era of"convergence"research in life sciences. Bulletin of the Chinese Academy of Sciences, 2018, 33(11): 1135-1149.

[2] Ran C, Yongbo Y, Huimin Z. Construction of a synthetic biology manufacturing plant. Science China:

Life Sciences, 2015, 45(10): 976-984.

[3] Ting T, Lihao F, Erpeng G, et al. Automated synthetic biology technology and engineering facility platform. Chinese Science Bulletin, 2021, 66(3): 300-309.

[4] Ting Z, Mengtian L, Fan J, et al. Overview of major scientific and technological infrastructure for synthetic biology research. Synthetic Biology, 2022, 3(1): 184-194.

[5] Cui J M, Zhang B Z, Ma Y F, et al. Engineering platform for synthetic biology research. Bulletin of the Chinese Academy of Sciences, 2018, 33(11): 1249-1257.

[6] Hillson N, Caddick M, Cai Y, et al. Building a global alliance of biofoundries. Nature Communications, 2019, 10(1): 2040.

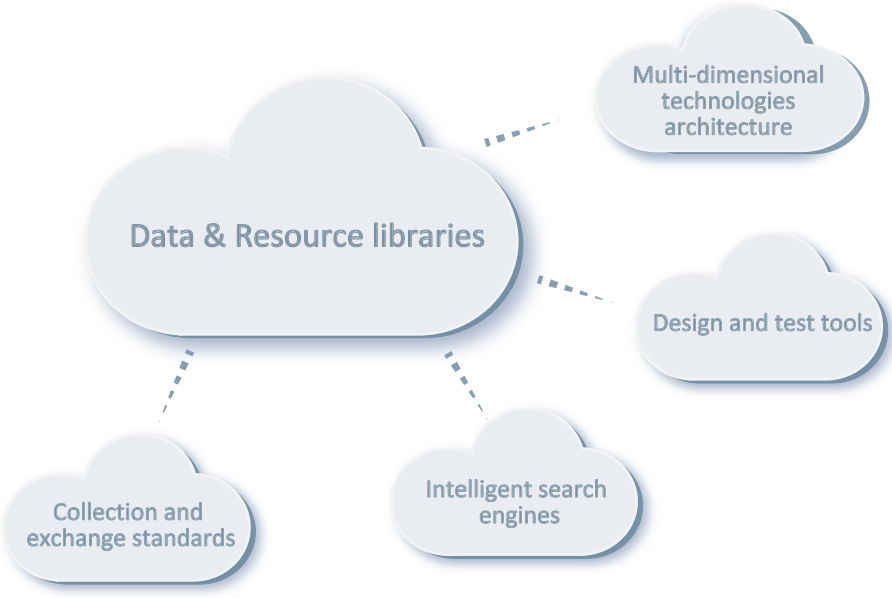
[7] Chao R, Liang J, Tasan l, et al. Fully automated one-step synthesis of single-transcript TALEN pairs using a biological foundry. ACS Synthetic Biology, 2017, 6(4): 678-685.

[8] Shapland E B, Holes V, Reeves C D, et al. Low-cost, high-throughput sequencing of DNA assemblies using a highly multiplexed nextera process. ACS Synthetic Biology, 2015, 4(7): 860-866.

[9] Si T, Chao R, Min Y, et al. Automated multiplex genome-scale engineering in yeast. Nature Communications, 2017, 8: 15187.

[10] Hamedirad M, Chao R, Weisberg S, et al. Towards a fully automated algorithm driven platform for biosystems design. Nature Communications, 2019, 10: 5150.

Biopart Data and Information Platforms



Authors Zhou Zhi-Hua, Yan Xing, Liu Wan, Song Hao

3.12 Biopart Data and Information Platforms

3.12.1 Abstract

Biological components refer to DNA, RNA, or amino acid sequences with specific functions; biological devices refer to functional modules constructed from multiple (two or more) biological components. Biological components and devices (refers to as bioparts hereafter) are the cornerstone of synthetic biology. The engineering nature of synthetic biology determines the necessity and importance of the construction of databases and entity libraries of bioparts. Currently, there is a scarcity of bioparts, most of them lack accurate characterization, and there is a lack of compatibility between bioparts in artificial systems or between bioparts and artificial systems. These are obstacles that must be overcome in the development of synthetic biology. The construction of biopart libraries and information platforms, including the establishment of standardized, high-capacity, and intelligent biopart databases and application platforms, as well as high-quality biopart and chassis entity libraries, will provide important support for the design, research, and application of synthetic biology. The biopart library and information platform will accelerate the convergence of biopart data, physical objects, and design tools based on the principles of data sourcing, multi-level review, resource sharing, information disclosure, information security, and authorized access, and further serve synthetic biology research.

3.12.2 Technical Overview

Biopart data and entity libraries involve establishing data standards for bioparts, collecting and preserving the entities of reported and constructed bioparts, achieving automated storage and retrieval of bioparts; integrating the previous accumulations of scientists in metabolite molecule databases, reaction databases, enzyme databases, omics data, and metabolic networks, organizing the sequence structure, characterization, and related literature information of bioparts, establishing an efficient management system for the biopart database, achieving the structural reconstruction and visual storage of biopart data; integrating new technologies for the prediction, design, assembly, and characterization of bioparts, as well as the correlation of biopart structure and multi-dimensional characterization information, forming corresponding tool packages, and constructing an online sharing application platform for bioparts.

3.12.2.1 Development of a Multi-dimensional Information Infrastructure Support System for Bioparts

Establish an information infrastructure support system related to biopart big data, including multi-dimensional information input, output, interaction query, and search engine platform for bioparts; develop a technical system that links biological catalytic bioparts with gene databases, phenotype databases, and related compound databases; develop data integration for the regulation of bioparts with metabolic networks, regulatory networks, interaction networks, and signal transduction networks, and a data system that maps the multi-dimensional information of bioparts. Therefore, the biopart information infrastructure support system includes the following two aspects.

The first is the construction of the biopart information platform infrastructure. Using the B/S structure and centralized server deployment, users can access the database through a browser. The system also uses a combination of the Linux system and Docker framework, utilizing development languages such as Java, Python, R, and a combination of relational database MySQL and document database MongoDB to achieve large data storage and fast retrieval.

The second is the development of data integration tools, search engines, and retrieval service interfaces for the data retrieval system. Based on the Solr search engine platform, using Solr for full-site search, field weight sorting, querying the MySQL database, and other technical means to achieve interactive query of bioparts, chassis, pathway, and compound information, and to achieve multi-dimensional information input and output of bioparts and chassis. Integrate databases of known molecules (substrates, products, intermediates, etc.), reactions, enzymes, genomes, and chassis cells to associate bioparts with corresponding databases.

3.12.2.2 Key Technologies for the Automated Storage and Retrieval of Biopart Entities

The biopart entity library, as an important infrastructure of synthetic biology, must adapt to the large-scale and automated needs of biopart splicing and assembly in synthetic biology research. The construction of the biopart entity library should focus on developing key technologies that can complete the automated storage and retrieval of hundreds to thousands of bioparts in a short period. These key technologies should include the innovative design of biopart storage units and the corresponding hardware and

software design for the automated retrieval of physical bioparts. In addition to including storage data of physical bioparts, the biopart database should also establish a close association with the functional data of synthetic bioparts and chassis and continue to update to provide important physical support for synthetic biology research and application.

3.12.2.3 Development of a Biopart Application Platform Interface with Multi-dimensional Knowledge Integration

With the in-depth study of the correlation between the function and structure of bioparts and the development of high-performance computing technology, molecular dynamics and big data analysis methods can quickly locate specific functional areas and co-evolutionary sites of bioparts. The computational design, assembly, and modification technology of bioparts will usher in a new stage of development. Use molecular fragment fingerprint software packages, protein sequence description software packages, and corresponding model construction methods, combined with machine learning, to explore mathematical models with better solutions for different technical needs from a large amount of biopart data, thereby establishing mathematical models for biopart pattern recognition, classification models for bioparts and molecular structural characteristics, establishing quantitative biopart performance prediction models, creating new types of biopart feature recognition and rational design technology; using human-computer interaction technology, achieve online testing and feedback of biopart prediction, design, construction, and characterization new technologies, and develop corresponding application interfaces based on this, forming software and tool packages for biopart prediction, design, construction, and characterization, and improve the level of information sharing and application of biopart resources.

3.12.3 Roadmaps

Current Status		
Data standards for bioparts have been established, but the speed of data collection is slow, and there are only relatively simple biopart query systems.		
Objective 1: Continuously Improve the Ability to Collect and Exchange Standardized Biopart Data in the Biopart Library		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Build an efficient standardized biopart data collection and exchange system and quality control system.	**Establish an efficient standardized biopart data collection and exchange system.** • Establish an efficient and automated literature review system. • Vigorously promote and upgrade the automated biopart submission system.	**Build a comprehensive biopart data quality control system.** • By relying on the establishment of a biopart ontology library, a complete quality control system for biopart data is built to improve the data quality of the biopart library.
Objective 2: Build an Intelligent Biopart Interaction Query and Search Engine		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Establish a biopart data interaction retrieval and intelligent search system that meets the needs of different users.	**Develop a biopart data interaction retrieval and intelligent search system that meets the needs of different users.** • For user needs based on text retrieval, develop corresponding retrieval and intelligent search systems. • For user requirements based on functional domain retrieval, develop corresponding retrieval and intelligent search systems. • For user requirements based on the interaction between bioparts and substrates, develop an intelligent search tool based on virtual screening.	**Improve the efficiency of biopart data query and retrieval.** • Improve the algorithm, optimize the intelligent search system for genetic bioparts, develop task partitioning strategies and scheduling pipelines for ultra-large-scale data, further optimize the local alignment module and sequence pattern screening engine, and achieve fast intelligent retrieval of biopart sequence data.

Objective 3: Integration and Mapping of Biopart Multi-dimensional Information Technology System Based on Knowledge Graph		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Integrate and map the multi-dimensional information technology system of bioparts based on the knowledge graph to achieve the mining and application of biopart data.	**Establish the ontology library of biopart data.** • Summarize the characteristics of bioparts through SBOL, SBML, biopart data standards, and other related materials. • Build an ontology library for bioparts to achieve standardized description of biopart information, especially standardized description of functional characterization information.	**Build a knowledge graph of bioparts.** • Establish seven major categories of entities related to catalytic bioparts: sequences, reactions, substrates, products, pathways, and qualitative and quantitative. • Suggest six major relationships: reaction-pathway, reaction-product, substrate-reaction, sequence-reaction, sequence-qualitative, sequence-quantitative • Establish a knowledge network of various attribute data of biopart sequences and functional, qualitative and quantitative.
Objective 4: Construction of a Multi-dimensional Knowledge Integration Biopart Design Application Platform		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
The establishment, continuously and iteratively updating and application of biopart prediction and design technology driven by big data.	**Online deployment and application services of new tools for biopart prediction and design.** • Complete the online deployment of new tools for biopart prediction and design. • Establish a connection between analysis tools and the biopart database to achieve online application services.	**Continuously and iteratively updating and application of biopart prediction and design technology driven by big data in the biopart library.** • Build a training dataset based on the data and physical resources of the biopart library. • Build an engineering platform, combine artificial intelligence to obtain a large amount of standardized data, expand the trial and error space. • Optimize quantitative characterization methods to effectively guide the design, construction, testing, and learning of synthetic biological systems.

Figure 1 Roadmap for construction of standardized, high-capacity, and intelligent biopart databases and application platforms

Current Level		
The number of bioparts and chassis is relatively small and generally lacks standardized functional data.		
Objective 1: Establish Standardized, Automated, and High-throughput Biopart and Chassis Functional Testing Methods		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Establish standardized, automated, and high-throughput biopart and chassis functional testing methods.	**Establish standardized testing methods for bioparts and chassis.** • Establish standardized functional testing methods for important types of regulatory elements (promoters, RBS, terminators, etc.) and catalytic elements (P450 and glycosyltransferases, etc.). • Establish standardized functional testing methods for chassis such as *E. coli*, yeast, Streptomyces, and filamentous fungi.	**Achieve automation and high throughput of biopart and chassis functional testing.** • With the help of automation and high-throughput equipment, upgrade the previously established standardized testing methods to automated and high-throughput versions.
Objective 2: Establish and Promote the Collection and Sharing Mechanism of Bioparts and Chassis		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Establish an efficient biopart sharing mechanism and promote its application.	**Establish an efficient biopart sharing mechanism and promote its application.** • Establish a hierarchical biopart and chassis physical objects and data management system to ensure the security and sharing of bioparts and chassis physical objects and data. • Set different security levels for bioparts and chassis physical objects and data at different levels for different ranges of publicity.	**Further innovate and optimize the sharing mechanism.** • Innovate the sharing mechanism of bioparts and chassis, for example, make comprehensive use of the advantages of blockchain technology, achieve secure sharing of biopart data through decentralized management of biopart and chassis data and metadata, and promote the circulation and exchange of bioparts and chassis physical between different research units, and continuously improve the value of bioparts and chassis during circulation.

Objective 3: Construction of High-quality Synthetic Biology Biopart and Chassis Entity Libraries		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Complete the construction of high-quality synthetic biology biopart and chassis entity libraries.	**Initially establish high-quality biopart and chassis entity libraries.** • Cooperate with advantage research units to focus on collecting regulatory elements and catalytic elements concerned in synthetic biology research, and carry out standardized functional testing. • Further accumulate the quantity and category of bioparts and chassis entity through carrying out standardized biopart and chassis testing services.	**Complete the construction of high-quality synthetic biology biopart and chassis entity libraries.** • Promote the cooperation and sharing mechanism of bioparts and chassis to establish a long-term mechanism for continuously accumulating high-quality biopart and chassis entity; establish the interrelationships between the above six major categories of entities.

Figure 2 Roadmap for building high-quality biopart and chassis entity libraries

3.12.4 Technical Pathways

3.12.4.1 Construction of Standardized, High-capacity, and Intelligent Biopart Databases and Application Platforms

Current Technologies: Catalytic biopart data standards have been established, and corresponding biopart data submission systems have been developed, with the ability to collect and exchange standardized biopart data. Standardized biopart data can be collected and exchanged through multi-database integration, literature review, and biopart data submission, but the efficiency of data review and exchange is still relatively low, and the speed of biopart data increase is still lagging behind the speed of biopart data publication[1]; a relatively simple sequence-based biopart query system has been established, and annotation processes for important bioparts such as P450 and glycosyltransferases have been developed, but it still cannot meet the needs of users of different categories.

Although preliminary collection of biopart data (including sequence, reaction, qualitative and quantitative functional data, etc.) has been carried out, the correlation between various biopart information has not been established, making it impossible to discover new knowledge from existing data. With the emergence of a large number of biopart prediction and design tools such as AlphaFold 2 and RoseTTA, the ability to design and develop bioparts has been significantly improved. However, these new tools often have problems such as difficult installation and high hardware resource requirements, which greatly limit their popularization and application. Integrating these design tools with the biopart database to form a multi-dimensional knowledge integration biopart design application platform will provide important support for synthetic biology research.

Objectives and Breakthroughs: Establish an efficient standardized biopart data collection and exchange system; build an intelligent biopart interaction query and search engine; integrate and map the multi-dimensional information technology system of bioparts based on the knowledge graph; construct a multi-dimensional knowledge integration biopart design application platform.

Challenges: At present, most researchers are used to submitting biopart sequence data directly to databases such as NCBI, but these databases do not accept functional data related to bioparts (such as catalytic reaction and kinetic data), and can only obtain functional data through literature review. Due to the slow speed of manual literature

review, the speed of biopart data collection and exchange is seriously delayed. With the rapid development of synthetic biology, a large amount of biopart data is generated and enters the biopart database. Due to the lack of effective data quality control means, it is impossible to effectively control biopart data including sequence, species, characterization information, etc., which is an urgent problem that needs to be solved to improve the quality of biopart data. In addition, single-function biopart intelligent search tools are difficult to meet the needs of different users. With the large amount of biopart data, the large increase in the number of users, and the development of various retrieval strategies, the current algorithm can no longer meet the needs of fast biopart retrieval.

Although synthetic biology open language (SBOL) and systems biology markup language (SBML) are related to the standardized description of biopart information, they focus more on the sequence and pathway description of bioparts, as well as the format conversion between different biopart data standards, and lack standardized description methods for qualitative and quantitative characterization information of bioparts, which hinders the quality control of biopart data as well as data mining and application [2,3]. Through the standardization of biopart data and the establishment of the ontology library, although it is possible to understand the various attributes of bioparts, there is still a lack of understanding of the correlation between various attributes, which limits the cognition and exploration of the intrinsic laws of biopart big data.

The existing codes for biopart-related prediction, design, construction, and characterization tools cannot be deployed and interactively used on the website, which greatly limits their promotion, use, and further optimization. Moreover, artificial intelligence requires a large amount of training datasets, and the current synthetic biology research has problems such as wide data sources, heterogeneous data forms, and insufficient high-quality training data, which leads to the difficulty of effective training of artificial intelligence models under sparse supervision with small datasets, thus hindering the effective application of artificial intelligence technology in biopart research.

Expected Progress Recently: Establish an efficient standardized biopart data collection and exchange system. Develop a biopart data interaction retrieval and intelligent search system that meets the needs of different users; establish the ontology library of biopart data; carry out the online deployment and application services of new tools for biopart prediction and design.

Expected Progress by 2030: Build a comprehensive biopart data quality control system; improve the speed and efficiency of biopart data query and retrieval; establish a knowledge graph of bioparts; achieve the continuous iteratively updating and application of new technologies for biopart prediction and design driven by big data in the biopart library.

Potential Solutions

Continuously improve the ability to collect and exchange standardized biopart data in the biopart library, establish an efficient automated literature review system, and combine deep learning and natural language processing technology to establish a review platform to efficiently obtain biopart data, especially function-related data, from the literature; vigorously promote and upgrade the automated biopart submission system to enable biopart data to directly enter the biopart database at the same time as the publication of the paper.

With the establishment of the biopart ontology library, build a comprehensive biopart data quality control system to improve the data quality of the biopart library. For text-based retrieval user needs, achieve interactive query of biopart data through the Solr full-text search engine based on Lucene and combined with the Boolean retrieval model; for function structure domain-based retrieval user needs, establish a search engine based on function structure domain annotation and retrieval; for biopart-substrate interaction retrieval user needs, develop a virtual screening intelligent search tool based on protein structure prediction and substrate interaction.

Improve algorithms and optimize the intelligent search system for genetic bioparts, for example, design a parallel computing model and corresponding data structure for homologous search execution of multiple sequence alignments, develop task division strategies and scheduling pipelines for large-scale data, deeply optimize local alignment modules and sequence pattern screening engines to achieve intelligent retrieval of biopart sequence data [4,5].

Summarize the characteristics of bioparts through SBOL, SBML, biopart data standards, and other related materials, and build an ontology library for bioparts to achieve standardized description of biopart information, especially standardized description of functional characterization information.

Build a knowledge graph of bioparts: by establishing seven major categories of entities related to catalytic bioparts: sequences, reactions, substrates, products, pathways, qualitative data, and quantitative data, and six major relationships: reaction-pathway, reaction-product,

substrate-reaction, sequence-reaction, sequence-qualitative data, sequence-quantitative data, to establish a knowledge network of various attribute data of biopart, such as sequences, qualitative and quantitative functional data[6].

Rewrite the existing tool codes to enable these tools to be interactively used and visualize results on the biopart library website; establish a connection between the biopart database and analysis tools, open interfaces, and allow users to select the biopart data they are interested in for process analysis to solve the problems of biopart structure simulation and functional analysis with one click.

Based on the data and physical resources of the biopart library, build a training dataset; build an engineering platform, combine artificial intelligence to obtain a large amount of standardized data, expand the trial and error space, optimize quantitative characterization methods, and effectively guide the design, construction, testing, and learning of synthetic biological systems.

3.12.4.2 Construction of High-quality Biopart and Chassis Entity Libraries

Current Technologies: Although different research institutions have developed a large number of biopart and chassis functional testing methods, due to the lack of unified standards, it is difficult to quantitatively compare the performance of biopart with similar functions from different sources, and it is also difficult to accurately judge the compatibility between biopart and chassis cells. It is necessary to repeatedly undergo trial and error to construct modules, pathways, and cell factories that meet expectations, which brings great difficulties to the selection and application of biopart and chassis. Therefore, the construction of high-quality biopart and chassis entity libraries needs to be based on standardized functional testing methods.

In terms of biopart and chassis sharing, there are still problems of small sharing quantity and narrow sharing range, which affect the development and utilization efficiency of biopart and chassis. The BioBricks Foundation and other biopart libraries are actively trying to use decentralized management methods to accelerate the sharing and distribution of biopart chassis data and physical objects [7].

High-quality biopart and chassis are the premise for synthetic biology to ultimately achieve the design of biological systems from scratch, and they have a fundamental role in synthetic biology research and application. However, due to the large amount of human and materia resources required for centralized management of biopart and chassis physical objects, and the cumbersome sharing process of these physical biopart and

chassis, biopart libraries are trying to build biopart and chassis entity libraries according to the principle of "decentralization", and manage some biopart and chassis physical objects in a "decentralized" manner through "information preservation" to allow biopart and chassis physical objects to be stored in different research institutions, but the physical information of biopart and chassis is saved in a unified format in the central biopart library, which is conducive to the collection and sharing of physical bioparts and chassis.

Objectives and Breakthroughs: Establish standardized, automated, and high-throughput biopart and chassis functional testing methods; establish and promote the collection and sharing mechanism of biopart and chassis physical objects; build high-quality synthetic biology biopart and chassis entity libraries in combination with "decentralized" management.

Challenges: It is necessary to establish standardized testing methods suitable for different types of bioparts and chassis to meet the needs of synthetic biology research and application. With the increase of biopart and chassis physical objects, the workload of functional testing is becoming larger and larger, and traditional instruments and equipment can no longer meet the measurement needs. Innovative bioparts and chassis have high application value, causing the current situation where bioparts and chassis are not easy to share. At present, the main biopart and chassis physical objects are still concentrated in their own projects or units, and only some bioparts and chassis are scattered in cooperative research units.

Expected Progress Recently: Establish standardized testing methods for bioparts and chassis; establish an efficient biopart sharing mechanism and promote its application; collect biopart and chassis cells that have important applications in the field of synthetic biology research and application, and carry out standardized functional testing on them to initially establish high-quality biopart and chassis physical libraries.

Expected Progress by 2030: Achieve automation and high throughput of biopart and chassis functional testing. Further innovate and optimize the sharing mechanism; establish a long-term mechanism for continuously accumulating high-quality biopart and chassis physical objects, and build a synthetic biology biopart and chassis entity library with significant influence in the field.

Potential Solutions

Establish standardized functional testing methods for important types of regulatory

elements (promoters, RBS, terminators, etc.), catalytic elements (P450 and glycosyltransferases, etc.), and chassis (*E. coli*, yeast, Streptomyces, and filamentous fungi); with the help of automation and high-throughput equipment, upgrade the previously established standardized testing methods to automated and high-throughput versions.

Through a hierarchical biopart and chassis physical and data management system, ensure the security and sharing of biopart and chassis physical and data. Set different security levels for biopart and chassis physical and data at different levels for different ranges of publicity, and maximize the use of the FAIR (Findable, Accessible, Interoperable, and Reusable) principle to share biopart and chassis data.

Innovate the sharing mechanism of bioparts and chassis, for example, make comprehensive use of the advantages of blockchain technology, achieve secure sharing of biopart data through decentralized management of biopart and chassis data and metadata, and promote the circulation and exchange of bioparts and chassis physical between different research units, and continuously improve the value of bioparts and chassis during circulation.

Cooperate with advantage research units to focus on collecting regulatory elements, catalytic elements, chassis concerned in synthetic biology research, and carry out standardized functional testing; combine the principle of "decentralization" to achieve flexible management of biopart and chassis physical objects, try to reduce the operating cost of the central biopart library as much as possible, and improve the sharing efficiency of bioparts and chassis; promote the cooperation and sharing mechanism of bioparts and chassis to establish a long-term mechanism for continuously accumulating high-quality biopart and chassis physical objects.

3.12.5 Summary

In the future, the construction of biopart libraries and information platforms will focus on two aspects: standardized, high-capacity, and intelligent biopart databases and application platforms, and high-quality biopart and chassis entity libraries, focusing on the standardization of bioparts and chassis data, promoting the collection, organization, and sharing of bioparts and chassis, providing more intelligent interactive retrieval and design tools, and carrying out regulatory research on the intrinsic attributes of bioparts and chassis based on big data of bioparts and chassis, providing the training datasets

needed for artificial intelligence research of bioparts, and gradually accelerating the convergence of biopart data, physical objects, and design tools on the basis of data sourcing, multi-level review, resource sharing, information disclosure, information security, and authorized access, so as to serve synthetic biology research and application.

References

[1] Liu W, Yan X, Shen X, et al. Research progress of biopart databases. Acta Microbiologica Sinica, 2021, 61(12): 3774-3782.

[2] Mao Z, Wang R, Li H, et al. ERMer: a serverless platform for navigating, analyzing, and visualizing *Escherichia coli* regulatory landscape through graph database. Nucleic Acids Research, 2022, 50(W1):W298-W304.

[3] McLaughlin J A, Beal J, Misirli G, et al. The synthetic biology open language (SBOL) Version 3: Simplified data exchange for bioengineering. Frontiers in Bioengineering and Biotechnology, 2020, 8:1009.

[4] Hucka M, Bergmann F T, Chaouiya C, et al. The systems biology markup language (SBML): language specification for level 3 version 2 core release 2. Journal of Integrative Bioinformatics, 2019, 16(2):20190021.

[5] Lu H, Diaz D J, Czarnecki N J, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature, 2022, 604(7907): 662-667.

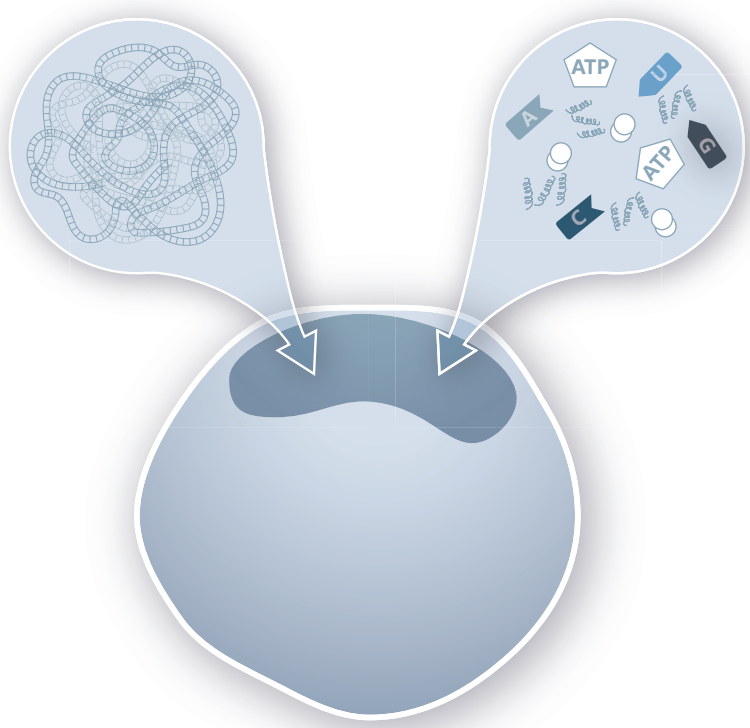
[6] Konermann S, Lotfy P, Brideau1 N J, et al. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. Cell, 2018, 173(3): 665-676.

[7] Kahl L, Molloy J, Patron N, et al. Opening options for material transfer. Nature Biotechnology, 2018, 36(10): 923-927.

4 Application Prospects

Based on the core concepts of "build to learn" and "build to use" of synthetic biology, applications of synthetic biology can be summarized in two aspects. First, synthesizing artificial cells from the bottom up with biological macromolecules and other components to understand how functions emerge during the origin of life and evolution. Second, driving iterative improvement of biotechnology and the transformation of the biomanufacturing industry, enabling synthetic biology to empower new quality productive forces, shaping the future bioeconomy, thereby contributing to human health and sustainable development.

Build to Learn



Authors

Liu Chen-Li, Miao Wei, Fu Xiong-Fei, Liu Xing-Guo, Zhong Chao, Hu Zheng, Jin Fan, Yan Fei, Lou Chun-Bo, Yu Tao, Gan Hai-Yun, Si Tong, Li Xue-Fei, Fu Mei-Fang, Qi Fei

4.1 *De Novo* Synthesis of Single Cells

4.1.1 Abstract

Cells are the basic units of life activities and synthesizing single-celled life artificially (it will be referred to "synthetic cell" in the following context) from non-living materials is a fundamental issue in life sciences. Cells possess highly dynamic and nonlinear regulatory characteristics, and *de novo* synthesis of single cells is an extremely challenging scientific and engineering issue, requiring the development of various enabling technologies in synthetic biology. How to rational design synthetic cell with partial or complete cellular functions is the core scientific question; how to efficiently synthesize biological macromolecules such as nucleic acids, proteins, and membrane lipids, and ultimately synthetic cells, are the key technical issues. At the same time, opportunities lie in predictable quantitative design and artificial intelligence, which provides theoretical guidance for designing synthetic cells, and standardized high-throughput biofoundries accelerate the iterative process of constructing synthetic cell. We point out that the integration of these technologies will help achieve this scientific and engineering goal. It is believed that the area of synthetic cells is approaching a tipping point, and it is expected that the single replication of synthetic cells can be achieved in the near future, and multiple autonomous replications can be achieved by 2030.

4.1.2 Technical Overview

4.1.2.1 Connotation of Synthetic Cells

It aims to synthesize structures with basic life characteristics, i.e. lipid vesicles encapsulating genetic material DNA, which is capable of autonomous growth, replication, and division, operating a limited number of their life cycles. The core proteins are synthesized autonomously by the synthetic cells, and the key chemical molecules (such as RNA, amino acids, etc.) can be provided exogenously. These synthetic cells have no complex organelles.

4.1.2.2 Enabling Technologies Required for Synthetic Cell

To achieve the self-replication of synthetic cells, each functional module (mainly

including cell growth, DNA replication, DNA separation, and cell division) needs to be reconstructed separately and then integrated. Currently, researchers have proposed various solutions for individual functional modules, and achieving function coordination is the biggest challenge. To address this challenge, the unified design and integration of individual functional modules should be under the umbrella of the goal of functional coordination. This process requires the integration of various enabling technologies: the theory of functional complex systems and the development of data and algorithm platforms provide theoretical guidance for designing synthetic cells; Technologies of DNA sequencing/ synthesis/assembly synthesize the necessary genetic sequences for synthetic cells; new generation gene editing technologies are expected to accelerate the analysis of natural cell coordination mechanisms, guiding the design of synthetic cells; protein engineering provides controllable, multifunctional building blocks for synthetic cells; genetic circuit engineering ensures the spatiotemporal orderliness of gene expression in synthetic cells, which is an important prerequisite for achieving functional coordination; cell-free systems provides a powerful tool for the efficient expression of proteins; the construction of membraneless organelles provides the necessary compartmentalization and programmability for complex cellular processes. In addition, construction of synthetic cells also needs to overcome the following technical bottlenecks: cell membrane artificial synthesis technology (artificially constructing cell membrane models with controllable shape and adjustable components, which will encapsulate various biological macromolecules); efficient material and energy synthesis technology, etc. The above technologies have been introduced separately in other chapters of this book, and this section will focus on the main tasks of constructing synthetic cells and the expectation/requirements for each enabling technology.

4.1.3 Roadmaps

Current Status		
In terms of cell membrane synthesis technology, various cell membrane models have been developed, the phospholipid synthesis pathway has been designed and reconstructed, and vesicle growth has been achieved through methods such as vesicle fusion; In terms of biomass growth, gene transcription and translation have been realized inside phospholipid vesicles [1].		
Objective 1: Autonomous and Controllable Growth of Synthetic Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Synthetic of cell membranes and biomass growth.	**Controllable synthetic cell growth.** • Establish a controllable and efficient technology to prepare phospholipid vesicles. • Reconstitute channel transporter proteins to achieve controllable material transport. • *In situ* synthesize phospholipids based on DNA to achieve controllable and efficient synthetic cell growth.	**Coordination of synthetic cell membrane and biomass growth.** • Reveal the coordination mechanism of membrane and biomass growth in living systems. • Reconstitute membrane and biomass growth coordination module. • Achieve coordination of synthetic cell membrane and biomass growth.

Objective 2: Continuous Synthesis of Core Components and Precursor Materials of Synthetic Cells		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Synthetic cell material supply system.	**Semi-synthesis of core metabolic products (non-DNA dependent).** • Develop a simple and efficient system for amino acids, sugars, and lipids synthesis. • Achieve semi-synthesis of core metabolic products (non-DNA dependent).	**Achieve *de novo* synthesis of core metabolic products (DNA dependent).** • Reveal the design principles and coupling mechanisms for quantitative allocation and efficient transformation of metabolic pathways in living systems. • Establish the technical system for synthesizing core cellular metabolites from simple compounds based on whole cell or cell-free systems. • Achieve *de novo* synthesis of core metabolic products (DNA dependent) .
Objective 3: Achieve Continuous Synthesis of ATP Driven by Chemical Energy or Light		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Synthetic cell energy supply system.	• Achieve the ATP synthesis system driven by cascade enzyme reaction mediator and electron transport chains.	• Achieve ATP synthesis driven by respiratory chain complexes and light.

Figure 1 Roadmap for synthetic cell growth

Current Status		
The *in vitro* DNA replication systems derived from *E. coli* and phage Phi29 have been established [4].		
Objective 1: Precise Control of DNA Replication Initiation Time and Dosage		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Precise, complete, and controllable replication of 100 kb circular DNA molecules; coupling of transcription-translation-replication to achieve self-sustained DNA replication.	• Construct a DNA-encoded replication control system to control the expression time and dosage of key proteins involved in replication initiation, achieving DNA replication initiation *in vitro* and its precise control.	• Construct an autonomous DNA replication system within phospholipid vesicles to achieve temporal regulation of DNA replication initiation.
Objective 2: Precise Control of DNA Replication Extension and Termination		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Develop controllable DNA replication termination strategies.	• Design and construct a DNA replication extension regulation system for developing synthetic single-cell life form.	• Coordinate DNA replication with transcription by prioritizing DNA replication over RNA transcription to reduce replication-transcription conflicts, and ensure genome stability.

Figure 2 Roadmap for synthetic cell DNA replication

Current Status		
Two methods are used to achieve synthetic cell division: *in vitro* reconstruction of cell division-related proteins, but so far, controllable phospholipid vesicle division has not been achieved; division of phospholipid vesicles by using physical, chemical, and mechanical methods [5]. In terms of DNA separation, the ParMRC system has been reconstructed *in vitro*, achieving plasmid DNA separation [6]; a theoretical model of entropy increase-induced DNA separation has been proposed [7].		
Objective 1: Rational Design and Construction of Synthetic Cell Division Complex		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve controllable synthetic cell division.	• Controllable synthetic cell division site determination and division septum construction.	• Identify key proteins and regulatory factors that mediate complete cell membrane division, achieving controllable synthetic cell division based on the cell division complex.
Objective 2: Rational Design and Construction of Membrane Proteins Mediating Synthetic Cell Division		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve controllable synthetic cell division independent of biochemical mechanisms.	• Synthesize and design membrane proteins, peptides, and their functional analogs to mediate synthetic cell division.	• Study the physicochemical properties of synthetic membranes and non-equilibrium chemistry to achieve controllable synthetic cell division independent of biochemical mechanisms.

Objective 3: Autonomous Separation of DNA within Synthetic Cell		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Achieve controllable DNA separation in synthetic cells.	• Achieve semi-synthetic DNA separation based on DNA separation complexes, as well as controlled DNA separation within vesicles based on the principle of entropy increase.	• Identify and reconstruct key proteins and regulatory factors in charge of the DNA separation process; Achieve controllable DNA separation within vesicles based on DNA separation complexes.

Figure 3 Roadmap for synthetic cell division

Current Status		
A digital whole-cell model based on the minimal cell JCVI-syn3A (493 genes) of Mycoplasma has been established [8].		
Objective 1: Analysis of Living Cell Functional Coordination Mechanisms and Synthetic Cell Reconstruction		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
By exploring living cell coordination mechanisms and thea utonomous interaction between intelligent agents and experiments, achieve an efficient closed loop of "design-build-test-learn", and ultimately achieve functional coordination in synthetic cells.	• Decouple the functional coordination mechanisms of living systems through cell-machine fusion; and reconstruct a functional coordination system responsive to DNA replication initiation *in vitro*.	• Achieve the complete functional coordination among cell membrane growth, protein synthesis, DNA replication, and cell division of synthetic cell.
Objective 2: Construction of a Theoretical Model and Standardized Characterization Data Form for Digital Cells with Periodic Coordination		
Expected Breakthroughs	**Expected Progress Recently**	**Expected Progress by 2030**
Build a standard database and establish digital cells.	• Establish a standard database and data-driven modeling software required for constructing synthetic cell and its theoretical modeling.	• Establish a digital twin cell and analyze the principle of function emergence.

Figure 4 Roadmap for functional coordination of synthetic cells

4.1.4 Technical Pathways

4.1.4.1 Growth of Synthetic Cells

Current Technologies: In terms of cell membrane synthesis technology, various cell membrane models have been developed, bacterial phospholipid synthesis-related proteins have been reconstructed *in vitro*, and electrostatic interaction and other methods have been used to induce phospholipid vesicle fusion to achieve vesicle growth. In terms of biomass increase, an *in vitro* transcription-translation system TXTL has been used to achieve gene transcription and translation with DNA as a template within phospholipid vesicles [1]. In terms of substances synthesis technology, efficient substances synthesis has been achieved through genetic regulation of living cells [2], In terms of energy synthesis technology, ATP has been artificially synthesized using various materials and reducing power [3].

Objectives and Breakthroughs: Controllable autonomous growth of synthetic cells; coordination of controllable synthetic cell membranes and biomass growth; continuous synthesis of core components and precursor of synthetic cells; establishment of a synthetic cell substance supply system that combines external supply and internal synthesis; achieve continuous synthesis of ATP driven by chemical energy or light energy; ATP generation through a cascade enzyme system loaded on the synthetic membrane.

Challenges: The prepared phospholipid vesicles are of uneven size, and the encapsulation efficiency is unclear; the artificial cell membrane lacks key ion channels or membrane proteins; the rate and efficiency of integrating phospholipid molecules into phospholipid vesicles are uncontrollable; the characterization method of phospholipid vesicles growth is unclear, and obvious phospholipid vesicle growth has not been achieved so far. The expression efficiency of multiple proteins and membrane proteins in cell-free systems is low, making it difficult to achieve efficient expression of genetic circuits with multiple functions within vesicles; controllable expression of genes in synthetic cells; the growth of synthetic cell membranes and the increase in biomass have not been coordinated. At present, the synthesis and modification of core metabolic products are mainly carried out in living cells, and how to achieve the synthesis of core metabolic products within synthetic cells requires innovative methods. A large number of components need to be systematically optimized and integrated, the final effect of the combination of multiple enzymes is unclear, and whether the integration level and

efficiency of the system can be improved needs further study; the combination of glycolytic cascade enzymes with multi-phase separation has not been tried before; there are still some technical risks in the ordered assembly of respiratory chain components into functional electron transport chains and research on ATP synthesis driven by respiratory chain complexes and light energy is still rare.

Expected Progress Recently: Controllable synthetic cell growth; semi-synthesis of core metabolic products (DNA in dependent), involves providing sustained material supply to synthetic cells by encapsulating core metabolic substances cascade enzyme system within liposomes; construct a ATP synthesis system driven by electron transport chain and mediated by cascade enzyme reaction.

Expected Progress by 2030: Coordination of synthetic cell membrane growth and biomass growth; *de novo* synthesis of core metabolic substances (DNA dependent), that is, achieve controllable expression of genes within synthetic cell membrane-bound vesicles, synthesizing biological macromolecules within synthetic cells from simple carbon sources; ATP synthesis driven by respiratory chain complexes and light energy.

Potential Solutions

Establish an efficient and controllable technology for preparing phospholipid vesicles; controllably integrate phospholipid molecules into phospholipid vesicles by using phospholipid transport proteins or channel proteins, which can be synthesized by protein modification, protein directed evolution, and other means; optimize exogenous substrate enzymatic reactions and *in situ* DNA-based phospholipid synthesis pathways to achieve efficient phospholipid vesicle growth.

Use automated platforms to screen and optimize experimental parameters for cell-free protein synthesis on a large scale; develop temperature-controlled/photo-controlled genetic circuit methods for membrane protein expression; reveal the coordination mechanism of membrane and biomass growth in living systems; verify and extract transferable coordination modules for membrane and biomass growth.

Systematically investigate biosynthesize pathways for different biomacromolecules, and divide the big system into different small systems, optimize them independently, and then couple and optimize them to form a systematic enzyme system for biomacromolecules synthesis.

Study the design principles and coupling mechanisms for quantitative allocation,

efficient transformation and utilization of metabolic pathways, analyze the transport, utilization, and transformation of metabolic substrates, intermediates, and terminal waste; establish a technical system for synthesizing core cellular metabolites from simple compounds in whole cell or cell-free systems.

Each enzyme in the system is purified, and a functional cascade enzyme system is finally obtained through continuous optimization of the quality and proportion of individual enzymes. A cyclic transfer chain is established to gradually realize the electron transfer process on the membrane.

Construct a respiratory chain complex system in primitive eukaryotes. Integrate and achieve the production of reducing power and ATP driven by respiratory chain complexes. Establish a light-driven system for the generation of reducing power and ATP.

4.1.4.2 DNA Replication of Synthetic Cells

Current Technologies: *In vitro* DNA replication systems derived from *Escherichia coli* and phage Phi29 has been established [4].

Objectives and Breakthroughs: An *in vitro* DNA replication system with controllable replication initiation time and dosage; precise, complete, and controllable replication of 100 kb circular DNA molecules; coupling of transcription-translation-replication to achieve self-sustained DNA replication; precise control of DNA replication extension and termination; develop controllable DNA replication termination strategies by learning replication termination strategies of living systems.

Challenges: Design primer-independent, controllable DNA replication initiation modules, which can precisely control replication initiation sites and replication activity; construct controllable and autonomous DNA replication within the membrane-bound vesicles, and achieve precise control of replication activity; establish a robust DNA replication speed control system; living organisms use complex DNA damage repair systems to deal with the impact of conflict events on the genome, which is difficult to reconstruct *in vitro*.

Expected Progress Recently: *In vitro* construction and precise control of DNA replication initiation and extension regulation system for developing synthetic cells.

Expected Progress by 2030: Construct an autonomous DNA replication system within synthetic cells and achieve temporal regulation of DNA replication initiation; coordinate DNA replication with transcription, reduce replication-transcription conflict events, and ensure genomic stability.

Potential Solutions

Study the replication initiation and activity regulation mechanism for the two *in vitro* replication systems of rolling circle and replication fork; construct and optimize the *in vitro* transcription-translation systems, and precisely control replication initiation sites and time; construct autonomous replication systems based on phi29, HSV-1, and other bacterial and viral replication complexes. Perform exhaustive testing on the DNA replication speed control systems; establish DNA replication systems within membrane-bound vesicles, and single and multi-factor replication extension speed control systems. Start from overall design, coordinate optimization of replication protein, genome structure, and transcription systems, and avoid the impact of conflict events on genomic stability from the source.

4.1.4.3 Division of Synthetic Cells

Current Technologies: Two methods are used to achieve synthetic cell division: The first one is *in vitro* reconstruction of cell division-related proteins, but so far, division of phospholipid vesicles has not been achieved; the second one is the division of phospholipid vesicles by using physical, chemical, and mechanical methods [5]. In terms of DNA separation, the ParMRC system has been reconstructed *in vitro*, achieving plasmid DNA separation [6]; a theoretical model of entropy increase-induced DNA separation has been proposed [7].

Objectives and Breakthroughs: Rational design and construction of synthetic cell division machinery; study the division mechanisms of living cell, and achieve controllable synthetic cell division by reconstructing cell division-related proteins *in vitro*; rational design and construction of membrane proteins that mediate synthetic cell division; achieve controllable synthetic cell division independent of biochemical mechanisms; autonomous separation of DNA within synthetic cell membranes, achieving controllable DNA separation in synthetic cells.

Challenges: It is difficult to precisely control the division location; *in vitro* reconstruction of FtsZ cell division proteins cannot achieve membrane division; rational design of membrane proteins, peptides, and their functional analogs that can mediate membrane division; use the physicochemical properties of synthetic membranes and

non-equilibrium chemistry to achieve controllable division of synthetic membranes; achieving equal DNA separation depends not only on the required key proteins and regulatory factors but also on precise spatiotemporal control within synthetic cells, which has a large degree of uncertainty; the biochemical mechanism of bacterial chromosome separation is currently unclear, and *in vitro* reconstruction lacks theoretical guidance.

Expected Progress Recently: Controllable division site determination and division septum construction; Achieve synthetic cell division through artificially designed membrane proteins and their functional analogs; achieve controllable DNA separation within synthetic cell based on the semi-synthesis of Par protein complexes and entropy increase principle.

Expected Progress by 2030: Identify and reconstruct key proteins and regulatory factors that mediate complete cell membrane division, achieving controllable artificial cell division based on cell division complexes; study and characterize the physicochemical properties and non-equilibrium chemistry of synthetic membranes to achieve controllable synthetic cell division independent of biochemical mechanisms; controllable DNA separation within synthetic cell based on DNA separation complexes.

Potential Solutions

Conduct large-scale screening of protein components and concentrations using synthetic biology facilities; develop new division site determination systems; use transcriptomics, proteomics, and molecular biology technologies to identify key proteins and regulatory factors that mediate complete cell membrane division, and reconstruct them *in vitro* to achieve synthetic cell division; study the mechanism of membrane invagination mediated by peptides and their analogs, and use this as a reference to rationally design and construct membrane proteins that mediate synthetic cell division; study the physicochemical properties and non-equilibrium chemistry of synthetic membranes to achieve controllable synthetic cell division independent of biochemical mechanisms; develop characterization technologies for cell deformation and synthetic cell membrane dynamic structure, providing quantitative basis for experiment design; use *E. coli* as a model organism to construct a semi-synthetic system for chromosome DNA separation; construct a physically constrained vesicle-enclosed model to regulate environmental crowding and DNA structure to achieve entropy increase-induced DNA separation; identify key proteins and regulatory factors of the DNA separation process,

and reconstruct protein complexes *in vitro* to mediate DNA separation.

4.1.4.4 Functional Coordination of Synthetic Cells

Current Technologies: A new model of "division permits" for bacterial cell division control has been proposed [8]; a digital whole-cell model based on the minimal cell JCVI-syn3A (493 genes) of Mycoplasma has been established [9].

Objectives and Breakthroughs: Analyze living cell functional coordination mechanisms, design and reconstruct them in synthetic cells; through exploring living cell coordination mechanisms, and the autonomous interaction between intelligent agents and experiments, achieve an efficient closed loop of "design-build-test-learn", and ultimately achieve functional coordination of synthetic cells; construct a theoretical model and standardized characterization data form for digital cells with periodic coordination; build a standard database and establish a digital cell.

Challenges: The specific proteins corresponding to functional coordination and their corresponding protein regulation and interaction are still unknown. There may be unknown coupling between the coordination components, leading to mutual interference of coordination mechanisms, and make it impossible to achieve complete coordination of multiple modules in synthetic cells. Large-scale automated platforms generate a large amount of data, but there is currently no effective technical method to systematically utilize these data for dimensionality reduction. Moreover, the existing whole-cell models do not consider the impact of coordination regulation on the cell cycle. The existing whole-cell models have few metabolic reactions and almost no regulatory processes; there are significant differences in the spatiotemporal scales of various processes within the cell, making it challenging to establish and analyze models. Different processes may require different mathematical representations to more accurately reflect their dynamic characteristics within cells. Therefore, in order to better understand and simulate biological processes within cells, future research needs to focus on addressing the above issues and developing more comprehensive and accurate whole-cell models.

Expected Progress Recently: Decouple the functional coordination mechanisms of living systems through cell-machine fusion, and reconstruct a functional coordination system responsive to DNA replication initiation *in vitro*; establish a standard database and data-driven modeling software required for constructing synthetic cell experiments and theoretical modeling.

Expected Progress by 2030: Achieve complete coordination of cell membrane

growth, protein synthesis, DNA replication, and cell division *in vitro*; establish a digital twin cell, that is, establish a digital version of the artificial synthetic cell, describe all cellular metabolic and regulatory processes with kinetic equations, and achieve simulation of the cell cycle.

Potential Solutions

Achieve an efficient closed loop of "design-build-test-learn" through autonomous interaction between intelligent agents and experiments; decouple the functional coordination mechanisms of living systems, and verify and extract transferable functional coordination modules.

Build an automated platform with cell process detection capabilities based on synthetic biology facilities; use high-throughput automated biofoundry to replace different components and characterize their degree of coordination; direct evolution and intelligent analysis for the complete coordination of cell membrane growth, protein synthesis, DNA replication, and cell division of synthetic cells.

Develop effective data analysis processes and methods, extract key regulatory nodes and patterns in the cell cycle based on the big data generated from automated platforms, and introduce relevant regulatory modules into the digital cell for reasonable parameter exploration, providing theoretical guidance for the introduction of coordination genetic circuits into synthetic vesicles; gradually establish a coarse-grained weak regulatory whole-cell model, a cross-scale functional coordination digital twin cell, and a coarse-grained functional coordination digital twin cell.

4.1.5 Summary

Whether it is possible to construct a self-replicating synthetic cell from non-living materials is a fundamental issue in life science. The artificial synthesis of macromolecules into single-celled life forms will break the boundary between "non-life" and "life" and help answer important scientific questions about the origin of life and biological evolution. However, the essence of the functional emergence from biomacromolecules to single-celled life has not yet been revealed, and how to coordinate different functional modules to form a complete "growth-replication-division" synthetic cell cycle is the bottleneck faced by the field. The integration of predictable quantitative design and

standardized high-throughput biofoundries provides a promising solution to these bottlenecks.

Expanded Reading

[1] Bhattacharya A, Cho C J, Brea R J, et al. Expression of fatty acyl coa ligase drives one-pot *de novo* synthesis of membrane-bound vesicles in a cell-free transcription-translation system. J Am Chem Soc, 2021, 143:11235-11242.

[2] Yu T, Zhou Y J, Huang M T, et al. Reprogramming yeast metabolism from alcoholic fermentation to lipogenesis. Cell, 2018, 174:1549-1558.

[3] Jia Y, Li J B. Reconstitution of FoF1-ATPase-based biomimetic systems. Nat Rev Chem, 2019, 3:361-374.

[4] Olivi L, Berger M, Creyghton R N P, et al. Towards a synthetic cell cycle. Nat Commun, 2021, 12(1):4531.

[5] Steinkuhler J, Knorr R L, Zhao Z, et al. Controlled division of cell-sized vesicles by low densities of membrane-bound proteins. Nat Commun, 2020, 11(1):905.

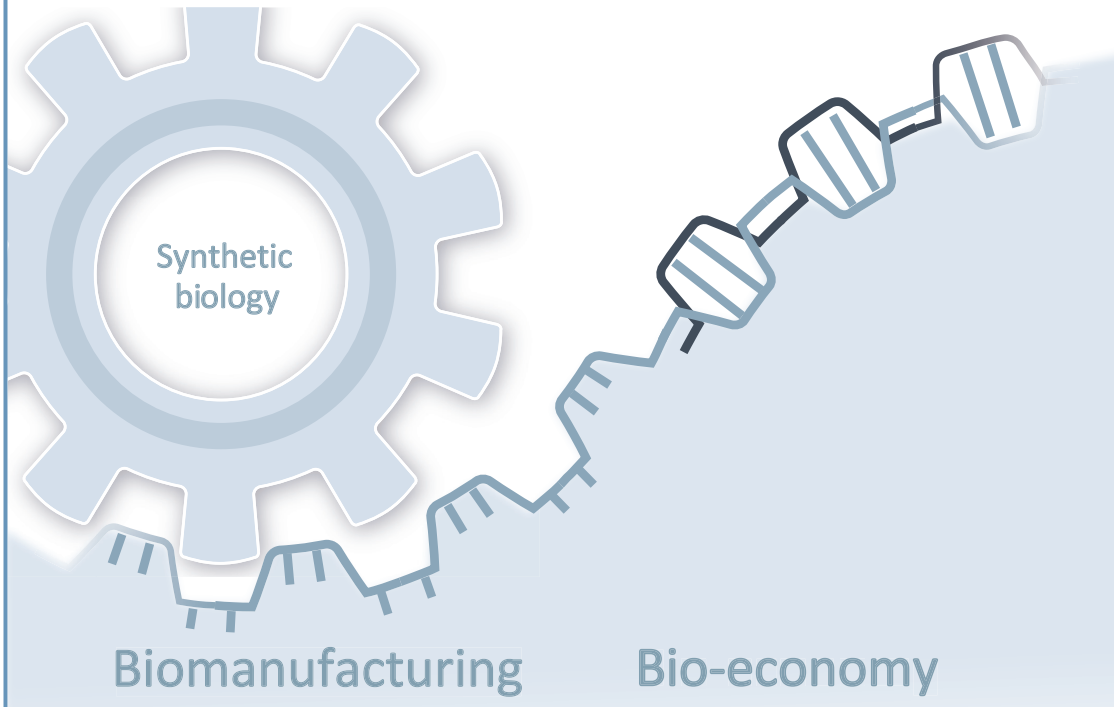
[6] Garner E, Campbell C S, Weibel D B, et al. Reconstitution of DNA segregation driven by assembly of a prokaryotic actin homolog. Science, 2007, 315(5816): 1270-1274.

[7] Jun S, Wright A. Entropy as the driver of chromosome segregation. Nat Rev Microbiol, 2010, 8: 600-607.

[8] Thornburg Z R, Bianchi D M, Brier T A, et al. Fundamental behaviors emerge from simulations of a living minimal cell. Cell, 2022, 185(2):345-360.

[9] Zheng H, Bai Y, Jiang M, et al. General quantitative relations linking cell growth and the cell cycle in *Escherichia coli*. Nat Microbiol, 2020, 5(8), 995-1001.

Build to Use



Authors

Lin Zhang-Lin, Wang Qin-Hong, Dai Zong-Jie, Zhu Zhi-Guang, Liu Tao, Bai Wen-Qin, Feng Jin-Hui, Zheng Hong-Cheng, Li Jin-Gen, Hu Qiang, Yuan Shu-Guang, Zhao Yong, Luo Xiao-Zhou, Wei Ping, Xie Zhen, Huang He, Zheng Hao, Zhong Chao, Ma Ying-Fei, Pan Hong, Qin Jian-Hua, Lin Min, Zhang Li-Xin, Wang Jin, Yao Bin, Zhu Jian-Kang, Zhou Jing-Wen, Jin Cheng, Song Mao-Yong, Xu Ping, Jiang Jian-Dong, Zhou Ning-Yi, Zhang Cheng-Cai, Shen Yue, Yang Yi, Zhang Xian-En, Jiang Hui-Feng, Wei Xin-Li

4.2 Industrial Applications

4.2.1 Abstract

The production method of material processing and synthesis using synthetic biology is characterized by cleanliness, efficiency, and sustainability. It can reduce the impact of industry on the ecological environment, reshape the development model of carbon-based civilization, trigger new industrial revolutions, and lead new industrial patterns and economic forms. The industrial application of synthetic biology is becoming an important driving force in biomanufacturing and global reindustrialization. It focuses on building a bio-economy, breaking through key core technology systems such as industrial chassis cell design, synthesis, regulation, and optimization, achieving disruptive technological innovations in high-performance industrial enzymes, fine and special chemicals, bulk renewable chemical products, bio-based degradable new materials, natural products, and artificial biological utilization of carbon dioxide, promoting the engineering and industrial application of synthetic biology.

It is expected that in the near future, a batch of important synthetic biology manufacturing industrial products will be formed, with an industrial scale reaching nearly 2 trillion yuan; it is expected that by 2030, a new generation of sustainable technology systems for synthetic biology industrial applications will be established, with a number of advantageous biotechnologies and products emerging in the international mainstream market, and the industrial scale will break through 5 trillion yuan.

4.2.2 Application Directions

4.2.2.1 Industrial Enzyme Design and Efficient Expression

Industrial enzymes have good application performance, high stability, heat resistance, acid resistance, and surface activity resistance, with broad application prospects. A relatively complete industrial enzyme protein efficient expression chassis cells have been constructed, with high enzyme activity and low fermentation costs. It is expected that by 2030:

Objective 1: Achieve AI-assisted iterative design of enzymes;

Objective 2: Achieve the creation of high-performance industrial enzyme products;

Objective 3: Achieve the creation of a new generation of industrial enzyme production strains.

4.2.2.2 Industrial Chassis Cell Genome Synthesis

The complete genome synthesis of *E. coli* and the construction of yeast genome Mb-level chromosomes have been achieved; in terms of functionality, the minimization of prokaryotic genomes, codon simplification to 7 codons, and synthetic chromosome rearrangement and evolution of eukaryotic unicellular yeast have been achieved. It is expected that by 2030:

Objective 1: Develop industrial chassis cell genome construction technology to achieve the synthesis and application of new industrial chassis cell genomes;

Objective 2: Develop high-throughput genome editing technology for non-model industrial chassis cells.

4.2.2.3 Fine Chemical Biological Synthesis

At present, chemical synthesis and biocatalysis are highly integrated, and biocatalysis has become the primary choice for the synthesis of fine chemicals, with high synthesis efficiency and low pollution. It is expected that by 2030:

Objective 1: Analyze and clarify the key enzyme structures, catalytic mechanisms, and chiral selection mechanisms for the synthesis of fine chemicals such as chiral amines and chiral alcohols, and achieve efficient synthesis;

Objective 2: Clarify the natural metabolic pathways of steroids and achieve efficient synthesis of steroid fine chemicals.

4.2.2.4 Microbial Recombination Synthesis of Natural Products

Microbial recombination synthesis of natural products has made significant progress, with the synthesis pathways of important complex natural products such as artemisinin and morphine being resolved and heterologous synthesis achieved; the artificial synthesis of farnesene, artemisinin, and stevia has been industrially applied. It is expected that by 2030:

Objective 1: Explore key enzymes in natural product biosynthesis and complete the resolution of important natural product biosynthesis pathways;

Objective 2: Construct heterologous microbial cell factories for natural products and

achieve large-scale production.

4.2.2.5 Biological Synthesis of Bulk Chemicals

The biological manufacturing of bulk chemicals has a complete technology development and industrialization system, and a relatively systematic intellectual property layout has been formed in key strains. It is expected that by 2030:

Objective 1: Achieve the transformation of chemical synthesis of new bulk chemicals towards biological synthesis;

Objective 2: Achieve iterative innovation and creation of core strains for traditional bulk chemicals, and the transformation of traditional fermentation raw materials to non-grain raw materials.

4.2.2.6 Biological Manufacturing of Biodegradable Materials

Progress has been made in the design, substitution, and biological synthesis of bio-based monomers for polymer materials such as bio-based polycarbonates, and the green production of degradable materials has been achieved. It is expected that by 2030:

Objective 1: Achieve efficient biological synthesis of various polymer material monomers;

Objective 2: Achieve efficient direct biological synthesis of multifunctional protein-like, polyamino acid, and other polymer materials.

4.2.2.7 Artificial Fixation and Transformation of Carbon Dioxide

Preliminary artificial customized carbon dioxide to more than a hundred complex molecules biological synthesis routes have been opened up, and large-scale application demonstrations from carbon dioxide to fuels and proteins have been achieved. It is expected that by 2030:

Objective 1: Form a batch of key technologies for efficient biological utilization and transformation of non-biological carbon, and achieve large-scale application of several artificial photosynthetic biological carbon fixation and artificial electrical energy biological carbon fixation systems;

Objective 2: Design key biological carbon fixation components from scratch, and achieve the design and construction of artificial carbon fixation pathways and systems;

Objective 3: Establish carbon biological manufacturing technology to achieve directed synthesis of chemicals from carbon dioxide as a raw material.

4.2.2.8 Algal Carbon Fixation Transformation

A research system and platform for bioenergy and high added-value molecules have been established, with the industrial microalgae synthesis biology model species represented by Chlorella, and an international research cooperation network for energy microalgae synthesis biology has been established, laying the foundation for the design and construction of efficient, low-cost, and scalable photosynthetic carbon fixation cell factories. It is expected that by 2030:

Objective 1: Achieve rational design and systematic modification of algal chassis cells, creating several kinds of traceless industrial microalgae cell factories;

Objective 2: Develop photobioreactors suitable for different engineering algal strains, and establish an algal large-scale efficient and stable culture process package;

Objective 3: Achieve demonstration and promotion of algal carbon fixation engineering and application.

4.2.3 Summary

The use of synthetic biology to efficiently synthesize bulk fermentation products, fine chemical products, scarce pharmaceutical products, etc., provides a new approach for traditional industries to break through resource and environmental constraints. The synthetic biology-supported green biomanufacturing industry is becoming a rapidly developing strategic emerging industry, leading new industrial patterns and economic forms. However, current synthetic biology research is still mainly focused on the design and modification of natural organisms, and most of its products are natural compounds. In contrast, the vast majority of chemicals do not have natural biosynthetic pathways, and there are very few reports on the creation of synthetic pathways from scratch. This will be a challenge faced by the future development and industrial application of synthetic biology, but it also highlights the huge potential of synthetic biology. The development of technologies such as artificial intelligence and big data will create new possibilities for synthetic biomanufacturing. The comprehensive integration of technologies such as synthetic component library construction, pathway large-scale analysis, high-throughput assembly and optimization of biological pathways, and tuning of artificial systems will

make it possible to construct entirely new artificial synthetic pathways and produce new types of compounds. The application of synthetic biology in industry will continue to reduce the costs of bulk fermentation products, fine and pharmaceutical chemicals, renewable chemicals and polymer materials, natural products, etc., promoting the sustainable development of human economy and society.

4.3 Medical Applications

4.3.1 Abstract

Synthetic biology offers transformative opportunities for the medical and pharmaceutical fields, with the potential to significantly improve disease prevention and cure rates, and a potential economic scale in the trillions. However, as the complexity of diseases increases, the difficulty and cost of new drug development have risen sharply, and the success rate is on a downward trend. Synthetic biology and artificial intelligence (AI) bring revolutionary technological means to new drug research and development. For example, AI empowers drug target discovery and compound screening processes, greatly enhancing the efficiency of new drug research and development; advanced metabolic engineering provides more controllable large-scale preparation for pharmaceutical compounds; gene editing can cure genetic diseases, and cellular engineering can construct more efficient cancer immunotherapy and stem cell therapy, etc.

It is expected that in the near future, breakthroughs will be made in ultra-large-scale AI-assisted drug design and screening, mRNA vaccine design and synthesis and delivery systems for various viruses and tumors, intelligent drug-controlled release, degradable and absorbable biomaterials, controllable intestinal flora treatment, engineered bacteriophage drugs, and new tissue engineering drugs. By 2030, it is expected that AI design of new functional pharmaceuticals and medical proteins will be achieved, ton-scale preparation of natural drugs (such as paclitaxel, artemisinin, etc.), genetic disease gene therapy, anti-viral infection and tumor treatment mRNA vaccines, cellular drugs and cell therapies (such as stem cells, CAR-T, etc.), new tissue engineering products (such as living biomaterials, etc.), and clinical applications of intestinal microorganisms and engineered bacteriophage drugs.

4.3.2 Application Directions

4.3.2.1 AI-assisted Design and Manufacturing of Protein Drugs

Teams represented by Google DeepMind, which created AlphaFold, have achieved great success in protein 3D structure prediction. AI-assisted protein *de novo* design is emerging, laying an important foundation for the screening of target proteins and the design of protein drugs. It is expected that by 2030:

Objective 1: Achieve precise AI prediction of target proteins, with the root mean square deviation (RMSD) of target protein prediction models compared to experimental values being less than 1.0 Å;

Objective 2: Achieve AI modification of protein functions, and AI design of new functional proteins;

Objective 3: Achieve tens of billions of efficient ultra-large-scale AI drug screenings.

4.3.2.2 Synthesis of Medical Biochemicals

Currently, the synthesis of some medical biochemicals has been industrialized. It is expected that by 2030:

Objective 1: Achieve efficient biosynthesis of products such as artemisinin and rapamycin at the ton scale;

Objective 2: Achieve efficient biosynthesis of products such as paclitaxel and dactinomycin through heterologous pathway reconstruction.

4.3.2.3 Efficient Synthesis of mRNA Drugs

mRNA drugs have broad application prospects in the treatment of tumors, rare genetic diseases, metabolic diseases, cerebrovascular diseases, and prevention of infectious diseases. mRNA vaccines developed by companies such as Moderna and BioNTech have been widely used during the COVID-19 pandemic. In addition to linear mRNA, new mRNA drug technologies such as replicating RNA and circular RNA are also rapidly developing. It is expected that by 2030:

Objective 1: Achieve differential expression of hundreds of cell types, tissues, organs, and tumors through the design and optimization of mRNA sequences;

Objective 2: Achieve nucleotide modification and efficient synthesis of mRNA, reducing synthesis costs by hundreds of times;

Objective 3: Develop new targeting delivery systems for mRNA, targeting dozens of human cell types, important tissues and organs, and typical lesion sites such as tumors.

4.3.2.4 Cellular Drugs and Cell Therapies

Cellular drugs have broad potential application value in the treatment of autoimmune diseases, tumors, metabolic diseases, infections, and age-related diseases. In terms of the safety and long-term compatibility of cell therapies, the industry is in the process of transforming from trial and error to rational design, and research institutions and clinical organizations have developed a few preclinical and clinical-related technologies. In terms of active cellular drug production, a complete industrial system and technological research and development capabilities are already in place; in terms of production quality control, close cooperation with regulatory agencies is in place. However, the current cost of cell therapy technology remains high, limiting its industrial development. It is expected that by 2030:

Objective 1: Initially achieve quality control of cellular drug production systems and promote the widespread clinical application of cell therapy;

Objective 2: Establish gene editing and genome stability optimization and efficient detection technologies for active cellular drugs;

Objective 3: Establish online detection and control technologies for active cellular drugs within the physiological and pathological environment of the body;

Objective 4: Achieve production quality control of active cellular drugs compatible with recipient cells to reduce production costs.

4.3.2.5 Intestinal Microbiota Regulation

Intestinal microbiota are closely related to health status and the development of various chronic diseases. Companies worldwide are developing engineered strains for chronic diseases related to metabolism and the digestive system and have launched clinical trials, with several products currently in the middle stage of clinical trials and achieving breakthrough progress. It is expected that by 2030:

Objective 1: Develop a gene engineering microbial drug technology platform for the treatment of major chronic diseases using engineered live biological agents, providing one-stop services from research and development to production for clinical needs;

Objective 2: Develop dozens of engineered live biological agents that can synthesize

and secrete drug molecules (such as protein drugs) in the intestine, achieving treatment of more refractory major chronic diseases.

4.3.2.6 Design and Application of Protein Medical Materials

The emergence of AI biological computing tools has laid the foundation for the *de novo* design of functional protein materials, which have been able to design intelligent materials with initial smart response capabilities, showing potential in fields such as stem cells and regenerative medicine, tumor targeted therapy, and drug delivery. It is expected that by 2030:

Objective 1: Achieve rational design of medical protein materials;

Objective 2: Achieve application of medical protein materials in fields such as drug delivery, tumor targeted diagnosis and treatment, molecular imaging and diagnosis;

Objective 3: Build preparation and production platforms for regenerative biomaterials of human tissues and organs, establish large-scale preparation and development platforms for bio-intelligent bionic materials, and achieve large-scale intelligent manufacturing of medical protein materials.

4.3.2.7 Bacteriophage Therapy

Bacteriophages represent a promising solution against bacterial drug resistance. By redesigning and constructing engineered bacteriophages, their biosafety and bactericidal value can be further improved. Current research is mainly focused on a limited number of model bacteriophages. It is expected that by 2030:

Objective 1: Achieve genome reconstruction of engineered bacteriophages for the treatment of refractory and chronic tuberculosis;

Objective 2: Complete clinical trials of engineered bacteriophage therapies and obtain market approval;

Objective 3: Achieve efficient mining, biosynthesis, clinical trials, and market approval of bacteriophage antimicrobial proteins.

4.3.2.8 Construction of Biomimetic Drug Delivery Systems

Create drug delivery systems based on natural or quasi-living entities (such as exosomes, liposomes, etc.) and use inherent *in vivo* pathways for targeted drug delivery. Their advantages include the ability to overcome complex *in vivo* environments and

multiple barriers, higher druggability, and the ability to achieve multiple targeting functions. It is expected that by 2030:

Objective 1: Achieve the construction of new quasi-living biomimetic drug delivery systems;

Objective 2: Achieve the construction of natural living entity biomimetic drug delivery systems;

Objective 3: Achieve the construction of synthetically modified living entity biomimetic drug delivery systems.

4.3.2.9 Regenerative Medicine and Organ Repair

Stem cell therapy has been experimentally and early applied in the treatment of many diseases, but the cell source and safety still lack necessary guarantees. Biological material substitutes such as bone, corneal have good biocompatibility and have been clinically applied. However, there are still deficiencies in aspects such as the degradation rate of materials and the matching of cellular functions. Existing tissue organ engineering products are mostly cell-free devices, and true tissue organ replacement is not yet achievable. It is expected that by 2030:

Objective 1: Establish a standardized stem cell bank and achieve industrial application of stem cells in regenerative medicine;

Objective 2: Achieve the application of transplant materials in organ repair through the organic combination of autologous cells, tissues, and biomaterials;

Objective 3: Combine synthetic biology to enhance the physiological compatibility, biological function, and intelligence of tissue organ engineering products, and achieve clinical application of tissue organ engineering products in organ repair.

4.3.2.10 Gene Editing Therapy

Gene editing shows great potential in disease treatment. However, overall, the safety, applicability, and clinical technology of somatic cell and stem cell editing still need to be strengthened, and a large-scale gene editing system is yet to be developed. It is expected that by 2030:

Objective 1: Develop clinical gene editing systems and achieve treatment of several known genetic diseases;

Objective 2: Based on successful somatic cell genome editing, develop delivery

vectors, gene editors, and editing systems for stem cell genome editing;

Objective 3: Achieve large-scale preparation of therapeutic gene editing systems;

Objective 4: Explore treatment pathways for polygenic hereditary diseases.

4.3.3 Summary

New drug development is facing the dual dilemma of high costs and unstable expected returns. Synthetic biology and artificial intelligence (AI) bring new technological means to new drug development. For example, AI empowers drug target discovery and compound screening processes, greatly enhancing the efficiency of new drug development; advanced metabolic engineering provides more controllable large-scale preparation for pharmaceutical compounds; new generations of delivery systems with clinical significance increase the possibilities of mRNA vaccines in disease treatment; cell therapy, living drug treatment, intestinal microbiota regulation, and bacteriophage therapy open up new avenues for disease diagnosis and treatment; regenerative medicine brings new hope for tissue organ repair, reconstruction, and regeneration, among which the integration of new techniques such as stem cells and synthetic biology will play an important role in the development of regenerative medicine. The rapid iterative development of synthetic biology will further stimulate its huge potential in the medical field and benefit human life and health.

4.4 Agricultural and Future Food Applications

4.4.1 Abstract

The application of synthetic biology in agriculture and food fields mainly includes drought (salinity) resistance, efficient nitrogen fixation/ photosynthesis, crop genetic editing breeding, artificial alternative proteins, and new types of synthetic food, healthy aquaculture, and agricultural formulations. Advanced cell factories and artificial microbial communities will facilitate the rational design and green manufacturing of future food, promoting the development of a new generation of agricultural formulations and healthy breeding technologies.

It is anticipated that in the near future, breakthroughs will be achieved in areas such as extremophile resistance element libraries, artificial resistance circuits, artificial nitrogen fixation enzyme systems, new photosynthesis systems, crop genetic editing

breeding technology, and the construction of food-grade high-version chassis cell factories, obtaining a batch of specialized microorganisms and enzymes for feed, fertilizers, biological control, and detoxification. It is expected that by 2030, breakthroughs will be made in the creation of new stress-resistant plant germplasm, the application of efficient intelligent joint nitrogen fixation systems, autonomous nitrogen fixation of plant cell factories, the construction of new intelligent carbon cycles for high-efficiency CO_2 recycling, the creation of new crop germplasm with high yield and nutrition, alternative protein recombination and product processing; and the demonstration of efficient and large-scale industrial manufacturing of key food components in synthetic food.

4.4.2 Application Directions

4.4.2.1 Efficient Stress Resistance and Disease Resistance

Significant progress has been made in the identification of plant stress signals, transmission mechanisms, signal compounds, stress resistance molecular regulatory mechanisms, construction of stress resistance element libraries, and the relationship between plant stress signals and yield and quality formation. This has laid an important theoretical and technical foundation for the artificial design of high stress resistance, high yield, and quality crops. It is expected that by 2030:

Objective 1: Achieve the design of intelligent plant stress circuits;

Objective 2: Achieve coordinated control of crop stress resistance, high yield, and quality;

Objective 3: Achieve breeding and application of crops and microorganisms with artificial high-efficiency and intelligent stress resistance circuits;

Objective 4: Achieve the design and breeding application of artificial circuits and modules for automatic detection and self-healing of plant infections.

4.4.2.2 Efficient Nitrogen Fixation

Significant research progress has been made in the utilization of nitrogen-fixing microbial resources, genome evolution, metabolic network analysis, interaction between root-associated microbial communities and hosts, construction of artificial nitrogen fixation systems, and the biology of nitrogen fixation structures, laying an important

theoretical and technical foundation for the agricultural application of artificial nitrogen fixation systems. It is expected that by 2030:

Objective 1: Create an artificial high-efficiency intelligent joint nitrogen fixation system at the root interface;

Objective 2: Create an artificial nodule nitrogen fixation system for non-leguminous grain crops;

Objective 3: Create autonomous nitrogen-fixing microorganisms and plants artificially.

4.4.2.3 Efficient Photosynthesis

Synthetic biology has made a series of important progress in improving light energy absorption, reducing light energy loss, enhancing carbon fixation efficiency, and creating photosynthesis system-new material composites, laying the foundation for the modification of photosynthetic pathways. It is expected that by 2030:

Objective 1: Achieve artificial modification and optimization of natural photosynthetic pathways;

Objective 2: Integrate artificial photosynthetic pathways with intelligent materials systems;

Objective 3: Reconstruct efficient photosynthetic pathways in chassis crops.

4.4.2.4 Crop Gene Editing Breeding

Gene editing technology, as an emerging biological breeding technique for the directed creation of new crop germplasm, has prominent advantages such as efficiency and precision targeting. Combined with synthetic biology principles, it is expected that by 2030:

Objective 1: Establish a set of innovative gene editing chassis tools, and form an integrated gene editing technology system;

Objective 2: Develop new crop germplasm with enhanced photosynthesis, biological nitrogen fixation, and stress tolerance in soybeans, wheat, corn, and potatoes;

Objective 3: Build and optimize green synthetic biology platforms for crops.

4.4.2.5 Future Foods and Food Ingredients

Artificial alternative proteins (such as whey protein, egg protein, microbial protein, insect protein, etc.) and new synthetic food and ingredient industries (such as fats, sugars,

starch, flavors, and additives, etc.) are developing rapidly, with a series of key manufacturing technologies, numerous listed companies and brands, and relatively sound risk assessment and policy regulations. It is expected that by 2030:

Objective 1: Explore and discover new food flavor components or structures;

Objective 2: Optimize cell factory production of future food ingredients to achieve low-cost fermentation or culture production of future foods and food raw materials;

Objective 3: Establish new food processing technologies, optimize food ingredient processing and structure, and achieve food processing applications of future foods and food raw materials.

4.4.2.6 Healthy Breeding

Currently, synthetic biology is relatively behind in breeding-related core strains and key technologies, and the application of new major products such as feed, fertilizers, and biological control is pending development. It is expected that by 2030:

Objective 1: Discover several new feed enzymes dedicated to animal intestine and body health regulation, and achieve their promotion and application;

Objective 2: Achieve larger-scale microbial synthesis of feed amino acids, proteins, and fats;

Objective 3: Develop several detoxifying synthetic microorganisms and enzyme preparations, and achieve their promotion and application.

4.4.2.7 Agricultural Formulations

Agricultural formulations refer to chemical or biological formulations used in agricultural production to increase crop yields, control pests and diseases, and regulate growth. They are widely used in the agricultural field, including pesticides, fertilizers, growth regulators, soil conditioners, etc. Currently, the construction of biological synthetic innovation systems for agricultural formulations lacks efficient universal chassis cells. It is expected that by 2030:

Objective 1: Achieve the creation of artificial pathways and efficient universal chassis cells for agricultural formulations;

Objective 2: Achieve synthetic biological manufacturing of major agricultural formulation products.

4.4.3 Summary

Synthetic biotechnology is expected to provide more effective solutions for the "old three issues" in agriculture (stress resistance, photosynthesis, nitrogen fixation) and the creation of new germplasm. Developing synthetic bioparts and chassis organisms for small molecule manufacturing and protein (enzyme) expression will contribute to the green manufacturing of a new generation of future foods, healthy breeding, and agricultural formulations. Sorting out relevant policies and regulations will broaden the commercial application and market opportunities of synthetic biotechnology in agriculture and food fields, thereby better ensuring food security.

4.5 Environmental Applications

4.5.1 Abstract

Environmental biotechnology mainly includes environmental biomonitoring, pollutant biodegradation, and environmental bioremediation. Synthetic biology has the advantages of customization and modularization, which can break through the long strain selection cycles, poor directionality, and instability of advantageous traits in traditional environmental biotechnology. By employing the research strategy of "component creation-circuit assembly-system reconstruction," to construct intelligent, efficient, and safe synthetic biological individuals or multicellular systems for the monitoring, degradation, and remediation of traditional pollutants (such as polycyclic aromatic hydrocarbons, pesticides, etc.) and new pollutants (such as drugs, estrogens, etc.); combined with the actual needs and bottleneck issues of pollution scenarios, provide customized solutions, achieve automated, high-throughput environmental monitoring and the transformation of pollutants from danger to safety, waste to treasure, and harmless strategies, promote technological innovation in related application fields, and provide new momentum for the development of a green economy.

It is expected that in the near future, the construction of ultra-evolutionary element libraries will be completed, forming a batch of relatively mature technical systems; it is expected that by 2030, synthetic biology will achieve large-scale application in the environmental field.

4.5.2 Application Directions

4.5.2.1 Environmental Biomonitoring

Artificial synthetic biological sensors have been successfully applied to the detection of various real environmental pollutant molecules, such as pesticides, plasticizers, synthetic hormones, heavy metals, and the novel coronavirus. It is expected that by 2030:

Objective 1: Achieve the construction of a library of high-sensitivity sensors for the rapid identification of new pollutants in real environments;

Objective 2: Utilize machine learning assistance to achieve the construction of efficient whole-cell biological sensors;

Objective 3: Achieve the construction of high-throughput biological sensor arrays and the connection with environmental health effect assessment systems.

4.5.2.2 Pollutant Biodegradation

Currently, bioparts have been created mainly for pollutants such as polycyclic aromatic hydrocarbons, halogenated aromatic hydrocarbons, alkanes, pesticides, and a few high molecular weight polymers, and metabolic pathways have been reconstructed to construct artificial multicellular systems. However, there is still a lack of effective general theories and tools. It is expected that by 2030:

Objective 1: Achieve safe degradation of difficult-to-degrade and new pollutants;

Objective 2: Achieve efficient intelligent synthesis of pollutant degradation microbial communities.

4.5.2.3 Environmental Bioremediation

Current research mainly focuses on the optimization and regulation of indigenous microbial communities in reactors, with few reports on the application of synthetic microbial communities. It is expected that by 2030:

Objective 1: Achieve customized construction and application of nitrogen and phosphorus removal synthetic microbial communities;

Objective 2: Achieve large-scale application of bioremediation of complex polluted soils rich in difficult-to-degrade pollutants and heavy metals.

4.5.3 Summary

Synthetic biology provides new opportunities for the revolution of environmental biotechnology. However, given the diversity, tenacity, and complexity of application scenarios of environmental pollutants, there are still technical bottlenecks in the application of synthetic biology in the environmental field. On the other hand, attention must also be paid to safety issues such as environmental release. Customizing degradation microbial formulations for characteristics of difficult-to-degrade and toxic pollutants and pollution areas can provide new solutions for biological remediation. Improving the stress resistance of synthetic biological products to actual sites will enhance their effectiveness in actual remediation. By using self-limiting systems or material coupling methods, reducing the escape ability of synthetic biology will ensure biosafety and break through the bottleneck of synthetic biology moving from the laboratory level to the application level.

4.6 Bio-Information Convergence Technology and Extraterrestrial Biology

4.6.1 Abstract

DNA storage and bio-sensing are considered typical intersections of biotechnology and information technology. Today, due to the rapid accumulation of data, storage technology has become a bottleneck. With the rapid development of DNA sequencing and synthesis technology, DNA storage is considered a disruptive new digital information storage method, which can be used to achieve long-term storage of massive amounts of data.

Bio-sensing is an important means for biological data collection and even the construction of Internet of Things (IoT) in health care, and the introduction of synthetic biology will empower bio-sensing with new capabilities. The development of living cell multi-dimensional molecular bio-sensing, rapid ultra-sensitive bio-sensing, wearable intelligent bio-sensing technology, etc., will provide advanced tools for life science research, disease diagnosis and treatment, biological manufacturing process control, on-site environmental pollution monitoring, etc.

In the field of space science, expanding into extraterrestrial space has significant

strategic importance. The extraterrestrial space environment is harsh, and whether life exists has remained uncertain, microorganisms are considered the most potential pioneers for future human expansion into extraterrestrial space. There are extreme conditions like those on Mars and other extraterrestrial planets in the Earth's environment. Relying on these conditions, constructing pioneer microorganisms with extraterrestrial survival and pioneering functions and synthetic microbial communities can help provide a theoretical foundation and reserve technological capabilities for human expansion into extraterrestrial space.

It is expected that in the near future, DNA storage system operation software will be built, and the establishment of industry standards related to DNA storage will be promoted; breakthroughs will be achieved in single-cell metabolic spatiotemporal dynamic imaging and analysis technology, integration and intelligence of biochemical sensor indicators, and wearable technology; at least one extraterrestrial pioneer microbial chassis will be constructed. It is expected that by 2030, a data operating system for DNA storage will be built; breakthroughs will be made in multi-detection, highly adaptable, and highly reliable point-of-care nucleic acid biosensors, achieving point-of-care testing (POCT) requirements; and at least one extraterrestrial synthetic microbial (group) will be constructed, which can use solar energy to synthesize organic matter from CO_2 in the air, with energy utilization efficiency and material transformation rate exceeding natural Earth organisms by 10 to 100 times.

4.6.2 Application Directions

4.6.2.1 DNA Storage

At present, in terms of encoding, theoretical information density of 215 PB/g has been achieved, and the number of DNA molecule copies required for data recovery is 10^4. In terms of functionality, random reading of information, image previewing, similar search functions, etc., have been realized. In terms of application, the construction of automated integrated principle devices for DNA storage has been realized, with reading and writing efficiency of 5 bytes per 21 hours; the largest reported application demonstration storage capacity is about 200 GB. It is expected that by 2030:

Objective 1: Achieve functional expansion of DNA storage and establish the operating system architecture to promote the formulation of industry standards related to DNA storage.

Objective 2: Achieve integrated integration and application demonstration of DNA storage.

4.6.2.2 Bio-sensing

In terms of living cell metabolic sensing, only a very small number of metabolites and metabolic pathways can be analyzed, and research on living cell metabolic phenomics is still in its infancy. In terms of nucleic acid sensing, although POCT instruments have appeared and been commercialized, nucleic acid detection technology that is fast, multi-dimensional, sensitive, and suitable for home use has not yet emerged. In terms of wearable biochemical sensing, continuous monitoring of indicators such as blood sugar and lactic acid has been achieved, but there are still huge challenges in multi-dimensional detection, equipment intelligence and miniaturization. It is expected that by 2030:

Objective 1: Develop multi-detection, highly adaptable, and highly reliable POCT immune/genetic biosensors to achieve pathogen detection POCT requirements.

Objective 2: Establish a cellular metabolic phenotyping analysis system to achieve the localization of genetically encoded biosensors within various subcellular structures.

Objective 3: Achieve real-time online monitoring of biochemical parameters of biological reaction processes.

Objective 4: Develop innovative sensors and sensor arrays for comprehensive health diagnostics.

Objective 5: Integrate bio-sensing with micro-nano robots for *in vivo* disease diagnosis and treatment.

4.6.2.3 Extraterrestrial Biology

With the help of extravehicular exposure devices on near-Earth orbital space stations, extraterrestrial survival experiments have been conducted on some Earth-origin microorganisms, such as bacteria, lichens, algae, bacteriophages, etc., under different time scales and conditions to evaluate their survival capabilities mainly in terms of phenotypic and physiological activity levels. Research on archaea and cyanobacteria synthetic biology is still in its infancy, and lichen synthetic microbiomes have not yet been reported. It is expected that by 2030:

Objective 1: Achieve the design and construction of pioneer extraterrestrial

microbial chassis.

Objective 2: Achieve the simulation application of extraterrestrial synthetic microorganisms.

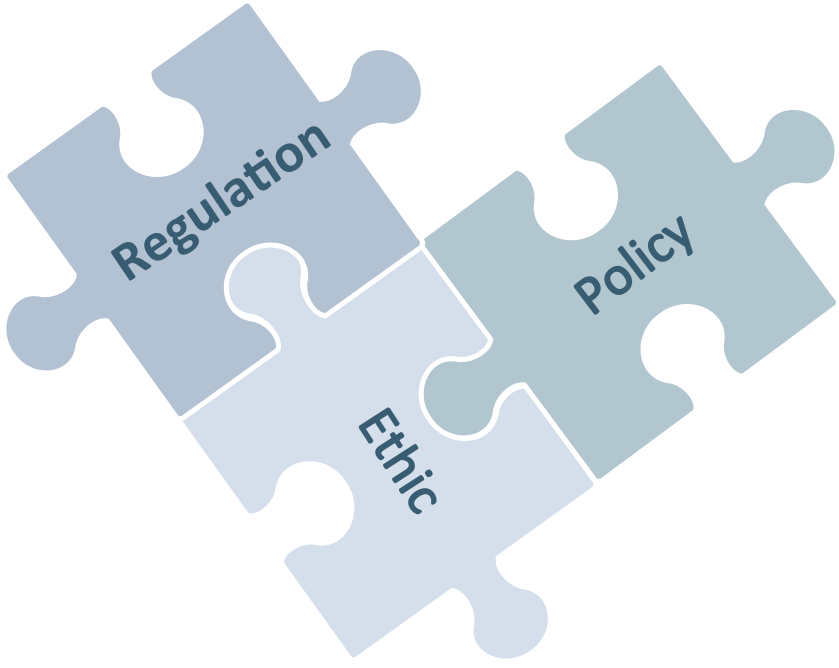
4.6.3 Summary

To achieve DNA storage, focus is needed on the efficiency of bit-base conversion, information density, and accuracy of data writing (synthesis) in terms of information writing; in terms of information reading, focus on error correction, decoding speed, and decoding accuracy; in terms of information preservation, focus on the stability of data recovery in long-term storage and extreme conditions; in terms of information manipulation and control, focus on the functional applications of molecular biology and nanoscience under the DNA storage system. Synthetic biology empowers bio-sensing technology in two aspects: one is the integrated application of multiple bio-sensitive elements to achieve multi-parameter measurement; the other is the intelligent design and modification of bio-sensitive elements to construct bio-sensing systems with good stability, ultra-sensitivity, and adaptability to different application scenarios. The subject of research is a multi-gene regulated spatiotemporal dynamic complex system, suitable for the development of new multifunctional bio-sensing technologies, which need to focus on practical application scenarios such as life omics technologies and disease diagnosis. There are still extreme conditions similar to those on Mars and other extraterrestrial planets in the Earth's environment, so extraterrestrial microorganisms can be constructed starting from existing Earth extremophile microorganisms, systematically building equipment and technological systems, and focusing on exploring new pathways for CO_2 fixation.

5 Guarantee Capability and Governance Principles

The high-quality development of synthetic biology is inseparable from a sound policy and regulatory environment, as well as a governance system that fits the characteristics of the discipline. This chapter will focus on aspects such as ethical considerations, legal regulation, talent cultivation, funding security, academic organizations, international exchanges, and public science popularization of synthetic biology. It also proposes governance principles for the development of synthetic biology based on international scientific consensus, to promote the healthy development of synthetic biology together with international peers and the public.

Guarantee Capability and Governance Principles



Authors Lei Rui-Peng, Wang Guo-Yu, Zhang Xian-En, Liu Chen-Li, Fu Xiong-Fei, Xiong Yan, Du Li, Chen Fang, Lin Zhang-Lin, Li Yu-Juan

5.1 Construction of Guarantee Capability

5.1.1 Ethical Considerations

As a future genetic biotechnology, synthetic biology also follows the conventional ethical considerations applicable to traditional genetic technologies. If the research and development process of synthetic biology involves animal experiments or human trials, ethical reviews must be completed in advance according to procedures. The description of synthetic biology as making "biology more easily engineered" often leads to an increase in the "dual-use" threat, but this often overlooks the professionalism of the discipline and the significant difficulties faced in seeking the design and production of artificial biological systems.

Globally, there is a high emphasis on biological ethics governance, with relatively comprehensive ethical governance systems formed at the international, regional, and national levels. At the international level, the updated Guidelines for Stem Cell Research and Clinical Translation by the International Society for Stem Cell Research in May 2021 provides practical and feasible recommendations for the scientific regulation of stem cell clinical translation. In 2021, the World Health Organization issued the *Human Genome Editing: A Framework for Governance* and *Human Genome Editing: Recommendations*, providing an ethical regulatory framework for human genome editing. China's current legal norms provide macro guidance and standards for ethical governance in the field of life sciences, establishing clear ethical requirements that should not endanger human health, violate ethical morals, or harm public interests. In addition, documents such as the *Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists* (2021), *Opinions on strengthening the governance of science and technology ethics* (2022), and *Measures for Ethical Review of Life Science and Medical Research Involving Humans* (2023) and *Guidelines for the Construction of Ethical Review Committees for Clinical Research Involving Humans* (2023) provide regulatory framework for science and technology ethics governance.

At present, synthetic biology is still in its early stages of development, and before identifying significant new risks, consideration may be given to applying existing biological ethics governance systems. For scientific research, innovation is encouraged; for clinical trials, strict review is implemented. Along with the development of the

discipline, continuous discussion of potential ethical issues is ongoing, and they are actively resolved in the process, balancing development with safety, and innovating with a sense of responsibility.

5.1.2 Legal Regulation

Research on the legal standards and regulation of biotechnology has been underway for some time, and existing laws and regulations have covered multiple areas including laboratory biosafety, pathogenic microorganisms, genetic engineering and transgene, human genetic resources and biological resource protection, ethical management, dual-use items and technology control, etc. It is worth mentioning that the "Criminal Law of the People's Republic of China" (2020 amendment) has added new crimes of illegal gene editing and embryo cloning, clarifying the national regulatory stance on illegal human genome editing and cloning activities. In 2021, the "Biological Safety Law of the People's Republic of China" (hereinafter referred to as the "Biological Safety Law") was officially implemented, establishing the basic regulatory legal framework for China's biological safety risk governance system. Among them, "biological technology research, development, and application" refers to scientific research, technology development, and application activities that recognize, modify, synthesize, and utilize organisms through scientific and engineering principles (Article 2, Paragraph 2). The research, development, and application of synthetic biology fall within the scope of "biological technology research, development, and application" as stipulated by the Biological Safety Law, and related activities are regulated by the Biological Safety Law. Moreover, the Biological Safety Law establishes an important principle that "biological technology research, development, and application activities should comply with ethical principles" (Article 34, Paragraph 2).

Laws, based on coercive power, can be universally adhered to over a considerable period, ensuring the law's high degree of stability to maintain the necessary order. However, the rapid development of technology presents regulatory challenges that far outpace the resolution of issues. Facing the uncertainty of scientific development (unknown innovations), legislation under scientific limitations may lead to excessive legislative costs and more regulatory risks. As synthetic biology is a rapidly developing field with an undefined development pattern, its empowerment boundaries are continuously expanding. Some unpredictable potential risks often emerge in the research

and application process.

At present, the focus of the synthetic biology regulatory system is on the "end product" or "application scenario." Some empowered applications of synthetic biology are subject to current legal restrictions under certain conditions. The emerging technology regulatory system based solely on risk assessment has the defect of being too narrow, and it still needs to take into account social economy and respond to social concerns while ensuring safety, balancing development and safety, and the research and formulation of legal norms need to keep pace with the development of synthetic biology.

Due to the uncertainty in the early stages of synthetic biology development and the actual political context, the formulation of new laws and regulations or significant amendments to existing laws and regulations require significant human and time costs. Emphasizing the transformation from governance to soft governance, which provides a new path for innovative regulation. The development and applications of synthetic biology involve a wide range of stakeholders, including the academic community, industry, social organizations, research institutions, and even the capital sector, media, and the public, constructing a multi-faceted synthetic biology regulatory system that fully leverages the subjective initiative of multiple stakeholders, mobilizes their participation enthusiasm, and carries out innovation with a responsible attitude, enabling this promising technology to develop rapidly while still fully preventing potential risks to human health and the environment.

5.1.3 Talent Cultivation

The development of a discipline is inseparable from the cultivation of talent. As a converging discipline, synthetic biology has received global attention to the cultivation of talents in the field. The International Genetically Engineered Machine competition (iGEM) is the world's highest-level student competition in the field of synthetic biology, initiated by the Massachusetts Institute of Technology, and has developed for more than ten years, cultivating a large number of outstanding talents in the field of synthetic biology. In China, more than 100 teams participate each year, with nearly 2,000 participants, and the teams have won about 1/3 of the gold medals for many years. In 2022, Division of Synthetic Biology, Chinese Society of Biotechnology initiated China's first Synthetic Biology Competition (Synbio Challenges), which got off to a good start and is gradually moving toward internationalization, aiming to provide a platform for

young students to exchange, learn, innovate, and create, cultivating a reserve army for synthetic biology, life sciences, and interdisciplinary studies.

In China, in 2018, the Ministry of Education launched the "Everest Plan for Basic Research in Higher Education Institutions," and Tianjin University was approved to establish a Frontier Science Center for Synthetic Biology. In 2019, the Ministry of Education listed synthetic biology as a self-set interdisciplinary subject, managed according to the second-level discipline. In 2019, Shenzhen University of Advanced Technology was officially approved to be included in the "13th Five-Year Plan" of Guangdong Province, starting to prepare for the establishment of the Synthetic Biology School as one of the six proposed schools. In 2020, the Ministry of Education officially approved the synthetic biology major reported by Tianjin University, and a number of universities have established synthetic biology majors.

As an emerging discipline, the overall talent reserve in the field of synthetic biology is relatively weak. For future development, it is necessary to strengthen the investment in synthetic biology education in universities and research institutes, further promote and optimize the interdisciplinary education model, construct a undergraduate-postgraduate-doctoral research system, and strengthen the cultivation of talents in the fields of "synthetic biology-biotechnology-biomanufacturing-bioeconomy" and related scientific and technological policies, ethical regulations, etc., to provide a high-quality talent reserve for the research, transformation, and industrial development of synthetic biology.

5.1.4 Funding Security

The influence of synthetic biology is increasingly prominent, driving the continuous expansion of the global bioeconomy, becoming a field that major economies such as the United States, the European Union, the United Kingdom, and China focus on planning and investment.

China attaches great importance to the development of synthetic biology. The Ministry of Science and Technology has made systematic deployments from the early layout of the 973 Program and the 863 Program to the focused special projects of the "13th Five-Year Plan" and the strengthened support in the "14th Five-Year Plan", with a total investment of about 5 billion CNY, greatly supporting synthetic biology research in China. In the new era, biomanufacturing and bioeconomy are listed as national strategies.

Several special projects such as synthetic biology, green manufacturing, and bioconvergence (BT and IT integration) have been deployed in the National Key Research and Development Program, and local governments have also set up special projects for synthetic biology and biomanufacturing. Some areas have even established policy-supported industrial parks to cultivate a large number of startups and even listed companies, forming a good situation of interaction between government investment and social capital, promoting the prosperous development of synthetic biology.

5.1.5 Academic Organizations and International Exchange

Academic organizations are important force in promoting academic research and exchanges, playing a pivotal role in promoting scientific and technological cooperation. With the acceleration of globalization, scientific and technological exchanges between countries are becoming increasingly close, and international cooperation is particularly important. Academic organizations provide an important platform for international scientific and technological exchanges and cooperation. In the field of synthetic biology, international, regional, national, and local organizations have formed a global innovation network system, promoting the joint development of the field.

5.1.5.1 Engineering Biology Research Consortium

The Engineering Biology Research Consortium (EBRC) is a non-profit public-private partnership dedicated to building an inclusive community that jointly promotes the development of engineering biology (synthetic biology) to address national and global needs. EBRC focuses on research related to synthetic biology routes, education, safety, and policy and international cooperation. Currently, EBRC has issued a series of roadmaps for the future bioeconomy of synthetic biology and directions such as materials science, microbiology, semiconductors, and defense, providing important references for the global development of synthetic biology.

5.1.5.2 Global Biofoundry Alliance

The Global Biofoundry Alliance (GBA) was officially established in Kobe, Japan, in 2019, jointly initiated by 16 top synthetic biology facilities from eight countries including the United States, the United Kingdom, Singapore, Australia, Canada, Denmark, Japan, and China. GBA aims to promote the construction of synthetic biology facilities globally,

sharing infrastructure, opening standards, sharing best practices, and interoperable data resources to jointly address global scientific challenges such as sustainable development. Currently, GBA is actively exploring and promoting mechanisms for the global sharing of major scientific infrastructure.

5.1.5.3 Asian Synthetic Biology Association

The Asian Synthetic Biology Association (ASBA) was founded in 2018 by academic institutions from China, Japan, the republic of Korea, and Singapore, headquartered in Shenzhen, China. ASBA is committed to carrying out international academic exchanges, improving talent cultivation and research levels, and promoting the development of Asian synthetic biology towards the international first-class level. Based on ASBA, eight countries including China, Japan, the republic of Korea, Singapore, Australia, Thailand, Malaysia, and India initiated the SynCell Asia Initiative in Shenzhen in 2023, focusing on global scientific challenges, exploring innovative institutional mechanisms for talent exchange, scientific cooperation, and other aspects among countries, and promoting the rapid development of Asian single-cell synthesis.

5.1.5.4 Division of Synthetic Biology, Chinese Society of Biotechnology

Division of Synthetic Biology, Chinese Society of Biotechnology, was officially established in 2018 and is affiliated with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, gathering the core strength of domestic synthetic biology, striving to become the "home of science and technology researchers" in synthetic biology, organizing various academic activities, undertaking scientific and technological consultation and popularization, organizing synthetic biology competitions (SynBio Challenges), and has been rated as an outstanding division of Chinese Society of Biotechnology for many times, becoming an active, cohesive, and influential academic organization. To further cultivate young forces and enhance the cohesion of the Division of Synthetic Biology, the Youth Working Group was established in 2021, cultivating outstanding young talents, strengthening the connection coverage of young scientific and technological groups, aiming to truly build a "club" for scientists in the field of synthetic biology.

5.1.5.5 Others

Shanghai, Shenzhen, Hubei, and other places have successively established

academic organizations, industry associations, and alliances related to synthetic biology, providing strong support for academic research and industrial development in this field. The establishment of organizations such as Shanghai Bioengineering Society, Shanghai Synthetic Biology Innovation Strategy Alliance, Shanghai Synthetic Biology Industry Association, Shenzhen Synthetic Biology Association, Shenzhen Synthetic Biology Industry Innovation Alliance, Hubei Synthetic Biology Society has played a significant role in the discipline and industrial development of synthetic biology.

In the next few years, international exchange and cooperation will further deepen, and regional and international non-governmental organizations will play an irreplaceable important role in international scientific and technological exchanges.

5.1.6 Public Science Popularization

Synthetic biology has broad application prospects in industries such as medicine, agriculture, food, environment, and bioinformatics, providing feasible solutions for global challenges such as climate crises, food crises, major disease diagnosis and treatment, and new public health emergencies, gradually becoming an important driving force for promoting bioeconomic and social economic development. However, the "bottom-up" research paradigm of synthetic biology and its grand goal of "creating life" have aroused public concerns. Misconceptions about the development of synthetic biology may invisibly exacerbate public worries about its potential ethical and safety risks. News media shape public understanding of science: on the one hand, hypothetical narratives may increase public concerns and trigger social safety risks; on the other hand, the media's lack of attention to the potential risks of synthetic biology can lead to public doubts. In the process of developing synthetic biology, we need to be vigilant against one-sided public opinion, avoid media hype, strengthen public science popularization based on positive guidance, media dialogue, construct scientific narratives, and promote the sharing of research achievements.

5.2 Governance Principles

Based on the investigation of ethical policies, regulatory policies, talent cultivation, international exchanges, and public science popularization of synthetic biology, the following governance principles for the development of synthetic biology are proposed,

including trustworthiness, solidarity, justice, respect, and dignity, aiming to promote the sustainable high-quality development of synthetic biology.

5.2.1 Trustworthiness

5.2.1.1 Ensure Safety and Controllability

Conduct safety assessments of synthetic biological enabling technologies to ensure the robustness and reliability, and ensure that the research activities involving the design and manufacturing of living organisms are safe and controllable.

Establish a transparent mechanism for the entire life cycle of synthetic biology research and translation applications to achieve explainability, traceability, comprehensibility, feedback capability; in addition to basic and applied research, develop biological containment technologies and safety devices to respond to biological safety and biosecurity.

5.2.1.2 Strengthen Risk Prevention

Enhance the thinking of bottom lines and risk awareness, strengthen the judgment of potential risks in the development of synthetic biology, accumulate knowledge of risk cognition and assessment. Conduct prospective risk assessments of the application research of synthetic biology in the fields of medicine/food, energy/materials, agriculture/environment, based on the principle of risk prevention, analyze and evaluate the uncertainties of constructing new biological entities, creating novel and enhanced functionalities, the novelty and complexity of synthetic materials, the genetic modification of wild populations, and the corresponding environmental and human health impacts.

5.2.1.3 Maintain Dynamic Balancing

Conduct risk-benefit assessments and judgments of synthetic biology innovation and applications, based on core theories and different technological solutions, balance expected benefits and potential risks to ensure an acceptable risk-benefit ratio, fully leverage the significant innovative value of synthetic biology in understanding the nature of life through “build to learn” and creating social economic benefits through “build to use”; based on reflective balancing methods, establish dynamic balancing assessment mechanisms across disciplines, fields, regions, and countries according to basic and applied research progress.

5.2.1.4 Strengthen Responsibility

Implement the value concept of responsible innovation and implement the main responsibilities in synthetic biology research and innovation; strengthen self-discipline awareness and strengthen self-restraint in synthetic biology research and development activities, actively integrate ethical principles/guidelines into each link of scientific research and technology development, strengthen scientific integrity, ethical research, and self-management, establish and improve accountability mechanisms for research and application, and promote the sharing of research achievements while balancing the maximization of benefits and minimization of risks. Actively undertake the responsibility of scientific popularization and dissemination, actively participate in public dialogue, and promote public understanding of science.

5.2.2 Solidarity

5.2.2.1 Enhance Common Welfare

Adhere to people-oriented values, follow common human values, respect fundamental human interests, and the innovation and application of synthetic biology should pursue public good, prioritize meeting major or urgent social public needs, and actively promote the welfare of vulnerable groups. Human welfare includes both the current generation and future generations, requiring people to be in a good state in social and environmental contexts. At the same time, minimize the risks that synthetic biology innovation may bring to people and the environment, maximizing benefits.

5.2.2.2 Achieve Sustainable Development

Synthetic biology should provide alternative technologies to address current societal challenges such as energy, materials, food, environment, health, and data storage, ushering in a new era of bioeconomy, preserving our digital heritage, and addressing unknown epidemic threats; promoting economic, social, and ecological sustainable development, and building a community of shared future for humanity.

5.2.2.3 Protect Biodiversity

Fully consider the potential positive and negative impacts and prudently use

synthetic biology to produce biological entities and products to assist in biodiversity protection, build a community of life on Earth, curb biodiversity loss. Strengthen the prevention of modified living entities and the potential threats of genetic escape, comprehensively consider and assess the impact of scientific and technological innovation on biodiversity.

5.2.2.4 Strengthen International Cooperation

Transform the mindset of zero-sum games, encourage extensive exchanges and cooperation among the international scientific community, explore win-win mechanisms in competition; jointly face the safety and ethical challenges that synthetic biology may bring, share best practices in biological safety, strengthen learning and exchange, establish an international science and technology ethics governance cooperation alliance covering early warning mechanisms and effective regulatory systems.

5.2.3 Justice

5.2.3.1 Promote Fair Accessibility

Adhere to universality and inclusiveness, effectively protect the rights and interests of all stakeholders, promote fair sharing of the research outcomes and expected products of synthetic biology across society, promote social justice and equal opportunities; when providing products and services, fully respect and assist disadvantaged groups and special groups, reduce social structural inequality, narrow the wealth gap.

5.2.3.2 Ensure Benefit Sharing

Utilize the components of biodiversity sustainably, fairly and justly use the benefits generated by genetic resources; through the exchange of biological resource data and information, promote the transfer of synthetic biological products such as modified living entities, knowledge sharing, and synthetic biology research capabilities, covering the fair sharing of research processes, research outcomes, and technical materials, while considering the needs of indigenous people and local communities.

5.2.3.3 Maintain Global Justice

Through technological innovation, narrow the existing global political and economic

order inequalities, on one hand, fully consider the impact of synthetic products and biofuels on agricultural planting in developing countries; on the other hand, make full use of synthetic biology to improve the global accessibility of vaccines and drugs, promoting global health justice.

5.2.4 Respect

5.2.4.1 Uphold Informed Consent

Respect the ethical principles that require respect for human autonomy, effectively safeguard the right to be informed and consent of clinical research participants; under the premise of full respect for research participants, conduct research activities based on the highest ethical standards; the form of informed consent may vary with the intervention or scientific research situation and its associated risks, balanced with the needs of scientific development.

5.2.4.2 Protect Personal Information

Respect the ethical principles that require the protection of personal information and privacy, handle personal information according to legal, proper, necessary, and honest principles, avoid personal privacy and data leaks, do not harm personal legal data rights, establish privacy safety assessment standards and protection plans, implement privacy protection throughout the entire life cycle of data/information collection, transmission, storage, processing, exchange, sharing, and destruction.

5.2.5 Dignity

5.2.5.1 Safeguard Human Dignity

Human dignity is sacred and inviolable, and every individual's existence is inherently valuable. Under the premise that technology is increasingly capable of intervening in human essence, it is necessary to face up to the inherent limitations and fragility of humans, safeguarding dignity and rights based on inherent value. Ensure the humane development of technology and avoid dehumanization trends.

5.2.5.2 Respect Life Value

Synthetic biology both challenges traditional concepts of life and helps understand

the nature of life. Through rational design and directed evolution, modification, and synthesis of existing biological entities, it should acknowledge the intrinsic value of life with sensory capabilities and cautiously distinguish between preserving natural life and artificial life.

5.3 Summary

Like many other high technologies, synthetic biology also has a dual-use nature. On one hand, synthetic biology has tremendous application value and benefits humanity. On the other hand, there are ethical considerations, technological misuse, or misuse risks. Existing regulations on genetic engineering and genetic manipulation can be applied to the potential biological risks of synthetic biology, but given that the enabling applications of synthetic biology are higher than existing genetic operations, it is necessary to establish international consensus to discuss ethics, laws, and regulatory systems from a broader perspective. During the development process of synthetic biology, it is necessary to strengthen public science popularization based on positive guidance, media dialogue, construct scientific narratives, promote the sharing of research achievements; strengthen the construction of synthetic biology disciplines and prepare reserve talents; strengthen international scientific and technological exchanges and cooperation to jointly promote the healthy and rapid development of synthetic biology.

In short, science popularization education, policy ethics, and regulatory formulation need to go hand in hand with the development of synthetic biology, continuously discuss and actively solve potential problems to escort the healthy development of synthetic biology.

Appendix: Glossary of Terms

Adult Stem Cell: Undifferentiated stem cells located in various differentiated tissues, with limited self-renewal and differentiation potential.

Aminoacyl tRNA Synthetase: An enzyme that recognizes specific amino acids and catalyzes the formation of an ester bond between the carboxyl group of the amino acid and the hydroxyl group of the corresponding tRNA. The aminoacylated tRNA then participates in protein synthesis in the ribosome.

Artificial Cofactor: A molecule with a structure similar to natural cofactor, capable of performing electron transfer functions.

Artificial Intelligence (AI): A branch of computer science that aims to understand the nature of intelligence and create machines that can respond in ways similar to human intelligence. AI research includes robotics, speech recognition, image recognition, natural language processing, and expert systems.

Artificial Neural Network (ANN): A mathematical model that mimics the behavior of animal neural networks, enabling complex information processing through intricate internal connections. ANN is a method of machine learning.

Base Editing: A CRISPR-Cas-based technology for targeted base substitution, primarily including cytosine base editors (CBE) and adenine base editors (ABE). CBEs and ABEs use Cas9 mutants fused to cytosine deaminase or evolved adenine deaminase, respectively, to precisely edit target bases, enabling C-to-T (G-to-A) or A-to-G (T-to-C) substitutions.

Biofoundry: A synthetic biology platform that integrates industrial automation and intelligent manufacturing concepts. It uses high-throughput, automated equipment combined with design software and machine learning to rapidly and cost-effectively

complete the "design-build-test-learn" cycle, enabling large-scale, high-throughput, and intelligent synthetic biology research and industrial applications.

Biopart: A functional amino acid or nucleotide sequence, such as regulatory elements (e.g., promoters, terminators, ribosome binding sites) or structural elements (e.g., enzyme genes in natural product synthesis pathways). These are the basic building blocks of organisms and the foundational modules for constructing artificial life forms in synthetic biology.

Bioprinting: An additive manufacturing technology that integrates medicine, engineering, electronics, and biology to precisely arrange bioactive materials and cells in space, creating functional living tissues. Bioprinting uses 3D printing-like technology to combine cells, growth factors, and biomaterials to produce biomedical components that mimic or surpass the features of natural tissues, enabling tissue repair and organ transplantation.

Biosensing: The use of biological materials (e.g., enzymes, proteins, DNA, antibodies, antigens, biomembranes, microorganisms, cells) as recognition elements to convert biochemical reactions into quantifiable physical or chemical signals, enabling the detection and monitoring of biological and chemical substances.

Biosynthesis Technology: A DNA synthesis technology that does not rely on existing DNA templates, instead using DNA terminal transferase and certain DNA polymerases to directly catalyze DNA strand synthesis.

Cell Factory: A metabolically engineered cell system designed for the production of specific substances. These modified cells utilize organic or inorganic substrates as carbon sources, driving energy-dependent activation of gene transcription, translation, and protein expression to produce specific biochemical products.

Cell Therapy: A therapeutic approach that involves isolating normal or functionally specific cells, expanding or processing them *in vitro*, and then transplanting or infusing them into patients. These cells can enhance immunity, kill pathogens and tumor cells, promote tissue regeneration, and aid in recovery. Stem cells or immune cells are commonly used in this therapy.

Cell-Free Protein Synthesis (CFPS): The process of achieving transcription and translation, as described by the central dogma of biology, in an open *in vitro* system.

Chassis Cell: A synthetic biology engineering platform where functionalized elements, circuits, and pathways are integrated into a host cell for rational design.

Chimeric Antigen Receptor T Cell (CAR-T Therapy): CAR-T therapy involves

genetically modifying T cells to express a chimeric antigen receptor that combines a specific antigen recognition domain with T cell activation signals. These modified T cells directly bind to specific antigens on tumor cells, activating them to release perforin, granzyme B, and other substances that kill tumor cells. Additionally, they release cytokines to recruit endogenous immune cells for tumor cell destruction. This therapy can also generate memory T cells, providing long-term, specific anti-tumor immunity.

Coenzyme Regeneration: In this strategic research, coenzyme regeneration refers to: ①The conversion of oxidized nicotinamide coenzymes to reduced forms, such as NAD(P)+ to NADPH; ②The conversion of adenosine diphosphate (ADP) to adenosine triphosphate (ATP).

Column Synthesizer: A low-throughput synthesis instrument that uses porous silica material packed in a column as the reaction carrier.

Covalent Protein Drug: A protein-based drug that forms covalent bonds with target molecules, often exhibiting stronger and more sustained effects.

CRISPR-Cas Gene Editing Technology: This technology uses a designed guide RNA (gRNA) to direct Cas nucleases (e.g., Cas9, Cas12a) or Cas complexes to cut specific genomic targets, causing double-strand breaks. The cell's DNA repair mechanisms, such as non-homologous end joining (NHEJ) or homology-directed repair (HDR), then enable precise cutting, deletion, insertion, or replacement of target sequences.

CRISPR-Cas System: An adaptive immune system in bacteria and archaea that defends against phage and other foreign DNA invasions. It consists of CRISPR (clustered regularly interspaced short palindromic repeats) and Cas (CRISPR-associated genes). When phages invade, foreign DNA is stored in CRISPR sequences. Upon re-invasion, CRISPR is transcribed into mature crRNA, guiding Cas effector proteins to target and eliminate the invading DNA.

Deep Learning: A type of neural network with many interconnected layers, capable of accurately extracting and representing complex biological patterns due to their more complex internal structure, thereby achieving precise prediction and generation tasks.

DNA Assembly: The artificial synthesis of double-stranded DNA sequences *in vitro* or *in vivo*.

DNA Sequencing: The process of determining the order of nucleotides (adenine (A), thymine (T), cytosine (C), and guanine (G)) in a specific DNA fragment.

DNA Synthesis: The artificial synthesis of single-stranded DNA sequences *in vitro*,

typically in the 3'→5' direction.

Electron Donor: A substance that donates electrons in electron transfer reactions and accept oxidation, also known as a hole scavenger.

Embryonic Stem Cell (ESC): Undifferentiated cells derived from the inner cell mass of a human blastocyst. ESCs can proliferate indefinitely and have the potential to differentiate into all cell types of the three germ layers.

Emergence: A phenomenon where the interaction of many small entities produces a larger entity that exhibits properties not present in its individual components.

Extracellular Matrix (ECM): A network of substances secreted by cells into the extracellular space, primarily composed of fibrous components (e.g., collagen and elastin), adhesive proteins (e.g., fibronectin), and filler molecules (e.g., glycosaminoglycans).

Fecal Transplantation of Intestinal Flora: The transfer of functional microbiota from a healthy donor's feces into a patient's gastrointestinal tract to reconstruct the gut microbiota and treat intestinal and extra-intestinal diseases.

Gene Editing (Genome Editing): The targeted modification of specific DNA sequences in a genome using nucleases or complexes with targeting capabilities, enabling cutting, deletion, insertion, or replacement of DNA sequences to alter genetic information and achieve new functions or phenotypes.

Genetic Circuit: A synthetic genetic device designed using regulatory elements and regulated genes based on known mechanisms of gene and protein regulatory networks. It controls the expression of target genes under specific conditions, including the reconstruction and design of genetic circuits.

Genetic Code Expansion: A technology that uses orthogonal aminoacyl-tRNA synthetases and corresponding tRNAs to encode unnatural amino acids via nonsense codons or newly created codons.

Glycosylation: Refers to the process under the catalysis of glycosyltransferase where sugars are transferred to proteins, forming glycosidic bonds with amino acid residues on the proteins. It is an important post-translational modification of proteins and plays a significant role in regulating protein functions.

Gold Nanoparticle (Colloidal Gold): A well-studied nanomaterial with particle sizes typically ranging from 1 to 100 nm. It exhibits strong light absorption, scattering, and photothermal conversion properties, high electron density, dielectric characteristics, and catalytic activity. Gold nanoparticles can bind to various biomacromolecules and have excellent biocompatibility, making them widely used in bioanalysis and

nanomedicine.

High-Throughput Chip Synthesizer: A high-throughput synthesis instrument that uses chips as synthesis carriers.

Immunogenicity: The ability of an antigen to induce an immune response, stimulating specific immune cells to activate, proliferate, and differentiate, ultimately producing immune effectors such as antibodies and sensitized lymphocytes.

Induced Pluripotent Stem Cell (iPSC): A type of stem cell artificially induced through reprogramming of human cells, possessing pluripotent differentiation potential similar to embryonic stem cells.

Knowledge Graph: A visual representation of the relationships between various concepts in a text, typically presented as a network.

Machine Learning: A core component of artificial intelligence that uses computer algorithms to simulate human learning, enabling the acquisition of new knowledge and skills. It typically uses mathematical models to fit patterns in data, allowing accurate predictions about the characteristics of new data or the generation of new data based on these patterns.

Mesenchymal Stem Cell (MSC): A type of adult stem cell with self-renewal and multi-lineage differentiation potential. MSCs are non-terminally differentiated cells with characteristics of mesenchymal, endothelial, and epithelial cells. Under specific induction conditions, they can differentiate into various tissue cells, including fat, cartilage, bone, muscle, nerve, liver, cardiac muscle, and endothelial cells, while retaining multi-lineage differentiation potential after continuous passaging and cryopreservation.

Metabolic Pathway Design: The redesign and engineering of microbial metabolic pathways using their diverse biochemical reaction networks to produce high-value products from low-value or renewable resources. This process focuses on efficient production, balancing metabolic flux and cofactor regeneration, and using gene editing and dynamic regulation to overcome feedback inhibition, creating an optimal synthesis pathway.

Mirror-Image Biology System: A system composed of mirror-image bioparts, which are chiral counterparts of natural biological molecules (e.g., proteins composed entirely of D-amino acids or nucleic acids composed entirely of L-nucleotides).

Multi-Enzyme Catalysis: A cascade reaction involving two or more enzymes, where the product of the first enzyme serves as the substrate or one of the substrates for subsequent enzymes.

Non-Equilibrium State: A steady state where system variables are not constant. A steady state refers to the final state that a system enters as it evolves over time from its initial state. A non-equilibrium state is a steady state other than an equilibrium state. It includes periodic motion (oscillatory state), quasi-periodic state (ergodic state), and chaotic state.

Nonsense Codon: Also known as a stop codon, it includes the amber codon (UAG), opal codon (UGA), and ochre codon (UAA). These codons were named by their discoverers.

One-Pot: A reaction method where all reactants are mixed in a single container in solution form.

Organ Repair: The use of tissue engineering methods to form or regenerate organs for repairing organ defects.

Organoid: A three-dimensional microtissue formed by the proliferation, differentiation, and self-organization of stem cells or organ-specific progenitor cells *in vitro*. Organoids contain multiple cell types and specific cellular arrangements, partially reflecting the physiological structure and function of the source tissue or organ.

Organoid Engineering: The use of engineering strategies to controllably design the cellular composition and self-organization of organoids, as well as to simulate complex tissue microenvironments, enabling the *in vitro* construction of highly reliable 3D organ models.

Organ-on-a-Chip: A miniature cell culture device that simulates the complex microenvironment of living tissues and organs *in vitro*, including multiple cell types, functional tissue interfaces, biochemical factors, and biomechanical forces (e.g., fluid shear stress, tension). It reflects the response of human organs to drugs or environmental stimuli.

Orthogonal Ribosomes: Artificially modified ribosomes that can recognize specific aminoacylated tRNAs without interfering with the endogenous system.

Phase Transition: The transformation of a substance from one state to another due to continuous changes in external parameters.

Phosphoramidite Triester Synthesis: The most widely used chemical method for oligonucleotide synthesis, where oligonucleotide monomers in solution form 3'→5' phosphodiester bonds and attach to a solid-phase carrier.

Plastid Engineering: In addition to traditional nuclear genome engineering, plastid genomes in plants can also be genetically manipulated. Since the first plastid-transgenic

tobacco was created in 1990, successful plastid transformations have been increasingly reported, making plastids a hotspot in plant synthetic biology research.

Polymerase Cycle Assembly Method: A method for assembling target DNA fragments directly from synthetic oligonucleotide primers without the need for additional DNA ligase.

Prime Editing Technology: A novel gene editing technology based on the CRISPR-Cas system. The prime editing system consists of two parts: ① A fusion protein of nCas9 (H840A) and engineered reverse transcriptase; ② A prime editing guide RNA (pegRNA) containing a reverse transcription template. Without requiring double-strand breaks or exogenous DNA templates, the pegRNA guides the nCas9-reverse transcriptase fusion protein to perform single-strand DNA cutting at the target site. The reverse transcriptase template carried by the pegRNA then reverse transcribes the designed DNA fragment, enabling precise insertions or arbitrary base substitutions at the target site.

Protein Prodrug: A precursor drug that undergoes chemical or enzymatic conversion in the body to become an active protein-based drug.

Protein-Drug Conjugate: A protein derivative formed by chemically linking biological or chemical molecules (e.g., small molecules, proteins, DNA, RNA, sugars, lipids) to a target protein, creating a drug candidate.

Pyrosequencing: A novel enzyme-based sequencing technology suitable for analyzing short known sequences. It offers reproducibility and accuracy comparable to Sanger sequencing but with significantly faster speed. Pyrosequencing has the capability to simultaneously sequence and analyze a large number of samples, enables high-throughput, low-cost, rapid, real-time and intuitive analysis of single nucleotide polymorphisms (SNPs) and clinical testing.

Pyrrolysine: A lysine derivative found in methanogenic archaea, known as the 22nd genetically encoded amino acid involved in protein biosynthesis.

Quadruple Codon: Through the modification of bioorthogonal elements, certain regions of mRNA are translated into a single amino acid using four nucleotides, i.e. the “quadruple codon”, which distinct from the “triple codon” in the natural amino acid translation system. It is an expanded genetic coding system that introduces additional base pairs (e.g., X and Y) to increase codon diversity, enabling the encoding of more amino acids or unnatural amino acids. This system can expand the diversity and functionality of protein synthesis.

Quantum Dot: A low-dimensional semiconductor material composed of group IV, II-VI, IV-VI, or III-V elements, typically spherical or quasi-spherical with diameters of 2–20 nm. Quantum confinement effects give quantum dots unique optical properties, including high fluorescence quantum yield, strong photostability, broad excitation spectra, narrow emission spectra, and tunable fluorescence wavelengths. They are widely used in life sciences and semiconductor devices.

Regenerative Medicine: An interdisciplinary field that applies principles and methods from life sciences, materials science, clinical medicine, computer science, and engineering to develop theories and technologies for replacing, repairing, reconstructing, or regenerating human tissues and organs.

Sample Loading Density of High-Throughput Sequencing: The density of DNA clusters or nanoballs per unit area on a high-throughput sequencing chip.

Sanger Sequencing (Chain Termination Method): A sequencing method invented by Sanger et al. in 1977, based on the incorporation of chain-terminating dideoxynucleotides (ddNTPs) during DNA synthesis by DNA polymerase. Each sequence determination consists of a set of four separate reactions, each sequencing reaction includes all four dNTPs and a limited amount of one ddNTP.

Self-Organization: A process in which a system spontaneously forms an ordered structure without external instructions, relying on mutually agreed-upon rules and coordination among its components.

Semi-Synthetic Organism: A life form created by introducing artificial, non-natural bases into genomic DNA, enabling the replication and transcription of genetic information. Currently, *Escherichia coli* is commonly used for this purpose.

Sequencing by Ligase chain reaction: A DNA sequencing method that uses DNA hybridization and ligation reactions to determine DNA sequences.

Sequencing by Synthesis: The most widely used high-throughput sequencing technology, which involves sequencing DNA by extending bases with DNA polymerase.

Sequencing Read: The length of sequence obtained in a sequencing reaction. In high-throughput sequencing, it is limited by the asynchronous extension of multiple copies of molecules, while in single-molecule sequencing, it is limited by the length of the extracted DNA fragment.

Single Molecule Sequencing: Sequencing individual DNA molecules in real-time without PCR amplification, primarily through single-molecule fluorescence sequencing or nanopore sequencing.

Spatial Omics: The analysis of gene expression patterns in individual cells and the spatial relationships and biological characteristics of cell populations across time and space.

Spatial Transcriptome Resolution: The minimum distance between capture probe clusters on the surface of a carrier.

Stem Cell: Cells with varying differentiation potential and the ability to self-renew in an undifferentiated state.

Synthetic Efficiency: The efficiency of coupling one base to the next in an oligonucleotide during a single synthesis cycle.

Synthetic Error Rate: A measure of synthesis fidelity, calculated as the ratio of incorrect bases to the total number of synthesized bases.

Synthetic Flux: The maximum number of oligonucleotide species that can be synthesized in a single run.

Synthetic Hybrid Biotic-abiotic System: A system or component composed of biological (e.g., nucleic acids, proteins, viruses) and non-biological (e.g., inorganic nanomaterials) elements. These components can be ordered through molecular assembly or biomineralization, leading to enhanced properties or emergent functions. Such systems include non-cellular hybrid systems and hybrid cellular systems.

Synthetic Hybrid Cellular Biosystem: An engineered cell system that carries or contains non-bioparts (e.g., inorganic nanomaterials) as functional units. These systems can be constructed by loading chemically synthesized non-bioparts into living cells or directly synthesizing non-bioparts within cells.

Synthetic Hybrid Non-Cellular Biosystem: Functional structures formed by the ordered combination of bioparts (e.g., proteins, nucleic acids, viruses) and non-bioparts (e.g., inorganic nanomaterials) through chemical crosslinking, adsorption, mineralization, or assembly.

Synthetic Yield: The total amount of oligonucleotide product from a single synthesis unit.

Transcription Activator-like Effector Nuclease (TALEN) Technology: A gene editing technology that uses transcription activator-like effector nucleases to target and edit specific DNA sequences. TALENs consist of a programmable transcription activator-like effector (TALE) protein for DNA recognition and the nuclease domain of FokI for DNA cleavage.

Universal Cell Therapy: A form of therapy that involves genetically modifying

human leukocyte antigen (HLA) or T cell receptor (TCR) molecules to avoid immune rejection in allogeneic cell therapy. This allows for the standardization, scalability, and universalization of therapeutic cells, enabling their characterization, quality control, and application like pharmaceutical drugs.

Universality Class: In statistical physics, critical exponents are used to quantitatively describe the critical behavior of physical quantities near phase transition points. Phase transitions with the same critical exponents belong to the same universality class.

Unnatural Amino Acid: Amino acids other than the 20 naturally encoded ones. Selenocysteine and pyrrolysine are also naturally encoded in specific systems, making them the 21st and 22nd natural amino acids. All other amino acids are considered unnatural.

Unnatural Base: Artificially designed and synthesized bases introduced into biological systems, including derivatives of the four natural bases (A, T, C, G) and novel bases (e.g., X, Y) that can pair complementarily and be used in transcription and translation.

Unnatural Protein: Proteins containing amino acids other than the 20 standard natural amino acids, including synthetic unnatural amino acids.

Vascularization: The formation of blood vessel-like structures in three-dimensional tissues constructed *in vitro*. Vascularization addresses the issue of oxygen and nutrient deficiency in organoids and is a key focus in current organoid research.

Virus Vector: A virus modified through genetic engineering to serve as a vehicle for delivering foreign genes into cells. Virus vectors efficiently transduce cells and enable long-term gene expression. Tool virus vectors are characterized by high transduction efficiency and high levels of foreign gene expression.

Zinc-Finger Nuclease (ZFN) Technology: A gene editing technology that uses zinc-finger nucleases to target and edit specific DNA sequences. ZFN consists of a programmable zinc-finger protein for DNA recognition and the nuclease domain of FokI for DNA cleavage.