

## Enabling technology and core theory of synthetic biology

Xian-En Zhang<sup>1,3\*</sup>, Chenli Liu<sup>1,2\*</sup>, Junbiao Dai<sup>1,2\*</sup>, Yingjin Yuan<sup>4\*</sup>, Caixia Gao<sup>5\*</sup>, Yan Feng<sup>6\*</sup>,  
Bian Wu<sup>7\*</sup>, Ping Wei<sup>1,2\*</sup>, Chun You<sup>8\*</sup>, Xiaowo Wang<sup>9\*</sup> & Tong Si<sup>1,2\*</sup>

<sup>1</sup>Faculty of Synthetic Biology, Shenzhen Institute of Advanced Technology, Shenzhen 518055, China;

<sup>2</sup>Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;

<sup>3</sup>National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;

<sup>4</sup>Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China;

<sup>5</sup>State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China;

<sup>6</sup>State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai 200240, China;

<sup>7</sup>State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China;

<sup>8</sup>Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China;

<sup>9</sup>Ministry of Education Key Laboratory of Bioinformatics; Center for Synthetic and Systems Biology; Bioinformatics Division, Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China

Received August 8, 2022; accepted October 4, 2022; published online February 6, 2023

Synthetic biology provides a new paradigm for life science research (“build to learn”) and opens the future journey of biotechnology (“build to use”). Here, we discuss advances of various principles and technologies in the mainstream of the enabling technology of synthetic biology, including synthesis and assembly of a genome, DNA storage, gene editing, molecular evolution and *de novo* design of function proteins, cell and gene circuit engineering, cell-free synthetic biology, artificial intelligence (AI)-aided synthetic biology, as well as biofoundries. We also introduce the concept of quantitative synthetic biology, which is guiding synthetic biology towards increased accuracy and predictability or the real rational design. We conclude that synthetic biology will establish its disciplinary system with the iterative development of enabling technologies and the maturity of the core theory.

**synthetic biology, quantitative synthetic biology, genome synthesis and assembly, DNA storage, molecular evolution, *de novo* design, computer-aided design, cell engineering, gene circuit, chassis cell, artificial intelligent (AI), biofoundry**

**Citation:** Zhang, X.E., Liu, C., Dai, J., Yuan, Y., Gao, C., Feng, Y., Wu, B., Wei, P., You, C., Wang, X., and Si, T. (2023). Enabling technology and core theory of synthetic biology. *Sci China Life Sci* 66, 1742–1785. <https://doi.org/10.1007/s11427-022-2214-2>

### Introduction

Synthetic biology, also referred to as engineering biology, is a newly emerging interdisciplinary subject. It integrates

biological sciences, chemistry, physics, material science, computer and information science, as well as engineering concepts to redesign or *de novo* design and construct biological systems, creating a new paradigm of biological research known as “Build to learn” and empowering current biotechnology a new driving force, called “Build to use” (Deng, 2019; Elowitz and Lim, 2010; Liu et al., 2013; Zhang, 2018; Zhang, 2019; Zhao, 2018). Synthetic biology consists of three interrelated aspects: theory, enabling technology,

\*Corresponding authors (Xian-En Zhang, email: [zhangxe@ibp.ac.cn](mailto:zhangxe@ibp.ac.cn); Chenli Liu, email: [chenli.liu@siat.ac.cn](mailto:chenli.liu@siat.ac.cn); Junbiao Dai, email: [junbiao.dai@siat.ac.cn](mailto:junbiao.dai@siat.ac.cn); Yingjin Yuan, email: [yjyuan@tju.edu.cn](mailto:yjyuan@tju.edu.cn); Caixia Gao, email: [cxgao@genetics.ac.cn](mailto:cxgao@genetics.ac.cn); Yan Feng, email: [yfeng2009@sjtu.edu.cn](mailto:yfeng2009@sjtu.edu.cn); Bian Wu, email: [wub@im.ac.cn](mailto:wub@im.ac.cn); Ping Wei, email: [ping.wei@siat.ac.cn](mailto:ping.wei@siat.ac.cn); Chun You, email: [you\\_c@tib.cas.cn](mailto:you_c@tib.cas.cn); Xiaowo Wang, email: [xwwang@tsinghua.edu.cn](mailto:xwwang@tsinghua.edu.cn); Tong Si, email: [tong.si@siat.ac.cn](mailto:tong.si@siat.ac.cn))

and applied research.

More than half a century ago, after understanding the primary structure of nucleic acids and proteins, scientists achieved a historic leap from the chemical synthesis of small molecules to the chemical synthesis of biological macromolecules. Pioneering studies included the total synthesis of crystalline bovine insulin (Kung et al., 1965), polynucleotide synthesis of the genetic code (Khorana et al., 1966), and amino acid transferase RNAs (Lapidot et al., 1969; Wang et al., 1983; Weber and Khorana, 1972). At the turn of the century, with the completion of the Human Genome Project and the advancement of DNA sequencing and synthesis technologies, synthetic biology achieved a new leap from nucleic acid synthesis to genome synthesis. From simple to complex, scientists have synthesized viral genomes (Cello et al., 2002), bacterial genomes (Gibson et al., 2010a; Ostrov et al., 2016), and yeast chromosomes (Annaluru et al., 2014; Dymond et al., 2011; Shen et al., 2017; Xie et al., 2017; Zhang et al., 2017a). These landmark approaches provided the foundation for synthesis of biological systems by demonstrating that synthetic genomes can fully perform natural biological functions.

During the same period, the concepts of engineering were introduced into the design and creation of biological systems, such as gene circuits, biological devices and modules, minimal genomes and chassis cells. These concepts were well explained by many artificial logical biological devices and genomes, for example, gene toggle switches (Gardner et al., 2000), synthetic oscillatory networks of transcriptional regulators (Elowitz and Leibler, 2000), quorum sensing based intercellular communication circuits (Miller and Bassler, 2001), synthetic multicellular systems for programmed pattern formation (Basu et al., 2005), RNA devices for performing logical operations (Win and Smolke, 2008), and programmable microbial kill switches (Callura et al., 2010). These pioneering explorations seek to achieve the standardization of biological components, the generalization of chassis cells, and the predictability of designing biological systems based on engineering principles.

However, biological systems are extremely complex systems due to their sustainable genetic variation, metabolic diversity and dynamics, and biomass flexibility. The knowledge we have acquired for biological design is far from sufficient to meet engineering standards. The high-throughput, multi-cycle and automated “trial and error” biofoundries came into being, and its efficiency is being exceedingly advanced by the embedding of artificial intelligence (AI). In particular, the advanced algorithms represented by DeepMind’s AlphaFold (Jumper et al., 2021; Senior et al., 2020) and the Baker lab’s RoseTTA fold (Anishchenko et al., 2021a; Anishchenko et al., 2021b) have revolutionized the prediction of protein 3D structures. This means that AI-assisted protein *de novo* design will flourish.

As a result, we can distinguish two *modus operandi*: a data-driven “black box” and a knowledge-driven “white box”. The vast amount of knowledge input and data output enables us to perform computer modeling that enables accuracy and predictability in biological design. Such modeling is termed *quantitative synthetic biology* which is becoming the core part of the theory of synthetic biology. Its validity and reliability have been demonstrated in investigations on the ability of bacteria to colonize new available habitats (Liu et al., 2019b) and elaborating quantitative relationship between growth and cell cycle of *Escherichia coli* (Zheng et al., 2020).

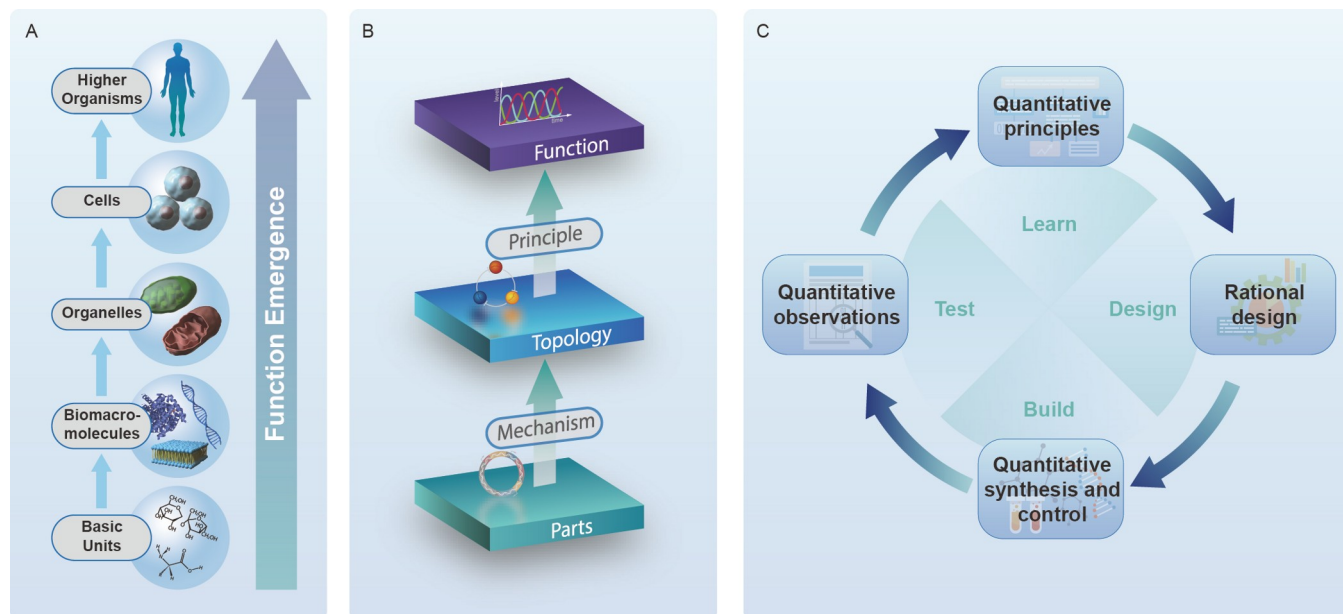
During 2009 and 2012, the National Science and Engineering Academies from China, USA, and UK held a series of joint symposiums on synthetic biology ([http://www.nap.edu/catalog.php?record\\_id=13316](http://www.nap.edu/catalog.php?record_id=13316)). The major topic of the event ocured in Shanghai was “Enabling technology of synthetic biology” ([http://www.sippe.ac.cn/yjdy/hcswx/hczxxx/201812/t20181214\\_5212243.html](http://www.sippe.ac.cn/yjdy/hcswx/hczxxx/201812/t20181214_5212243.html)). Enabling technology of synthetic biology is a collection of a series of novel or iterative technologies, such as genome sytheis and assembling, gene editing and genome reprograming, evolution and *de novo* design of proteins, chassis cells, computer-aded design and modeling of new biomacromolecules and biosystems, DNA information storage, production automation of biological systems, genetic code expanding and semisynthetic organisms, etc. With the creating and development of the enabling technologies, synthetic biology will widely realize its goals, establish a new paradigm for biological science research, and have a revolutionary impact on biotechnology and its applications.

On December 24, 2021, we successfully held the Forum on “Enabling technologies and core theory of synthetic biology”. Here, we summarize the discussion held at this Forum. However, due to the large number of topics and limited space, this paper does not cover some critical topics, such as orthogenetic systems based on genetic code expansion, as well as bioethics, biosafety and biosecurity, which we hope to discuss at other opportunities. See the link for: <https://blog.sciencenet.cn/video.php?mod=vinfo&pid=1530>.

## Quantitative synthetic biology

### *Hierarchy of synthetic biology research*

Living systems are considerably complex, as a result of their hierarchical organization and layers of interaction (Channon et al., 2008) (Figure 1A). Properties displayed by complex systems are often referred to as “emerging properties”, which are the result of interactions between components that cannot be predicted, even with a full knowledge of all parts (Aderem, 2005). For example, identifying the genes and proteins in an organism is analogous to listing all parts of an



**Figure 1** Understanding function emergence with quantitative synthetic biology. A, Function emergence in living systems across their hierarchical organization, which cannot be understood by knowing individual parts alone. B, Hierarchy of the methodology. C, Principle and rational design in quantitative synthetic biological research promote each other. The proposed research paradigm of quantitative synthetic biology integrates rational design, building, and testing.

airplane, which is not enough to understand how an airplane flies. Therefore, understanding how the properties of life emerge in hierarchical organization remains one of the most fundamental questions in life sciences.

Using biological and engineering principles, synthetic biology aims to design biological modules/reactions/systems to achieve desired functions/products (Kai and Schwille, 2019). From biomolecular engineering, genome engineering to artificial cell design, synthetic biologists can access hierarchical biological systems at any level. At every level, synthetic biologists aim to create a more synthetic and complex entry, thus advances in synthetic biology often results in the most complex and unnatural systems (Channon et al., 2008). In general, synthetic biology studies how properties of life emerge in hierarchical living systems using a bottom-up strategy which we “understand by construction”. As Richard P. Feynman’s famously stated “What we cannot create, we cannot understand.”

Synthetic biology has proven to be a valuable tool for exploring all three dimensions of the properties of life, namely, functions that previously existed, functions that exist, and functions that do not yet exist. It allows the creation of artificial systems, such as silicon-based life (Kan et al., 2016) and single-chromosomal yeast (Shao et al., 2018), to explore the boundaries of life and reveal the fundamental question of “what is life”. In contrast, these artificial systems can surpass the capabilities of natural cells, which will considerably facilitate the development of biotechnology to transform our daily lives (Voigt, 2020), i.e., to explore non-existent life functions. By engineering/rebuilding biological

systems, synthetic biology enables us to extract inaccessible information from conventional biology and helps us understand the principles of life (existing functions) (Elowitz and Leibler, 2000; Liu et al., 2019b). Finally, synthetic biologists study complex biological systems through simplification and modeling. These efforts have advanced our understanding of its underlying physics, which will help elucidate possible evolutionary routes (existing functions) in early Earth environments (Budin and Szostak, 2011).

Despite these considerable advances in understanding the life system, a well-established synthetic biology approach remains missing. To address this issue, we first summarize the general paradigm of previous systems. An integrative biology research practice usually consists of three levels, termed partial, topological, and functional. Functionality comes from parts, but with complex systems it is difficult to obtain an end-to-end relationship between them. Therefore, a topological intermediate layer is added. How topology emerges from the part is called the mechanism, and how the function emerges from the topology is called the principle (Figure 1B). The principle is more general, while the mechanism is specific to the part used. For example, Alan Turing proposed the reaction-diffusion principle, which broadly regulates pattern formation in nature (Turing, 1952). Based on this general principle, researchers constructed chemical Belousov-Zhabotinsky reactions (Yoshida, 2010), biological circadian oscillations (Nakajima et al., 2005), and engineered DNA/enzyme reaction networks (Senoussi et al., 2021) of artificial reaction-diffusion systems to follow their respective mechanisms. Therefore, understanding these

principles lays the foundation for rational design, which in turn accelerates synthetic biology towards increased complexity and efficiency, advancing our understanding of the fundamental principles of living systems (Figure 1C). An ideal synthetic biology practice, which we call rational design, is to design basic components based on the principle of interest to achieve the intended function.

### ***Quantitative synthetic biology: a research paradigm to investigate function emergence***

Successful rational design is limited to a few well-characterized systems, such as toggle switches and oscillators in earlier studies (Elowitz and Leibler, 2000; Gardner et al., 2000). Recent examples include morphogen-mediated artificial cell differentiation (Tian et al., 2019), and artificial photosynthetic systems that enable photodynamic carbon dioxide fixation (Miller et al., 2020a). However, the principles governing biological functions of interest are often elusive. In this case, researchers can devise novel principles based on quantitative analysis of the interest function and then test them; or they can rely on well-defined components and fine-tune to explore their potential functions. Thus, when the rationale for a function is unknown, routine work in synthetic biology research involves tedious trial and error and luck in fine-tuning. While this trial-and-error research paradigm has been successful in generating critical information and expanding our understanding of complex biological systems, the rational design of synthetic biology will be in high demand in the coming decades to effectively explore the fundamental principles of living systems.

How can we achieve rational design without the principle of interest? To design a system rationally is to design for predictability, that is, the ability to predict outcomes based on input components and parameters. This requires us to quantify natural phenomena, freeing us from the ambiguity and subjectivity of qualitative descriptions, allowing us to develop theories and make predictions. There are two models that guide us to quantify natural systems: knowledge-driven “white-box” models and data-driven “black-box” models. A white-box model is established based on macroscopic experimental observations. Through synthetic analysis, we can describe these complex observations in mathematical formulations, thereby extracting general theoretical frameworks and underlying principles. For example, Liu et al. developed a reaction-diffusion model consisting of multiple partial differential equations describing cell growth, cell movement, etc., in a range expansion system of bacterial colony, reproducing the spatial patterns that spontaneously emerge in this system (Liu et al., 2011); Zheng et al. discovered a linear relationship between the cell mass and the rate of chromosome replication-segregation, which provided a quantitative basis for understanding cell cycle regulation and program-

ing cell size (Zheng et al., 2020). Terence Hwa’s group formulated a number of bacterial growth laws and established a principle of proteomic resource allocation (Erickson et al., 2017; Scott and Hwa, 2011), providing a predictive model for understanding the response of bacteria to physiological perturbations and design synthetic metabolic pathways. In contrast, black-box modeling focuses on the direct correlation between input and output. A large amount of known input and output data will be used to train and improve the algorithm, which can then be used to predict the outcome of the associated system. For example, AlphaFold, developed by DeepMind, successfully predicted 98.5% of human protein structures, which were trained with known amino acid sequence-protein-structure relationships (Jumper et al., 2021).

Both white-box and black-box modeling have been demonstrated to be valuable tools in biodesign: By developing mechanistic models and systematically analyzing the models, Chau et al. identified network topologies that can achieve spontaneous cell polarization. Based on the models’ predictions, they successfully built synthetic gene circuits that generated polarized protein distribution in yeast (Chau et al., 2012). Lu et al. developed a machine learning algorithm to predict how PET (a major type of plastics) hydrolases could be mutated to improve their efficiency and robustness. Guided by this algorithm, the team engineered the wild-type enzyme and obtained a mutant with much superior PET-degrading activity (Lu et al., 2022). Although taking different routes, both approaches aim to construct natural phenomena through quantitative relationships, enabling predictive design. Therefore, we propose the concept of quantitative synthetic biology as an urgent research paradigm facing the current bottleneck of rational design.

Quantitative synthetic biology is the intersection of quantitative biology and synthetic biology. It studies synthetic biology systems from the bottom up and uses simplified quantitative relationships to describe complex biological phenomena. Guided by both white-box and black-box modeling, a quantitative understanding of living systems can be obtained, from which we can develop theories of complex biological systems and explore their fundamental principles. An understanding of the underlying principles will facilitate rational design, thereby accelerating the realization of true synthetic biology engineering. Both white-box and black-box modeling deal with large amounts of data, which can be achieved by building automated and high-throughput experimental facilities and standardized protocols, algorithms, and workflows (tests). Finally, enabling technologies should be further developed to precisely control/rebuild biological systems, such as efficient and precise DNA synthesis techniques, genome editing, gene circuit design, and protein directed evolution (construction). Therefore, we propose a future research paradigm for synthetic biology, including



design, construction, and testing (Figure 1C). We envision that this research cycle will revolutionize the current qualitative, descriptive, and limited synthetic biology research into a new phase with quantitative, theoretical, and systematic features. This revolution will push the frontiers of biology by answering fundamental questions about how life functions, which in turn will help us design synthetic systems with improved predictive power.

## Synthesis and assembly of a genome

### Brief history of genome synthesis

The history of synthetic genomics dates back to 1970, following a 5-year effort to synthesize the 77 bp yeast tRNA gene (Agarwal et al., 1970). In 2002, a 7.5 kb poliovirus complementary DNA was chemically constructed (Cello et al., 2002). One year later, the 5.5 kb genome of phage X174 was created in just two weeks (Smith et al., 2003). Encouraged by this success, a number of groups set out to construct the 583 kb *Mycoplasma genitalium* JCVI-1.0 genome, which was achieved in 2008 (Gibson et al., 2008a). Subsequently, the 1.1-Mbp *Mycoplasma mycoides* JCVI-syn1.0 genome was synthesized and demonstrated to be functional (Gibson et al., 2010b). To date, synthetic genomes have mostly mimicked natural template DNA. In 2016, scientists minimized the 1.1 Mbp JCVI-syn1.0 genome to a functional 531 kbp JCVI-syn3.0, using three design-build test cycles (Hutchison III et al., 2016).

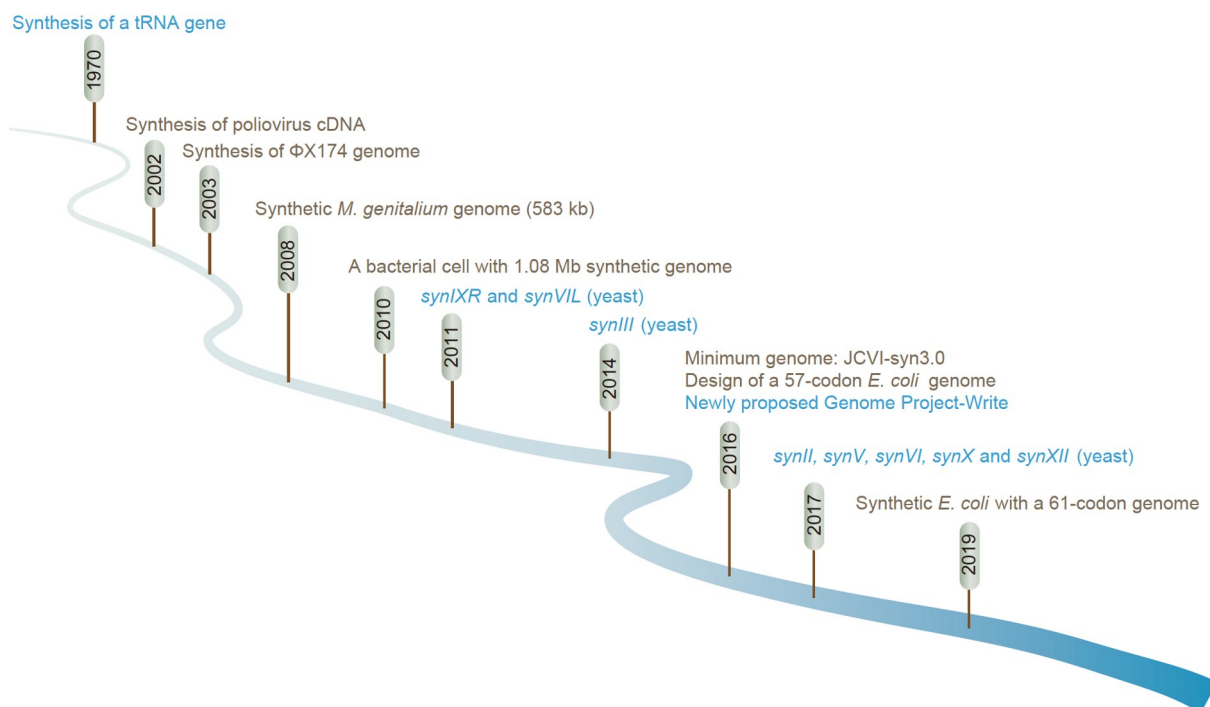
In addition to assessing the plasticity of gene content, genome synthesis allows for reprogramming of the genetic code. In 2016, a 3.97-megabase, 57-codon *Escherichia coli* genome was designed and experimentally validated on 63% of the synthetic genome (Ostrov et al., 2016). Three years later, an *E. coli* variant with a synthetic genome of 61 codons was created, enabling the first compression of the sense codons of the entire genome (Fredens et al., 2019) (Figure 2).

When synthesizing viral and bacterial genomes, an attempt to synthesize eukaryotic genomes, called the Synthetic Yeast Genome Project (Sc2.0), was initiated, an effort that is only now nearing completion (Jiang et al., 2020a; Luo et al., 2020; Luo et al., 2018; Zhang et al., 2020b). Sc2.0 aims to synthesize the entire genome of *Saccharomyces cerevisiae* (~12 Mb, divided into 16 chromosomes), with numerous changes to explore fundamental biological questions about genome function. In 2016, a more ambitious project, the Genome Project Write (GPW), was proposed to rewrite complex genomes of gigabase size (Boeke et al., 2016). However, the current capacity and the cost of DNA synthesis are the major limiting factors in the construction of such large genomes, thus there is an urgent need to make breakthroughs in DNA synthesis and genome assembly.

### Technology development for gene synthesis and genome assembly

#### Oligonucleotide synthesis

Currently, the most commonly used technique for oligonu-



**Figure 2** Milestones in synthetic genomics. Brown represents progress in viral and bacterial genome composition. Blue indicates milestones in eukaryotic genome composition.

cleotide synthesis is the solid-phase phosphoramidite chemical synthesis method developed in the 1980s (Beaucage and Caruthers, 1981). In this method, the addition of each nucleotide monomer proceeds through a cycle of four steps: deprotection, coupling, capping, and oxidation (Hughes and Ellington, 2017; Hughes et al., 2011). The cycle is then repeated for the next base by removing the protecting group. The robustness and fidelity of this approach enables it to be automated and industrialized. Since the 1990s, DNA synthesizers based on this method have been developed to synthesize 96–1536 different oligonucleotides simultaneously (Cheng, 2002; Rayner et al., 1998). In comparison, arrayed base oligonucleotide synthesis technology could, in theory, significantly reduce costs and increase yields (Hughes and Ellington, 2017). The authors, however, pointed out that synthesis quality typically decreases with increasing oligonucleotide length due to inevitable synthesis-related errors in stepwise multiple reaction systems. Despite continuous efforts to optimize the synthesis process, synthetic oligonucleotides are typically no longer than 200 nt in length (Hughes and Ellington, 2017; Kosuri and Church, 2014).

Enzymatic *de novo* synthesis of oligonucleotides, proposed as early as the 1960s (Bollum, 1962), has emerged as a promising alternative due to the length limitations and hazardous waste of chemical synthesis. Currently, terminal deoxynucleotidyl transferase (TdT) is the best option (Barthel et al., 2020; Eisenstein, 2020; Lee et al., 2020; Lee et al., 2019; Palluk et al., 2018). After years of efforts, it has been reported that enzymatic synthesis can generate about 300 mers, outperforming chemical synthesis (Eisenstein, 2020). To date, several companies have been established to advance the commercialization of enzymatic DNA synthesis. Achieving rapid, high-throughput on-demand synthesis of long DNA molecules in the future will considerably accelerate the design-build-test cycle in systems biology.

### Gene synthesis

The term “gene” in gene synthesis refers to a long double-stranded DNA sequence rather than the classical definition of a gene (Kosuri and Church, 2014). Commercially synthesized genes are usually between 200 and 3000 bp in length. Single-stranded oligonucleotides with complementary overlapping sequences are the raw materials for the assembly of this double-stranded synthetic DNA. Earlier methods were ligation-based in which adjacent oligonucleotides were enzymatically ligated by DNA ligase (Agarwal et al., 1970; Au et al., 1998; Bang and Church, 2008; Sekiya et al., 1979). Since the invention of the polymerase chain reaction (PCR) in the 1980s, PCR-mediated methods have been widely used to assemble desired DNA sequences from oligonucleotides (Hughes and Ellington, 2017; Hughes et al., 2011; Stemmer et al., 1995). In addition, Gibson and colleagues developed *in vitro* (Dormitzer et al., 2013; Gibson

et al., 2010b) and *in vivo* (Gibson, 2009a) one-step methods for the direct assembly and cloning of oligomers into plasmids. Currently, the above methods have been iteratively improved and used in most commercial and academic applications (Kosuri and Church, 2014). Furthermore, due to the need for inexpensive synthetic DNA, methods for gene synthesis using microarray-based oligonucleotide pools have also been developed (Borovkov et al., 2010; Kosuri et al., 2010; Quan et al., 2011; Tian et al., 2004).

In addition to genes, various applications require longer DNA molecules over 10 kb or even 100 kb in length, which has led to the development of a range of methods to assemble short DNA, such as BioBrick (Shetty et al., 2008), BglBrick (Anderson et al., 2010), iBrick (Liu et al., 2014), and HVAS (Li et al., 2014). However, the “scar” sequences produced by these endonuclease-based techniques may affect the function of the final constructs or introduce undesired variations. Type IIS restriction enzymes are characterized by a cleavage site that is only a few bases away from the recognition site, making them an ideal solution for “scar-free” assembly. Based on this principle, the Golden Gate method and toolkit were developed and gained significant popularity (Cermak et al., 2011; Engler et al., 2008; Guo et al., 2015). Furthermore, to obviate the need for restriction enzymes, several seamless assembly methods such as Gibson assembly (Gibson et al., 2009b), ligase cycling reactions (de Kok et al., 2014), sequence and ligation independent cloning (Li and Elledge, 2007), circular polymerase extension cloning (Quan and Tian, 2009; Quan and Tian, 2011) and yeast assembly (Gibson et al., 2008b; Jiang et al., 2022; Shao et al., 2009), have been established. Currently, which assembly technique to use is a matter of preference. Importantly, most of the above methods can be automated to increase the throughput of constructing long synthetic DNA (Hughes and Ellington, 2017).

### Genome assembly

To synthesize small genomes, restriction cloning or polymerase circular assembly (PCA) methods are usually sufficient (Cello et al., 2002; Smith et al., 2003), while combinations of different tools are employed to construct larger scale (more than 100 kilobases) synthetic chromosomes or genomes (Annaluru et al., 2014; Fredens et al., 2019; Gibson et al., 2008a; Gibson et al., 2010a; Hutchison III et al., 2016; Richardson et al., 2017; Zhang et al., 2020b; Zhang et al., 2017a).

Although Gibson assembly has been reported to assemble DNA molecules up to several hundred kilobases (Gibson et al., 2009b), the efficiency of *in vitro* procedures declined with the increase of DNA length, making it mainly a commercialized tool in constructing synthetic DNA of tens of kilobases (Gibson et al., 2008a; Gibson et al., 2010b; Zhang et al., 2020b). By comparison, the upper limit of yeast assembly appears to be much higher. Beyond the high effi-

ciency of DNA assembly within or around 100 kb (Fredens et al., 2019; Gibson et al., 2008a; Jiang et al., 2022), one-pot yeast transformation has generated several synthetic genomes of hundreds of kilobases or even over 1 Mb, such as the two *Phaeodactylum tricornutum* chromosomes (497 kb and 441 kb) from 5 fragments (Karas et al., 2013), the 583 kb bacterial genome generated with 25–25 kb fragments (Gibson et al., 2008b), the 786 kb *Caulobacter ethensis*-2.0 using 16 mega-segments (38–65 kb in size) (Venetz et al., 2019) and the 1.08 Mb JCVI-syn1.0 genome via 11 pieces of 100 kb overlapping intermediates (Gibson et al., 2010a). Yeast homologous recombination is also critical for the assembly of the synthetic chromosomes in Sc2.0, which share high similarity to its own genome (Annaluru et al., 2014; Luo et al., 2018; Richardson et al., 2017; Zhang et al., 2020b; Zhang et al., 2017a). Combinations of regular cloning, Golden Gate, Gibson assembly or yeast assembly were used to generate the “megachunks” (30–60 kb), which were sequentially introduced into yeast to replace their native counterparts through the strategy called switching auxotrophies progressively for integration (SwAP-In). Together, these results highlight the considerable capacity of the yeast host for DNA uptake and assembly. The fact that all 16 yeast chromosomes can be reorganized into a single linear or circular chromosome suggests that budding yeast may be capable of constructing DNA molecules over 10 Mb (Jiang et al., 2022; Shao et al., 2019; Shao et al., 2018).

In addition to *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Salmonella typhimurium*, and *E. coli* are three alternative hosts for *in vivo* genome assembly (Fredens et al., 2019; Itaya et al., 2008; Itaya et al., 2005; Lau et al., 2017). A 3.5 Mb genome has been assembled into a *B. subtilis* genome by the “inchworm” method (Itaya et al., 2005). Using a *B. subtilis* genome (BGM) vector, a 16.3-kb mouse mitochondrial genome and a 134.5-kb rice chloroplast genome were successfully integrated by homologous recombination into a *B. subtilis* genome (Itaya et al., 2008). A 200 kb *S. typhimurium* segment was replaced by synthetic DNA through a process called stepwise integration of rolling circle amplified fragments (SIRCAS) (Lau et al., 2017). In *E. coli*, a conjugation-based strategy, coupled with repetitive replicon execution for enhanced genome engineering through programmed recombination (REXER), enabled the synthesis of a ~4 Mb recoding genome (Fredens et al., 2019; Wang et al., 2016).

### **Perspective on synthetic genomics**

Bottom-up genome synthesis enables the simultaneous integration of dense and complex genome-wide changes. Synthetic genomics not only return valuable discoveries in the life sciences. But also may lead to a new industrial revolution for food, medical and chemical production (Jiang et al., 2020b; Venter et al., 2022). For example, synthetic

viruses have altered vaccine design and production, and synthetic genomes are being used to save lives through humanized pig organs for transplantation (Venter et al., 2022).

Currently, however, the cost of gene synthesis remains prohibitive for genomes megabases or more in length. In addition, a number of technical hurdles remain to be addressed. First, currently designed genomes are usually assembled in microbial hosts such as *E. coli* or *S. cerevisiae*; however, toxicity of certain DNA sequences to the host often leads to assembly failure. Second, the stepwise assembly of long DNA fragments is limited by sequence repeatability, such as centromeres and telomeres in higher eukaryotes. Third, the transfer of assembled DNA fragments from the host to the target organism remains challenging. Currently, only the *Mycoplasma circular* genome has been successfully transplanted.

Recent developments in high-throughput DNA synthesis and assembly technologies should considerably accelerate the construction of synthetic genomes. The emergence of new DNA synthesis technologies using high-density microchips, enzymatic DNA synthesis, and automated gene assembly using microfluidics will continue to drive down the price of gene synthesis. In addition to assembling large DNA fragments *in vivo*, new techniques for synthesizing and amplifying large DNA *in vitro* will emerge. Within the next five years, the cost of gene synthesis is expected to reach 0.001/base for DNA lengths over 1 Mb. Meanwhile, chromosomes of about 1 Mb in size can be fully synthesized *in vitro*, transferred into the target organism and restart the host, opening a new era of synthetic genomics.

One research direction related to genome assembly is genome simplification, which aims to identify a minimal genome of a living organism. For example, J. Craig Venter’s team removed nearly half of the mycobacterial genome to study the genome composition necessary for cell survival (Hutchison III et al., 2016). A genome compaction strategy based on synthetic chromosomal rearrangements and modifications by *LoxP*sym-mediated evolution (SCRaMbLE) revealed that at least 60% of the genes on the synthetic chromosomal arms (*synXIII*) of budding yeast are dispensable for cell viability (Luo et al., 2021). These studies considerably improved the feasibility of constructing a microbial chassis with minimal genomes and desirable properties. Future exploration of genomic minimization of multiple synthetic chromosomes or entire genomes will considerably expand our knowledge of the core functions of eukaryotes (Dai et al., 2020).

### **DNA data storage technology: a paradigm of BT-IT integration**

#### ***The emerging DNA data storage***

Information storage has always been the driving force of hu-

man civilization, a necessary condition for the accumulation of knowledge, the transmission of culture, and the transmission of technology from generation to generation. The techniques used to preserve information can be traced back to the beginning of papermaking and knot-tying in ancient China, and later in history, to paper and printing. Until nearly half a century ago, storage technology based on magneto-optical silicon, such as hard disks, solid-state disks, and magnetic tapes, has constantly changed the means of information storage.

Modern data storage and processing technologies have brought humanity into the digital age, and the total amount of digital data on the planet has grown exponentially. However, the current storage media are facing a number of challenges: theoretical limits of density, short duration, high energy consumption, and environmental pollution. So, a new generation of information archiving technologies needs to be developed (Ceze et al., 2019; Zhirmov et al., 2016). Surprisingly, DNA, a natural medium for preserving genetic information, has been found to be a potential medium for artificial data storage with high density, long-term durability, and low maintenance costs (Church et al., 2012; Dong et al., 2020; Grass et al., 2015; Bancroft et al., 2001). The use of synthetic DNA for high-density and long-term data storage has become a highly promising area of research, attracting considerable interest worldwide from both governments and industrial investors. The Semiconductor Industry Association, Defense Advanced Research Projects Agency, National Science Foundation, Semiconductor Research Corporation, and Intelligence Advanced Research Projects Activity all contribute to U.S. DNA data storage technologies and related semiconductors. The European Commission has also specifically funded DNA data storage and launched the Horizon 2020 program. The China Association for Science and Technology listed DNA data storage as one of 60 major scientific and engineering technology issues in 2018. China's 14th Five-Year Plan explicitly promotes the development of cutting-edge technologies such as DNA storage. Microsoft and Western Digital, along with Twist Biosciences and Illumina announced in 2020 the creation of the "DNA Data Storage Alliance". Its "common goal is to enable the full potential of DNA Data Storage as a new storage medium across existing and emerging archival storage use cases". To date, more than 50 companies and academic institutes have joined the Alliance (<https://dnastoragealliance.org/>).

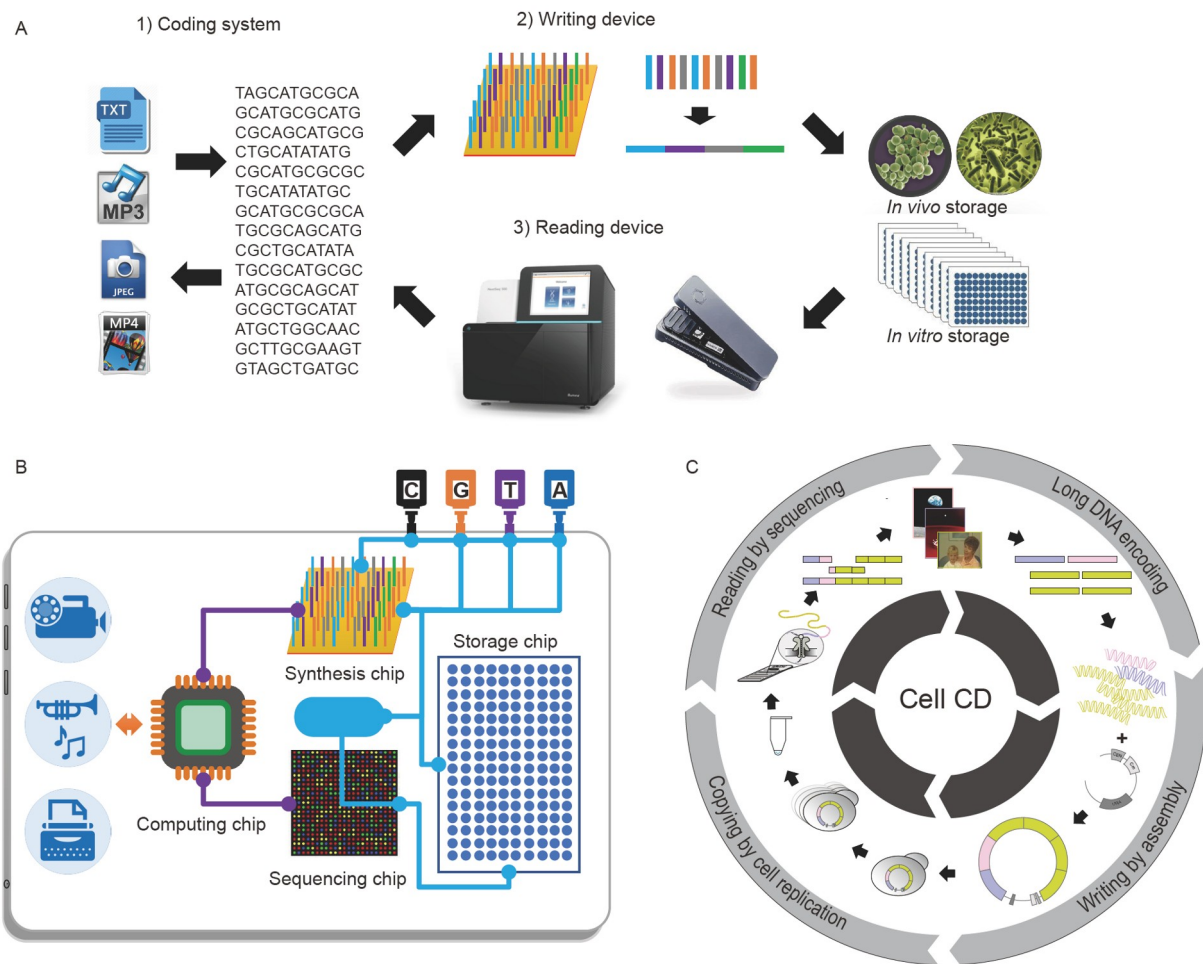
### **Concept and working modes of DNA data storage**

As shown in Figure 3A, the basic concept of DNA data storage consists of three basic components: (i) a coding system that can encode binary strings to DNA strings and adapt the reverse process—decoding DNA strings to binary strings; (ii) a writing device that can make actual DNA molecules with a specific sequence or structure; (iii) a

reading device capable of reading the sequence of the DNA molecules (Han et al., 2021; Meiser et al., 2020). It is worth mentioning that, so far, there have been two different strategies for digital information storage in DNA. For the first strategy, DNA molecules with specific sequences are used to record information. DNA molecules with specific sequences are generated by *de novo* DNA synthesis or assembly for data writing (Church et al., 2012; Goldman et al., 2013; Grass et al., 2015; Organick et al., 2018). The second strategy uses pre-existing DNA molecules as the backbone for data recording. The information is then stored at preset locations on the double-stranded DNA (dsDNA) or single-stranded DNA backbone by gene editing or DNA hybridization to generate precise sequence or structural modifications on the backbone (Chen et al., 2019a; Chen et al., 2020a; Shipman et al., 2017; Tabatabaei et al., 2020). The first strategy offers higher storage densities but is rather expensive to write due to the need for DNA synthesis. The second strategy is expected to be less expensive to write than the first because it bypasses the costly stage of DNA synthesis. However, the reduction of its storage density may limit future applications (Chen et al., 2019a; Chen et al., 2020a). So, DNA data storage technology can be divided according to the technical details of writing, copying, storage and reading, namely into *in vitro* "hard disk mode" and *in vivo* "CD-ROM mode" (Figure 3B and C).

Figure 3B shows the *in vitro* "hard disk mode", which uses high-throughput DNA synthesis to write data and has the potential for high-density data storage, similar to ordinary hard disks. The data writing and reading process in this mode is relatively simple because there is no cell membrane barrier (Church et al., 2012). However, according to studies (Gao et al., 2020; Heckel et al., 2019), *in vitro* storage has been shown to be associated with DNA strand loss during replication, high replication costs, and DNA degradation throughout storage. The *in vivo* "CD-ROM mode", as depicted in Figure 3C, uses artificial chromosomes to store and distribute large amounts of data (Chen et al., 2021b). This *in vivo* model has a protective environment in which efficient replication and repair enzyme systems emerge naturally, offering significant advantages in terms of durability, fidelity, and low replication costs. The main advantage of the "CD-ROM mode" compared to the *in vitro* "hard disk mode" is the low-cost, reliable replication of chromosomal DNA as part of cellular replication, which can be used for fast, low-cost data replication and dissemination. In addition, *in vivo* storage makes it easier to realize more advanced storage functions by constructing complex intracellular biological circuits, such as random reading and writing through gene editing, encryption and decryption, and communication with the flow of biological information. These additional options open up more possibilities for *in vivo* mode, allowing for a wider range of potential application scenarios such as cel-





**Figure 3** The basic concept and storage modes of DNA data storage. A, The basic concept of DNA data storage. To achieve basic data writing and reading operations, DNA data storage requires three basic components of a coding system, a writing device, and a reading device. B, “Hard disk mode” based on *in vitro* synthesis and sequencing of massive DNA fragments. C, “CD-ROM mode” based on the manipulation of *in vivo* chromosomal DNA (redrawn with reference to [Chen et al., 2021b](#)). The details of the two storage modes are described in the main texts.

lular event recording, environmental toxicant detection, and disease marker monitoring.

### Major progresses and the current state of DNA data storage

The feasibility of the *in vitro* “hard disk mode” has been demonstrated on the lab scale. Researchers at Columbia University have introduced Fountain codes into DNA data storage to improve coding efficiency and prevent GC-rich, complex DNA sequences that are difficult to construct and sequence. At a data scale of 2 MB (megabytes,  $10^6$  bytes), a high storage density of 215 PB/g of DNA (PB, petabytes,  $1 \times 10^{15}$  bytes) has been achieved ([Erlich and Zielinski, 2017](#)). In late 2018, researchers at the University of Washington achieved reliable random data access at a scale of 200 MB ([Organick et al., 2018](#)). Furthermore, a fully integrated DNA storage system was built, enabling the automatic writing, storage, and reading of a single word “hello” ([Takahashi et](#)

[al., 2019](#)). Startup CATALOG has taken a different approach, using “DNA movable type” for high-speed data writing. They announced in 2019 that all 16 GB (gigabytes,  $10^9$  bytes) of Wikipedia text can be written in DNA within 12 h ([Versai, 2019](#)), which is nearly 1000 times faster than any other currently used technology. Researchers at the Technion Institute of Technology in Israel have devised a concept of compound DNA letters, which minimize the cost of writing by exploiting the base composition information to improve the ability to write per synthesis cycle ([Anavy et al., 2019](#)). Gao et al. accomplished low-bias DNA strand amplification using isothermal amplification ([Gao et al., 2020](#)). Chen et al. of Tianjin University used Low Density Parity Check (LDPC) and Reed-Solomon (RS) algorithms to encode video clips and text with a total size of 3 MB in DNA ([Chen et al., 2020b](#)). To address the sequence compatibility issue for DNA storage, Ping et al. devised a “yin-yang” encoding system ([Ping et al., 2022](#)). Early proof-of-concept studies of the *E. coli* DNA CD-ROM paradigm used plasmid storage

data (Bancroft et al., 2001; Davis, 1996). Later research has focused on the implementation of genetic circuits, such as toggle switches, for data storage (Gardner et al., 2000). However, the storage capacity of such systems is clearly limited. Shipman et al. utilized CRISPR-Cas9 technology to store digital movies in bacterial cells and enable them to be decoded using high-throughput sequencing (Shipman et al., 2017). Later, Tang and Liu recorded a large number of cellular activities in the cell population by using two CRISPR-mediated analog multi-event recording apparatus systems (Tang and Liu, 2018).

Recently, Chen et al. of Tianjin University designed and synthesized an artificial chromosome with a length of 254,886 bp from scratch for data storage. This study shows for the first time that assembled artificial chromosomes can be used for large-scale data distribution through reliable and low-cost cellular replication (Chen et al., 2021b). New concepts and ideas such as “DNA-of-things”, “bio-orthogonal information storage”, “true random number generation”, “data encryption in DNA”, have also been proposed (Banal et al., 2021; Bee et al., 2021; Fan et al., 2021; Koch et al., 2020) and lead the way for a wide range of potential applications for DNA storage and computation. A recent review (Meiser et al., 2022) provides an excellent summary of these topics.

### ***Future perspective***

DNA data storage involves a range of key technologies, including DNA synthesis, sequencing, microfluidics, micro-nano fabrication, and requires multidisciplinary efforts to achieve the ultimate goal of transforming DNA storage into practical applications. Although previous research has made significant progress in data volume, stability and random access, cost, especially write cost, has become a major obstacle to practical application of DNA data storage (Ceze et al., 2019; Dong et al., 2020; Ping et al., 2019). It is estimated that DNA data storage will need to reduce write costs by 7–8 orders of magnitude over the currently used tape-based storage technologies (Meiser et al., 2022). Despite several attempts, such as non-terminated TdT (Lee et al., 2019), DNA punch cards (Tabatabaei et al., 2020), DNA movable type (Versai, 2019), compound DNA letters (Anavy et al., 2019), and low-quality synthesis (Antkowiak et al., 2020), the competitive route to cost reduction remains unclear. Every information storage medium faces the same high production cost challenges in its early stages.

Modern storage technologies have been widely used in an accordance with Moore’s Law for over decades. It is worth mentioning that DNA synthesis and sequencing, two key technologies in DNA data storage, are developing faster than Moore’s Law predicts. In a brief history of DNA storage, since the first publication of chip-based DNA data storage by

Church et al. in 2012 (Church et al., 2012), the data size has expanded more than 300 times, showing a rapid upward trend.

In conclusion, the author argues that with the continuous development of enzymatic DNA synthesis, data writing and reading methodologies, practical DNA data storage technologies will be available in the near future. As an environmentally friendly, high capacity and long-term storage medium, DNA is expected to compensate for the insufficiency of current storage media.

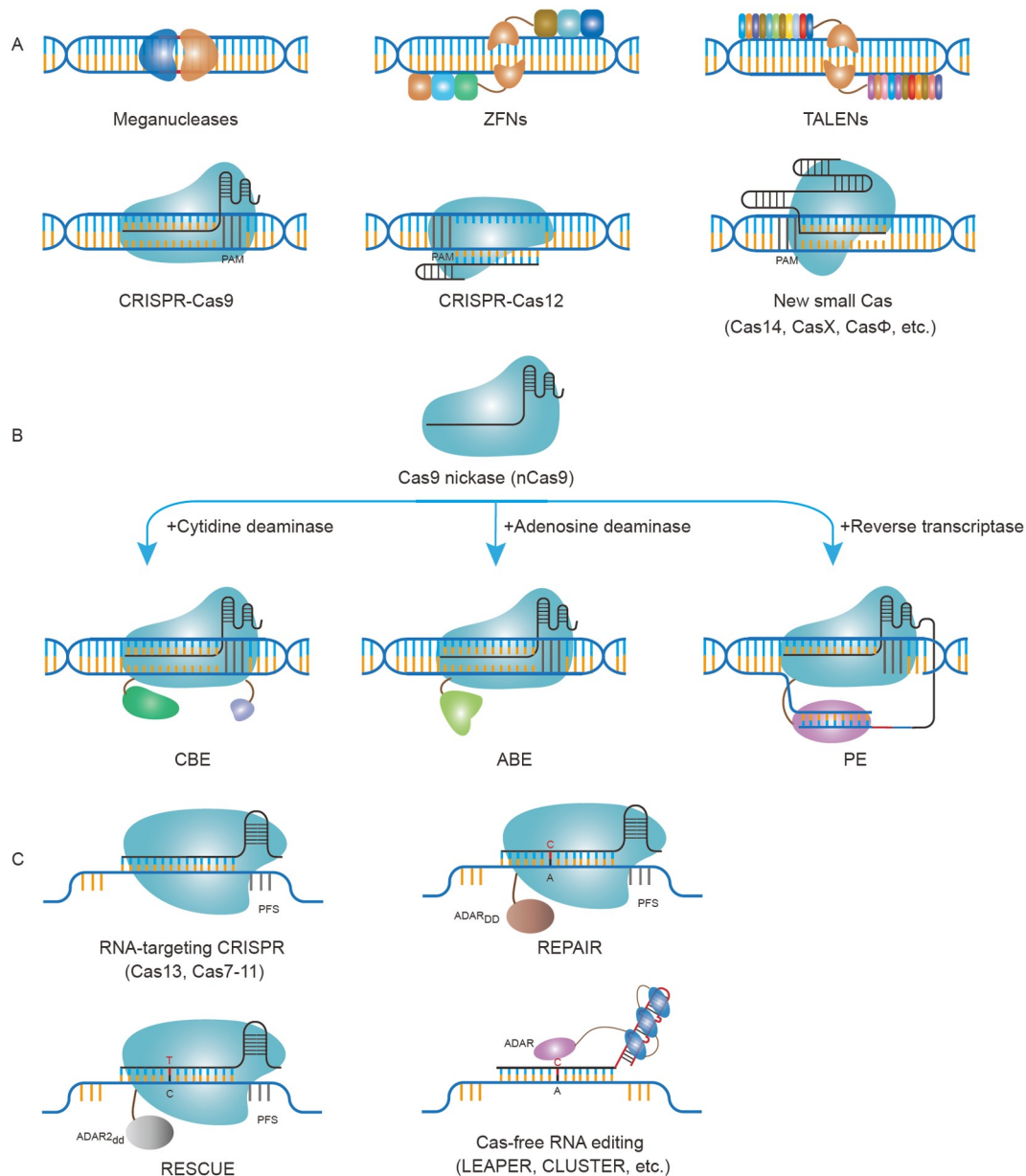
### **Gene editing**

In the life sciences, there has long been a goal of being able to programmably, specifically and efficiently edit the DNA sequence of all living cells, which has unlimited value in gene research, gene therapy, genetic breeding, and synthetic biology. Previous approaches, such as meganucleases, zinc finger nucleases (ZFNs), and transcription activator-like effector nucleases (TALENs), rely on complex and specific protein-DNA interactions to target protein effectors to desired DNA sequences. While effective for targeting specific loci, it is difficult to rapidly and simply reprogram the targeting of these protein domains to new genomic loci of interest. The discovery and engineering of the clustered regularly interspaced short palindromic repeats (CRISPR) system has sparked a new and exciting renaissance in the field of genome editing.

### ***Strictly protein-based genome editing systems***

Meganucleases, ZFNs, and TALENs are powerful biological tools that can be used for genome editing (Figure 4A). Meganucleases (also called homing endonucleases) are large protein complexes that recognize specific DNA sequences. These proteins rely on a complex network of interactions between the protein itself and the target DNA sequence. Although previous efforts had successfully applied meganucleases to new, user-defined genome sequences (Epinat et al., 2003; Silva et al., 2011), the process was extremely laborious, time-consuming, and technically challenging. The application of meganucleases fundamentally relies on the large-scale reprogramming of entire protein complexes, enabling them to recognize new DNA regions of interest. Therefore, there is urgent need to identify novel variants against newly defined protein sequences using high-throughput methods that require variant libraries. Thus, more programmable methods are required for thermostable DNA targeting.

Zinc finger proteins are small protein modules capable of recognizing specific sequences of three DNA bases. These proteins are commonly found in nature, and previous studies



**Figure 4** Overview of genome editing technologies. A, Nuclease-based genome editing technologies that target DNA, including meganucleases, ZFNs, TALENs, CRISPR-Cas9, CRISPR-Cas12 and new small Cas variants. B, Precision DNA genome editing technologies including the cytosine base editor, adenine base editor, and prime editor. C, RNA editing technologies including CRISPR-Cas13, CRISPR-Cas7-11, and RNA base editing approaches like REPAIR, RESCUE, and other Cas-free RNA editing approaches.

have identified key components of individual zinc finger modules that determine specific 3-base pair DNA-binding sequences. A modular zinc finger array can be fused together to enable DNA targeting based on specific DNA sequences. In addition, the researchers cleverly fused these larger zinc finger proteins involved in DNA targeting with the *FokI* protein, which can cut DNA. To minimize all undesired random DNA cleavage in living cells, the researchers cleverly split the *FokI* protein into two halves, each recruited to target regions of DNA using specific zinc fingers (Kim et al., 1996; Urnov et al., 2005). Therefore, the combination of two

zinc finger nucleases (ZFNs) can specifically and precisely cut DNA. These ZFNs have been shown to function in human, animal, and plant cells, and thus play an important role in programmable genome editing.

Following the discovery of zinc finger proteins, researchers identified transcription activator-like (TAL) effectors from plant pathogens. Unlike ZFs, each TAL effector (TALE) binds to a single DNA base. This effect can be programmed to bind to specific DNA sequences. TALEs bind to *FokI* dimers to generate TALE nucleases (TALENs), a fully protein-based programmable genome editing tech-

nology (Christian et al., 2010; Li et al., 2011b). Compared with ZFNs, TALENs exhibit better programming ability because each DNA base is recognized by a single unit, rather than the triplet-encoded property of zinc fingers, but TALENs are larger than ZFNs, thus delivery remains challenging. In addition, protein complexes need to be constructed, which is not easy when seeking to extensively edit the genome of living cells.

### **CRISPR-Cas systems**

While studying bacterial genomes, researchers identified a repeating stretch of DNA named as a clustered regularly interspaced short palindromic repeats (CRISPR) array (Ishino et al., 1987; Jansen et al., 2002). Through subsequent studies, the researchers demonstrated that CRISPR arrays and their nearby proteins, CRISPR-associated (Cas) proteins, function as bacterial immune systems against foreign invading nucleic acids (Barrangou et al., 2007; Marraffini and Sontheimer, 2008). When bacteria are exposed to pathogenic DNA fragments, the immune system isolates a small portion of the foreign DNA and integrates that sequence into the CRISPR array in the bacterial genome itself. This discovery is critical to the development of CRISPR-Cas as a revolutionary genome editing technology.

The CRISPR arrays were identified to encode RNA sequences associated with Cas proteins and target a protospacer-based nucleic acid sequence in the DNA. Subsequent engineering demonstrated that targeting sequences in CRISPR RNAs can be easily replaced and programmed with user-defined sequences (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2012; Mali et al., 2013) (Figure 4A), which would completely alter and reprogram the Cas protein targeting sequence. This discovery plays an important role in the field of genome editing, as this is the first time that genome editing reagents can be easily reprogrammed by replacing nucleic acid sequences, unlike previous approaches that required complex and high-throughput protein engineering. Once bound to a target DNA sequence, the Cas protein initiates cleavage of double-stranded DNA, creating damage in the genome of living cells.

### **New CRISPR proteins**

The Cas protein is a key component in the CRISPR genome editing technology. *Streptococcus pyogenes* (Sp) Cas9 was the first engineered Cas protein for genome editing applications and will remain widely used when developing new editing techniques. All Cas proteins are known to require a protospacer-adjacent motif (PAM), a small stretch of DNA located directly adjacent to the target genomic locus. This targeting range limitation of Cas proteins remains a challenge when trying to edit at other positions in a cell's gen-

ome. Researchers have discovered a large number of novel Cas proteins with diverse PAM requirements (Ran et al., 2015; Zetsche et al., 2015), thereby expanding the targeting range of CRISPR genome editing technology. In addition, protein engineering and directed evolution efforts have successfully altered the PAM requirements of Cas proteins (Miller et al., 2020b; Nishimasu et al., 2018; Walton et al., 2020), which has contributed to the development of a range of genomic targets using CRISPR Cas.

Recently, many new small Cas proteins have been discovered. The length of SpCas9 is 1368 amino acids, which is further extended by effector proteins in the base editor and prime editor. The stability and delivery of genome editing proteins are negatively affected by increased length. New CRISPR-Cas proteins, such as Cas12f (Kim et al., 2022; Wu et al., 2021; Xu et al., 2021a), formerly called Cas14 (Harrington et al., 2018), CasΦ (Pausch et al., 2020), CasX (Liu et al., 2019a), are smaller than many previously discovered Cas proteins (Figure 4A). However, further engineering, discovery and evolution efforts are required to improve the editing efficiency of these new Cas proteins.

### **Genetic knockout**

Meganucleases, ZFNs, TALENs, and DNA-targeting CRISPR-Cas systems all operate by cleaving double-stranded DNA. Following the generation of DNA double-strand breaks (DSBs), a cell's endogenous repair machinery rapidly repairs the lesions. Perfect repair can serve as a substrate for additional editing until non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ) repair results in random small DNA insertions or deletions (INDELs) around the target site (Moore and Haber, 1996). INDEL results in gene knockout, which is useful in certain cases but lacks precision. Homology-directed repair (HDR) is a competitive repair process in which a nucleic acid donor template is used to repair DNA (Liang et al., 1998). Although programmable, HDR is extremely inefficient when compared to NHEJ/MMEJ repair. Therefore, new genome editing technologies are needed to edit DNA sequences efficiently and precisely.

### **Base editing**

Base editing is a programmable, efficient and precise genome editing technology built on the ability to localize DNA-binding proteins to sequences of interest. The first class of base editors designed, called cytosine base editors (CBEs), exploited the ability of Cas proteins to bind and unwind target regions of DNA into a single-stranded DNA state (Figure 4B). CBEs consist of a single-strand-specific cytidine deaminase fused to Cas proteins that deaminate regions of endogenous DNA targeted by Cas proteins (Komor et al.,



2016). Deamination of cytosine bases in DNA produces uracil, which can be replicated and repaired to thymine by endogenous cellular processes. To improve editing efficiency, CBE also includes uracil glycosylase inhibitor (UGI) to inhibit endogenous uracil N-glycosylase (UNG), which specifically recognizes the presence of uracil bases in the cell genome (Acharya et al., 2003; Komor et al., 2017; Krokan et al., 2002). The presence of localized UGI would further prolong the lifespan of the uracil intermediate, thereby promoting the permanent incorporation of thymine after repair.

To further facilitate editing, the Cas protein is converted into a nickase that cleaves the strand opposite the edited strand, using the opposite side chain of the DNA containing the new base-edited uracil as a repair template, manipulating the cellular repair machinery to replace the nicked trace chain. This finally achieves permanent editing from one DNA strand to two DNA strands, significantly increasing the efficiency of base editing.

Adenine base editors (ABE) were the second class of base editors developed. ABE consists of a laboratory-evolved adenosine deaminase that converts adenine bases in DNA to inosine (Gaudelli et al., 2017) (Figure 4B). Inosine is subsequently recognized as guanine by endogenous cellular polymerases. Advanced directed evolution methods have further improved editing efficiency and expanded the utility of adenine base editing (Gaudelli et al., 2020; Richter et al., 2020).

The original base editing methods used Cas proteins to unwind DNA and expose single-stranded DNA sequences as substrates for deamination. A new class of base editors called DddA-derived cytosine base editors (DdCBEs) utilize a naturally occurring double-stranded DNA cytidine deaminase called DddA to perform base editing without unraveling the DNA (Mok et al., 2020). DNA-binding proteins, such as TALEs or ZFs, can be fused to split DddAs and UGIs to direct cytosine base editing to target DNA sequences in the absence of Cas proteins. Further, by fusing catalytically impaired DddA variants with an adenine deaminase TadA8e, targeted A-to-G editing was achieved in human mitochondria (Cho et al., 2022).

CBE, ABE, and DdCBE can all edit DNA precisely and efficiently to create CG-to-TA (CBE and DdCBE) or AT-to-GC base (ABE) conversions. However, there are many other types of genome editing, such as other base conversions and programmable insertions and deletions, which require newer precision editing techniques.

### Prime editing

Prime editing is a precision genome editing technology that uses a Cas protein's ability to bind DNA and nick one strand of DNA (Anzalone et al., 2019) (Figure 4B). In contrast to base editors, prime editors nick the unwound single-stranded

R-loop DNA following Cas binding. Following the specific nicking of this strand, the released DNA can serve as a primer to perform subsequent DNA polymerization.

Another critical component of prime editors is the prime editing guide RNA (pegRNA), which encodes a primer binding site that is complementary to the released single-stranded R-loop from the Cas protein nick and a template region encoding a particular desired DNA editing event. Following RNA-DNA hybridization, a reverse transcriptase protein fused to the Cas protein can use the pegRNA as a template to extend the genomic DNA. An orthogonal Cas protein guide RNA which targets a region 3' of the prime edit can be used to further enhance editing. Following subsequent DNA replication and repair, the newly synthesized DNA sequence can permanently be integrated into the genome, resulting in a programmable and versatile edit dictated by the pegRNA sequence. Initial demonstrations of prime editing were relatively low in editing efficiency (Lin et al., 2020), however, subsequent modifications of the procedure, such as optimal primer binding melting temperatures (Lin et al., 2021), the use of two pegRNAs (Anzalone et al., 2021; Choi et al., 2022; Lin et al., 2021; Wang et al., 2022), DNA repair manipulations (Chen et al., 2021a), and RNA stability motifs (Nelson et al., 2022), and modification of the reverse transcriptase enzyme (Zong et al., 2022) have greatly improved prime editing efficiencies.

### RNA editing

Genome editing in RNA can avoid permanent changes to the genome, thereby reducing the risk of off-target DNA editing. A class of RNAs targeting Cas proteins, such as Cas13a (Abudayyeh et al., 2017; Cox et al., 2017; Xu et al., 2021b) and Cas7-11 (Özcan et al., 2021), programmably target RNA sequences determined by CRISPR guide sequences (Figure 4C). Similar to the development of DNA base editors, researchers developed RNA base editors by fusing RNA-specific adenosine deaminase to RNA-targeting Cas proteins (Figure 4C). The adenosine deaminase RNA-specific (ADAR) protein is fused to Cas13a through the "RNA editing with programmable A-to-I replacement (repair)" technique, which converts adenine to inosine (similar to ABE DNA bases) (Cox et al., 2017). Similarly, engineered ADAR proteins that can deaminate cytosines in RNAs were used to develop "RNA editing for specific CU exchange (RESCUE)" technology, which converts cytosine bases to uracils in RNA (similar to CBE DNA base editor) (Abudayyeh et al., 2019).

New CRISPR-free RNA editing systems have been developed to perform site-specific RNA editing by exploiting the ability of RNA nucleic acids to recruit endogenous proteins to chemically react with RNA (Merkle et al., 2019; Reautschnig et al., 2022) (Figure 4C). Furthermore, a par-

allele technique demonstrated that longer RNAs can naturally recruit ADARs for A-to-I editing of RNAs (Qu et al., 2019). Recently, aggregation design and loop design of RNA have considerably improved the editing efficiency and specificity of RNA editing techniques (Katrekar et al., 2022; Reautschnig et al., 2022; Yi et al., 2022).

The past decade has been marked by the rapid development of new genome editing technologies. From initial protein-based approaches to precise genome editing techniques such as prime editing, the ability to manipulate the genomes of living cells and organisms is increasingly exciting. RNA editing techniques are also starting to become more precise and efficient. The continued development of smaller, precise, accurate, and efficient genome editing tools is urgently required, especially for application in areas such as therapeutics, agriculture, and biological research.

### **Applications of genome editing**

The development of genome editing technologies has enabled great advances throughout biomedicine and agriculture. There has been a flurry of advances from biotechnology companies to generate new genome editing medicines. Recently, researchers have advanced *in vivo* genome editing technologies such as CRISPR-Cas9 and base editing to treat genetic disorders like sickle cell anemia (Newby et al., 2021), progeria (Koblan et al., 2021), transthyretin amyloidosis (Gillmore et al., 2021) or genetic conditions like hypercholesterolemia (Verve Therapeutics, 2022. <https://ir.vervetx.com/news-releases/news-release-details/verve-therapeutics-doses-first-human-investigational-vivo-base>).

The application of genome editing in agriculture has sparked new excitement for future biological crop breeding (Gao, 2021). Disease resistance and herbicide resistance are two of the most developed across many crop species. Recently, researchers demonstrated that four simultaneous multiplex editing events enabled the creation of disease resistance and increased yield wheat plants (Li et al., 2022). Furthermore, many endogenous edits have been shown to generate robust herbicide resistance (Zhang et al., 2019a; Zhang et al., 2021c). Genome editing will continue to enable the creation of valuable agricultural crop species.

### **Molecular evolution of proteins**

*In vitro* molecular evolution of proteins accelerates the natural evolution of protein in a test tube, creating an infinite opportunity for protein science and application. The original contributor of the method, Frances H. Arnold, shared the Nobel Prize in Chemistry (2018). In recent years, great efforts have been made to build more efficient method for

directed evolution of proteins, which not only contribute to a deep understanding of fundamental science of proteins, but also can create enzymes and antibodies that are superior to natural or non-existent ones, and promote the application of synthetic biology.

### **Structure-based evolution**

A strategy based on a deep understanding of protein structure-function relationships, called rational design, can generate desired mutants in a short period of time. The precise definition of mutational “hot spots” is the key to achieving the desired results. Furthermore, building smaller but intelligent mutation libraries can considerably speed up the evolutionary process (Chica et al., 2005).

With the rapid development of bioinformatics, ‘hot spot’ predictions have become popular due to the release of mutational limits at certain residue positions that can have a significant impact on the specific function of enzymes (Lutz, 2010). Computational tools have been developed to identify and assess favorable hotspots (Ofra and Rost, 2007). For example, the ConSurf web server (Ashkenazy et al., 2010) can analyze evolutionarily conserved patterns of protein structures, the LigPlot+ program (Laskowski and Swindells, 2011) can generate schematic diagrams of protein-ligand interactions, and CAVER 3.0 (Chovancova et al., 2012) can visualize protein structures in tunnels and channels. The web server PoPMuSiC (Dehouck et al., 2011) can estimate recent protein stability changes, and the algorithms ASRA and Innov’SAR are well suited as guides for saturating mutations at sites within the binding pocket to enhance stereoselectivity and activity (Cadet et al., 2018; Li et al., 2019a).

Various robust strategies focusing on active site engineering have subsequently been developed and have been used in lipases (Liu et al., 2015), glucanases (Niu et al., 2016), xylanases (Wang et al., 2021a) and other major achievements in the evolution of enzymes. Structure-directed mutational screening of multiple residues in the substrate-binding pocket of thioesterase TesA strongly alters its substrate selectivity (Deng et al., 2020). The active site stabilization (ACS) strategy effectively enhances the enzymatic kinetic stability of the lipase CalB by increasing the rigidity within the directed active site (Xie et al., 2014). Non-standard amino acid (ncAA) technology has significantly expanded the functional scope of synthetic polypeptide materials by incorporating new chemical functions that may facilitate material fabrication (Wu et al., 2013). Click-reaction modification was applied to protein modification to increase the molecular weight of protease, and dextran was used as a modifier to successfully optimize the application of protease in wool biofelting processing (Shao et al., 2019).

Through structural and phylogenetic analysis, loop remodeling reconstituted a phosphotriesterase (PTE) with

PTE-like lactose activity within several mutational steps, demonstrating the potential role of loop remodeling for rapid differentiation of new enzyme functions (Afriat-Jurnou et al., 2012). The stepwise loop insertion strategy (StLois) identifies target regions through structural and functional analysis of the corresponding enzymes, effectively expanding the residues in the loop regions to provide new structures of the enzyme active site for new catalytic properties (Hoque et al., 2020). Domain swapping helps reveal structural and functional information on important regulators, such as  $\beta$ -repressors (Ghosh et al., 2019) and decay accelerators (Panwar et al., 2016).

Semi-rational design introduces random mutations at selected residues, as saturation mutation creates a small library of mutants containing all possible mutations at the selected residues, a small number of which may be beneficial in the mutated protein. Notably, with the help of codon degeneracy, extended versions of the combinatorial active site saturation assay (CAST) (Reetz et al., 2005) and iterative saturation mutagenesis (ISM) (Reetz and Carballeira, 2007) were efficient construction. Considerable progress has been made with the creation of “smart libraries” (Qu et al., 2020). These approaches have been reported to successfully improve enzyme properties such as thermostability, catalytic activity, and enantioselectivity (Fan et al., 2022; Tan et al., 2022; Tong et al., 2022). The combination of enzyme engineering and systems metabolic engineering has also significantly increased the metabolic flux of target products (Qian et al., 2019; Yang et al., 2020).

### **Random mutagenesis**

Directed evolution, which does not rely on the structural information of enzymes but on the sequence information of enzymes, offers a promising way to obtain desired mutants in the laboratory over months rather than millions of years. The sequence space for variation is very large, for example, mutating at four residues may yield 160,000 ( $20^4$ ) sequences. One of the key issues in directed evolution is how to efficiently generate mutant libraries.

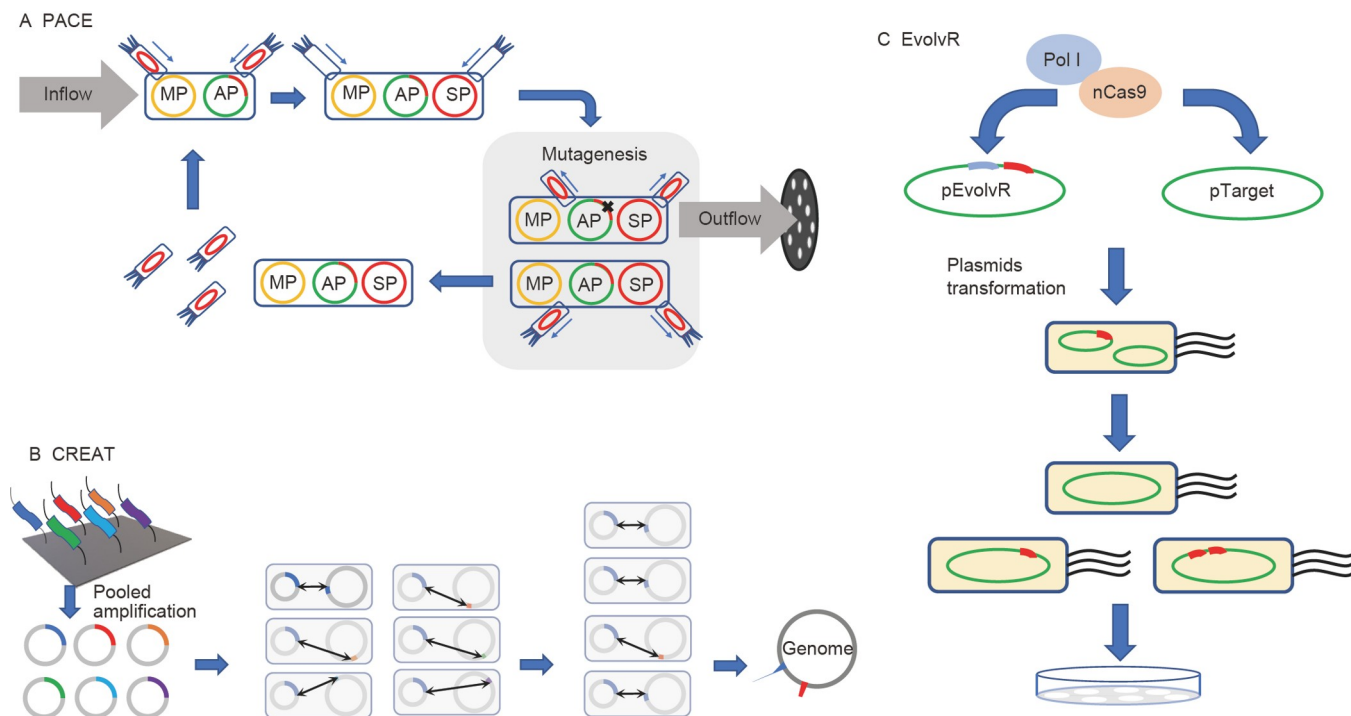
The general method is error-prone PCR (epPCR), which introduces changes in genes. The researchers considerably increased the mutation rate by changing the PCR reaction conditions. Zaccolo et al. re-scaled the mutation rate to one mutation every five base pairs by changing PCR conditions and the number of mutation PCR cycles (Zaccolo and Gherardi, 1999). So far, epPCR has achieved many successes, such as improving the activity, affinity and stability of enzymes and substrates (Brands et al., 2020; Ruan et al., 2022; Zhou et al., 2021; Zhu et al., 2019). Most importantly, epPCR also represents a powerful method for studying molecular evolution by analyzing large-scale sequence diversity (Konno et al., 2022).

DNA shuffling mimics natural homologous recombination, another mechanism of natural evolution. During DNA shuffling, two or more related starting genes are recombined, resulting in a pool of variant genes with new combinations of random sequences. Compared to epPCR, DNA shuffling combines fragments of related functional proteins, resulting in novel sequences with a relatively high probability of being compatible with the desired protein structure and function. An example is the generation of active halogenase variants from catalase with altered selectivity to expand the enzymatic halogenation capacity of unactivated C–H bonds (Neugebauer et al., 2021). Similarly, motif shuffling based on BRC repeat modularity was used to generate stronger chimeras that bind to RAD51 (Lindenburt et al., 2021).

Recently, a number of promising techniques have been developed for *in vivo* protein evolution (Figure 5). These methods generate multiple random mutations directly within the host organism by localizing mutant enzymes or nucleases into DNA (Kim et al., 2019). CRISPR/Cas9 ushered in a new era of genome editing, which has also been applied to protein engineering. The EvolvR system consists of Cas9-nickase, and an error-prone DNAP I that continuously generates mutations in tunable windows under the guidance of gRNAs (Halperin et al., 2018). More specifically, a new *in vivo* mutation method, CRISPR-Enabled Traceable Genome Engineering (CREATE), utilizes the CRISPR/Cas9 system and barcoded tracking cassettes to mutate multiple sites and track them (Garst et al., 2017). It can form single-base libraries for entire protein sequences to construct a saturated library in which every amino acid residue is substituted (Liang et al., 2017; Reynolds et al., 2017). Phage-assisted sequential evolution (PACE) is another strategy for *in vivo* evolution. It exploits the survival of the M13 phage to mutate genes in *E. coli*. In general, PACE is able to evolve any protein associated with basic phage gene expression (Miller et al., 2020c). Due to the rapid generation time of the phage life cycle, dozens of rounds of evolution can occur in a day without human intervention. In addition, T7 RNA polymerase is used in several *in vivo* protein evolution systems because of its binding affinity to DNA. MutaT7 is a chimeric protein containing T7 RNA polymerase and cytidine deaminase that can edit or mutate specific genes downstream of the T7 promoter (Moore et al., 2018a). More recently, targeted *in vivo* diversification via T7 RNAP (TRIDENT) has been developed based on an evolutionary platform of T7 RNA polymerase, exploiting increased mutational diversity and higher *in vivo* mutation rates (Cravens et al., 2021).

### **High-throughput screening**

High-throughput screening is a technique used to obtain desired mutants from large variant libraries. Microtiter plate-based screening methods are the most commonly used



**Figure 5** Overview of emerging *in vivo* mutagenesis methods. A, Schematic of PACE. PACE utilizes the survival of M13 phage to mutate genes in *E. coli*. MP, mutagenic plasmid; AP, accessory plasmid; SP, selection plasmid. B, Schematic of EvolvR. EvolvR utilizes a chimeric protein of error-prone Pol I and nicking variant of Cas9 to specifically mutate genes targeted by gRNA. nCas9, nicking variant of Cas9. C, Schematic of CREATE. CREATE utilizes CRISPR-Cas9 with barcode-tracking of mutations for multiplex genome engineering.

methods in enzyme-directed evolution. These systems have the advantages of simple installation, convenient operation and strong versatility. However, the screening capacity is relatively low, usually limited to  $10^3$ – $10^4$  colonies per day. To increase the speed of screening, automated equipment such as robotic liquid handling units and colony picking systems have been developed (Nirantar, 2021). In high-throughput screening methods, colorimetric or fluorogenic substrates are often used to measure enzymatic activity (Giger et al., 2013). This screen can also be combined with pH indicators (Bornscheuer et al., 1999; Moris-Varas et al., 1999) or enzymatic cascades that generate absorbance or fluorescence signals to create high-throughput screening methods (Malhotra et al., 1996).

Growth-complementary selection is a powerful screening method whenever the target enzyme is critical for host cell survival. This approach has been widely applied to enzymes associated with major metabolic pathways, including tRNA synthetases (Zhao et al., 2021), proteases (Verhoeven et al., 2012), amino acid synthesis isomerases (Jürgens et al., 2000), and more. Similarly, enzymes with desired functions, such as base editors, can be screened for by rescuing defective antibiotic resistance genes containing point mutations at key positions (Gaudelli et al., 2017).

Fluorescence-activated cell sorting (FACS) (Aharoni et al., 2006) and fluorescence-activated droplet sorting (FADS) (Agresti et al., 2010) have screening throughputs larger than

$10^6 \text{ h}^{-1}$ , making them ultra-high-throughput screening technical benchmarks. In pioneering studies, fluorogenic substrates for glycosyltransferases were designed that can move freely in and out of cells, and the fluorescent products can be captured in cells and screened by flow cytometry (Tan et al., 2019; Yang and Withers, 2009). For cells that cannot absorb desired substrates or retain fluorescent signals, FADS uses droplets as enzymatic microreactors to separate individual cells. Microfluidic chip systems allow for multiple operations such as droplet production, cell lysis, reagent addition, incubation, fluorescence detection, and dual-channel screening (Ma et al., 2018b). Recently, a method combining FACS and FADS was invented in which intact double-emulsion droplets can be selected using a commercial FACS instrument (Brower et al., 2020).

Protein display technology is an important platform for screening protein or peptide binding activity. Phage surface display was first used to study antigen-antibody binding (Smith, 1985). Various cell display methods were subsequently invented, such as bacterial display (Charbit et al., 1986) and yeast surface display (Schreuder et al., 1993). Cell display methods are also widely used for directed evolution, such as improving the stability of  $\beta$ -lactamases (Kather et al., 2008) and expanding the substrate spectrum of DNA polymerases (Chen et al., 2016). Likewise, cell-free display methods such as ribosome display (Hanes and Plückthun, 1997) and mRNA display (Amaral et al., 2017; Wilson et al.,



2001) have accelerated the directed evolution of enzymes. Compared to phage display, cell surface display provides a larger display surface and can also be screened by FACS/FADS if relevant fluorescence assays are available. In addition, the cell-free display system overcomes the limitations of cell-based display methods on transformation efficiency as it can handle libraries of up to  $10^{14}$  members and is also suitable for generating toxic or unstable proteins.

## Computer-aided design of functional proteins

Proteins are the main cellular macromolecules with a plethora of biological functions and constitute the basic building blocks of biological systems. However, since the sequence structure-function space of protein systems is considerable, it is extremely challenging to mathematically solve protein-related problems. The efficient design of proteins, one of the core tasks of synthetic biology, significantly compresses the search space at the expense of acceptable accuracy. The goal of computational protein design is to employ algorithms to create proteins that can fold into specific structures and have desired functions (Brini et al., 2020). With breakthroughs in the computational prediction of protein structures and the continuous emergence of sequence design algorithms, it has become possible to develop computational protein design platforms that support synthetic biology (Huang et al., 2016).

### *Algorithms for designing protein structures*

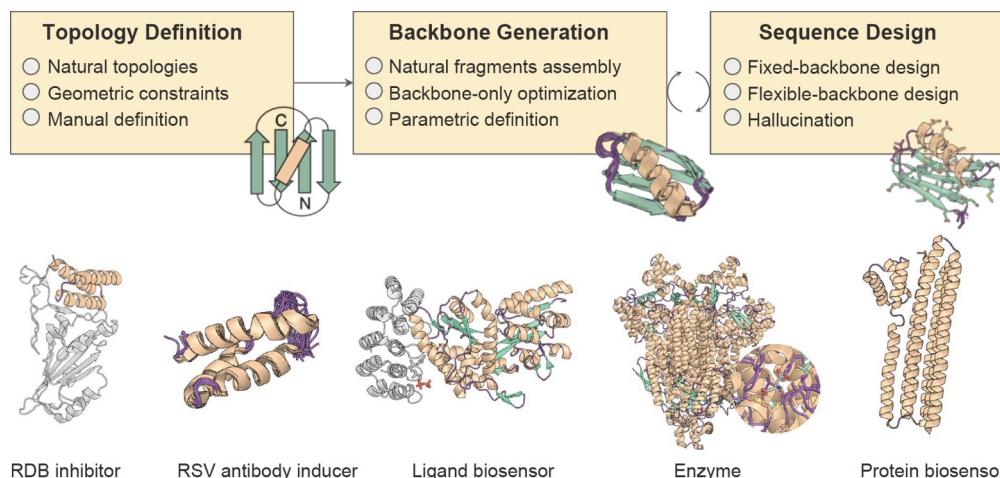
Currently, protein sequences are usually designed with a fixed backbone according to data from the existing protein structures. Compared with the given narrow structural space, the corresponding protein sequence space is considerable, and the epistatic negative influence can significantly weaken the foldability of the designed protein. Therefore, sequence design requires the development of targeted algorithms and strategies. Commonly used methods for computational protein design can be divided into the following categories (Richter and Baker, 2013). (i) Backbone generation: construct the backbone conformation model according to the requirements for sequence design. (ii) Side chain layout: according to the structure of a given protein framework, a set of suitable amino acid side chain conformations are selected to meet the requirements of the backbone structure. This requires the actual design of the sequence, also known as protein sequence design. (iii) Rigid body placement: fix the relative spatial position and orientation between proteins/proteins or proteins/small molecules. (iv) Negative film design: increase the energy of non-target states and achieve effective folding, which can be considered as an optimization and supplement to the sidechain layout algorithm.

Computational design of proteins typically involves three

steps. First, discrete side chain conformations are placed on the main chain. Next, the energy between the inserted side chain and the native side chain and between the side chains and the backbone is calculated. Finally, the combination of sequence and conformation is optimized by a search algorithm (Figure 6). The entire process involves the optimization of a series of sequence combinations and their corresponding structures through a search algorithm. Fixed backbone frameworks are given in advance (e.g., derived from native protein structures). The type of amino acid residues at each backbone position and their side chain conformations are unknown and should be calculated. The possible combinations of structural states and the choice of residues at different positions make up the amino acid sequence and side chain structure space. Energy functions defined in this space are used to evaluate specific sequence and conformation combinations. Search algorithms automatically search an unknown number of sequence spaces and side-chain conformations to find the lowest-energy solution to design protein structures. In order to correctly mimic the mutated side chain conformation, it is necessary to redesign the existing structure. This step is usually performed using a backbone-dependent rotor molecule library of the software, while the optimization of the side chains is energy-dependent.

The energy function is the basis for characterizing the different conformational structures of each sequence combination. Different algorithms use different energy functions, mainly including physical energy terms (non-covalent van der Waals interaction, electrostatic energy, hydrogen bond energy, solvation free energy) and statistical energy terms (main chain dihedral angle, side chain twist). The most widely used energy functions are the Rosetta energy function (mainly determined by the physical energy term) (Leman et al., 2020) and the backbone-based amino acid usage survey (ABACUS)/side chain-unknown backbone arrangement (SCUBA) energy function (mainly determined by the statistical energy term) (Xiong et al., 2014; Huang et al., 2022).

In protein design with a fixed backbone, the lengths and angles of covalent bonds are usually constant and the main interactions to be considered are non-covalent. In the Rosetta energy function, the Lennard-Jones potential is used to calculate the van der Waals interaction energy. The electrostatic energies were calculated using the raw atomic charge distribution of the CHARMM molecular force field and adjusted by group optimization. The energies of hydrogen bonds are calculated using the electrostatic model and a special hydrogen bond model, and the hydrogen bonds are classified into four types: long-range main chain hydrogen bonds, short-range main-chain hydrogen bonds and hydrogen bonds between main-chain and side-chain atoms. Hydrogen bonds between side chains are calculated separately. The Lazaridis-Karplus implicit Gaussian exclusion model



**Figure 6** Principles and examples of computer-aided protein design.

can include isotropic and anisotropic solvation free energies to describe solvation effects. The statistical energy term represents the energy obtained by transforming the probability distributions present in the database. From the point of view of statistical thermodynamics, in the equilibrium state, the energies and probabilities of the different microstates of the system obey the Boltzmann distribution.

An alternative view is that from a purely statistical perspective, assuming the distribution of amino acid sequences for a given backbone structure can be written as conditional probabilities, the problem to be solved by sequence design is to find the sequence with the largest conditional probability. Thus, ABACUS incorporates different structural features: structural type of amino acid positions; backbone dihedral angles; solvent accessibility; relative positions; and statistical information between residues to obtain side chain rotamers (rotor isomers) and atomic packing energies. In addition, SCUBA utilizes neural networks to learn explicit energy terms from the backbone-centered structural variable energy landscape. Together, SCUBA and ABACUS provide comprehensive solutions for the design of artificial proteins.

Search algorithms are also critical for protein sequence design to avoid traversing all conformational combinations in considerable sequence spaces and even larger conformational spaces. Therefore, as a stochastic software, Rosetta was designed based on the Monte Carlo method to perform statistical analysis of conformations generated by multiple simulations and then obtain numerical solutions. Rosetta first uses a random number generator to generate random images. A random perturbation is then confirmed, and new conformations are scored, accepting all conformations with higher scores and those with lower scores with a certain probability, until the best score is selected within a given number of cycles. However, such iterative algorithms are often trapped at local minima. To obtain the global energy minimum conformation, in addition to molecular dynamics

simulations, the physical concept of momentum was used. Imagine a small ball rolling down from a high-energy state. When the momentum is high enough, the ball will not be trapped in a small pit, but will rush towards the final canyon. The iterations take into account not only the current energy, but also previous energy changes.

Several algorithms based on statistics and machine learning were proposed. Inspired by the success of the algorithm trRosetta for structure prediction, Baker et al. further developed the Hallucination protein *de novo* design methods (Anishchenko et al., 2021b). First, a random sequence is fed into trRosetta as input to predict the residual contact map. Then, the amino acid sequence space was sampled using Monte Carlo methods and the KL scattering between sequences was calculated to obtain foldable sequences and predicted structures. The Hallucination method proposes the DeepDream algorithm, based on convolutional neural network, which transforms the input into the training data space and produces (*note tense*) a dream-like illusion. Thus, the Hallucination method can be used to rapidly design protein sequences that are similar to the input sequence and conform to the sequence structure relationships learned by trRosetta, yet differ considerably from the natural sequence.

### **Protein design in synthetic biology**

Sequences designed from protein structures cannot directly meet the needs of synthetic biology for the desired functional proteins. The computational design of proteins mainly includes the design of protein self-skeletons, protein-macromolecule interactions and protein-small molecule interactions. These interactions can be engineered to optimize the function of native proteins as components of synthetic biology, while creating biosensors, biocatalysts, and vaccines with desired functions.

Protein frameworks are designed to enhance the robustness

of native proteins, stabilize vaccine epitopes, and modify protein stability under specific conditions. To develop novel coronavirus inhibitors, and basing these on the structure of a complex of the novel coronavirus S protein and human angiotensin-converting enzyme 2 (ACE2), Baker et al. used the helical fragment of ACE2 bound to the S protein receptor binding region as a starting point. An attempt to stabilize the structure was made by adding two extra helices. In addition, using protein molecular docking and protein interface design in the microprotein library, small proteins capable of inhibiting 2019-nCoV at picomolar concentrations were designed (Cao et al., 2020). Correia et al. developed the TopoBuilder system for the *de novo* design of proteins capable of stabilizing complex pre-defined building blocks. For different epitopes, the authors enumerated suitable two-dimensional protein topologies and constructed tertiary structure models using ideal secondary structures. This method was used to design proteins that present three antigens simultaneously (Sesterhenn et al., 2020). Combining physical energy terms, statistical energy terms, and bioinformatics analysis, Wu et al. developed a greedy accumulated strategy for protein engineering (GRAPE strategy) based on the fusion of a single-point prediction algorithm and a “greedy” algorithm to computationally reshape PET plastic hydrolase, which superimposes a single point mutation that increases the thermal melting temperature of the final mutation by 31°C (Cui et al., 2021a).

Designing protein-macromolecule interactions can be used for signal transduction and regulation in synthetic cells. Computationally designed biosensors by Baker et al. can take advantage of naturally occurring interacting proteins in signaling pathways. In the absence of a detection target, the locking domain of the sensor’s lucCage protein binds to the cage domain. In contrast, in the presence of the detection target, the terminal region of the lucCage domain binds to the detection target, and the lucCage protein opens and binds to the sensor’s lucKey protein, activating luciferase to emit fluorescence (Quijano-Rubio et al., 2021). The same group also designed logic gates to regulate protein binding, constructed *de novo* backbone helical frameworks, and constructed hydrogen-bonding networks to optimize sequences. Multiple protein pairs with specific heterodimers were designed, using monomers or linker monomers as input. The gating unit is constructed to accept different inputs through a designed hydrogen bond network encoding for specific binding (Chen et al., 2020d).

The interaction design of proteins and small molecules can obtain new enzyme catalytic components, transcription factors and small molecule sensors. By designing enzymes with substrate selectivity, new biochemical reactions can be generated for direct use in bioindustrial catalysis as well as new pathway design. In this context, Kortemme et al. screened four residue-binding modules for the binding of

farnesyl pyrophosphate (FPP) to the structure of the native protein. They then designed biosensors in which FPPs could be regulated by interfacing with various frameworks enhancing further optimization (Glasgow et al., 2019). Ranganathan et al. used direct coupling analysis to extract statistical constraints on the space of implicit sequence structure functions in multiple sequence alignment (MSA). They designed a chorismate translocase with comparable activity to the native enzyme (Russ et al., 2020). Wu et al. used a fixed backbone design, combining multiple parallel short-duration kinetic simulations to compensate for uneven sampling of the fixed backbone and sidechains. Thus, aspartate lyase-catalyzed hydroamination of unnatural amino acids was obtained (Cui et al., 2021b).

### **Short summary**

Over the past decade, impressive progress has been made in computationally creating functional proteins with tailored activities and specificities. The astonishing rate of algorithm development continues to improve researchers’ ability to manipulate protein structure and function. Looking ahead, there are numbers of key trends expected to accelerate the discovery, design and application of function proteins. Advances in computational methods for predicting protein structure through AI have raised the confidence of the biomolecular community, subsequent function design may provide access to the demand of target reaction with the help of combination of model-based and data-based methods. As the protein structure databases and standard experimental data continue to grow, more advanced computational methods, will create further research opportunities for interpreting the underlying catalytic mechanisms, eventually leading to a clearer perception of structure-function relationships of function proteins. Based on the considerable success of computational protein design, the future is expected to witness the generation of more efficient, customized proteins for synthetic biology.

### **Cell and gene circuit engineering**

Whether using traditional bioengineering or current synthetic biology, designing cells for beneficial functions has presented a considerable challenge. In the era of synthetic biology, a hallmark of engineered cells is the emphasis on designing and recreating unnatural cellular behaviors at the systems and quantitative levels, which often require more than one component to form interactive networks with specific topology and function. These designable biological networks consist of macromolecules such as proteins, DNA, RNA, or any genetic part within each cell, called gene circuits. It is worth noting that such a network can logically go

beyond the single-cell level, in other words, an interactive multi-cell system is formed through direct or indirect inter-cell contact or communication, which is called a cell circuit. Engineering cellular and gene circuits faces two fundamental challenges: (i) an available genetic component that emphasizes orthogonality and modularity, and (ii) knowledge of circuit modular design principles that provide theoretical guidance for predictable circuit behavior. Furthermore, the design process is highly dependent on sophisticated computational modeling capabilities to analyze and predict circuit behavior in larger circuits and parameter spaces. Therefore, computational aided design of synthetic cells will further facilitate automation and artificial intelligence in future cell engineering, which we will discuss next.

### ***Synthetic gene circuits and quantitative cellular behavior***

Gene circuits conceptually originate from electronic circuits, but substantially differ from electronic circuits due to the enormous complexity arising from the biochemical or biophysical interactions of a large number of components and the nonlinear connections between these components. Similar to gene circuits in natural cells, synthetic gene circuits comprise two basic types: (i) protein-based signaling circuits (or protein circuits) (Chen and Elowitz, 2021; Gao et al., 2018) and (ii) transcriptional gene regulatory circuits (or genetic circuits) (Gardner et al., 2000; Hasty et al., 2002). However, these two types differ little, and work coordinatively to control cellular function. Specifically, protein circuits process environmental signals on faster time scales (from seconds to minutes) through membrane receptor proteins (or sensors), and then transmit the signals to downstream gene regulatory circuits, occurring at a longer time scale (from minutes to hours) (Kiel et al., 2010).

In the past few decades, extensive research on synthetic circuits resulted in successful construction of genetic circuits with integrated functions such as logic gates, bandpass, oscillation, adaptation, and polarization (Anderson et al., 2007; Basu et al., 2005; Chau et al., 2012; Elowitz and Leibler, 2000; Shen-Orr et al., 2002). While many of these studies are still in the proof-of-concept stage, the increasing complexity and scale of these synthetic circuits has considerably advanced our ability to design and build complex genetic circuits with increased efficiency and accuracy. A major development in the direction of synthetic circuits is to take full advantage of computer-aided design and automation. To do this, extensive research work is necessary, including well-characterized and standardized genetic components, experimentally validated algorithms and software for building and simulating silicon circuits, and custom-developed automated experimental equipment (Chen et al., 2020c; Jones et al., 2022; Lim et al., 2013; Ma et al., 2009; Nielsen et al., 2016).

Notably, circuit engineering in mammalian cells is un-

derdeveloped due to the difficulties of genetic manipulation and the limitations of various protein or nucleic acid tools. For example, the number of promoters used in mammalian circuit engineering is often in the single digits. In the existing promoter toolboxes, the transcriptional strength of target genes is difficult to continuously regulate, which becomes a major obstacle to experimental verification of circuit parameter conditions in circuit design. In addition, the dynamic range of gene transcription of many inducible mammalian promoters is extremely low, which is not conducive to the construction of circuits that require low basal but highly inducible gene transcription. Similar to bacterial cells, it is difficult to predict the consistency of promoter strength and induction in different mammalian cell lines.

Protein engineering is more challenging than promoter engineering. The protein function is determined by a three-dimensional structure consisting of 20 amino acids, which is considerably more complex than a one-dimensional sequence consisting of 4 nucleic acids. As for sensors, receptor engineering has emerged as an important area for establishing orthogonal cell-to-cell signaling that results in either sensing a given extracellular signal, such as synthetic cytokines and growth factors, or redirecting cells to specific disease signals (Engelowski et al., 2018; Schwarz et al., 2017; Sockolovsky et al., 2018; Williams et al., 2020). Chimeric antigen receptor (CAR)-activated T cells have been an important example of anticancer therapy (Jackson et al., 2016). Many protein types in mammalian cells exist which establish signaling pathways at different levels, including protein kinases/phosphatases, proteases, adaptor/scaffold proteins, transcription factors or epigenetic regulatory proteins. Several protease tools adopted from viruses have been repurposed to control many levels of cellular function (Cella et al., 2018; Fernandez-Rodriguez and Voigt, 2016; Gray et al., 2010). Recent studies have also shown that protein circuit construction of complex logical functions is mainly based on these engineered proteases (Chen and Elowitz, 2021; Gao et al., 2018). Finally, *de novo* protein design is becoming increasingly powerful, especially as a tool for engineering programmable protein-protein interactions (Chen et al., 2019b; Silva et al., 2019). Notably, recent developments in AI algorithms should play a vital role in future protein engineering (Baek et al., 2021). Undoubtedly, the development of protein tools remains a difficult but essential task for mammalian synthetic biology.

Another challenge for mammalian synthetic biology is the complex behavior controlled by the “black box” of natural evolution. These complex behaviors exhibit quantitative properties, the rationale for which remains unclear. These principles govern nearly all important cellular processes, including the cell cycle, control of size and number, robustness and heterogeneity, homeostasis and growth, cell differentiation and death, and more. To date, few synthetic



biology studies have been able to cover these enigmatic questions of life. Encouragingly, the bottom-up approach of synthetic biology has shown new avenues to understand the construction of complex biological systems in far greater detail than previously thought. A striking example is the oscillatory circuit that controls many fundamental biological processes (e.g., cell cycle, circadian rhythms, signaling responses, somitogenesis) (Elowitz and Leibler, 2000; Stricker et al., 2008; Zhang et al., 2017c). As a next step, this oscillatory circuit is expected to act as a “central processing unit,” intelligently controlling the function of engineered cells (Figure 7).

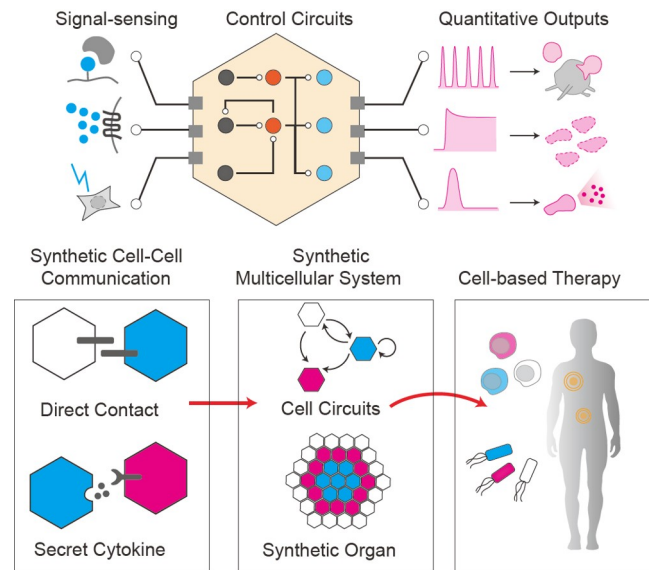
We envision that mammalian cell engineering, along with new tools and techniques, will be one of the next key steps in synthetic biology.

### Cell-cell communication-based cell circuits

A newly emerging field for mammalian synthetic biology is to engineer multicellular systems. This will form interactions based on cell-to-cell communication with specific circuit structures and functions. For bacterial cells, a clear direction is to reconstruct microbial communities ranging from diverse natural environments and disease-associated guts to agriculturally important soils (Bano et al., 2021; Khan et al., 2021). As for mammalian cells, they naturally exist in the context of multicellular interactions, even in well-structured organs. Therefore, cell engineering at the multicellular level represents another major avenue for synthetic biology (Figure 7).

Communication between cells in natural systems occurs in three ways: (i) proteins or small molecules produced and secreted by sender cells diffuse and induce activation of surface receptor proteins or intracellular sensors in recipient cells (Altan-Bonnet and Mukherjee, 2019); (ii) signaling molecules (usually small second messenger molecules) are transported via channel proteins to neighboring recipient cells in direct contact (Hervé and Derangeon, 2013); (iii) membrane ligands on the sending cells and direct interactions between membrane receptors on recipient cells (Kotsias et al., 2019). It is likely that a signal from the sender triggers a transcriptional event in the receiving cell. Regardless, these cell-level circuits would lead to highly complex population behaviors that would not function at the single-cell level. Spatial organization patterns in bacterial and mammalian cells can be formed by typical bandpass circuits or logic gates. Recently, synthetic quorum sensing circuits have been successfully deployed to control cell population size in bacterial and mammalian cells (Ma et al., 2022b).

However, this cellular circuit remains at an early developmental stage. Two major challenges need to be overcome in the future. First, too few signaling molecules are used in



**Figure 7** Synthetic cell and gene circuits for therapeutics. At the single cell level, engineer efforts will focus on three major aspects: (i) sensors that can recognize diseases or environmental signals as biochemical reactions; and then (ii) gene circuits function as the “central processor” to process various input signals, yields (iii) quantitatively-defined output functions to control cell function. As the multicell level, synthetic cytokine secretion or direct ligand-receptor interactions enable various cell-cell communications to form topologically-organized cell circuits or spatially-organized organ-like patterns. These single or multiple engineered living cells may act as the powerful drug platform to cure complex diseases, such as cancer and metabolic diseases.

the current studies. In contrast, hundreds of cytokines and growth factors exist in humans that are involved in multiple regulations in a large number of cell types. Therefore, it is attractive to engineer synthetic cytokines or other factors to construct future cellular circuits. Second, orthogonal pairs of receptors and ligands for direct contact communication are difficult to design. The most successful example is synthetic Notch (synNotch) signaling, which enables the binding of any ligand through the extracellular recognition domain and triggers programmable downstream gene transcription (Morsut et al., 2016). Several demonstrative studies have applied the synNotch system to spatially organized multicellular structures (Toda et al., 2020).

In addition to the technical challenges similar to gene circuits, the principles by which structure determines circuit function is difficult to understand, especially given the increasing complexity of the spatiotemporal regulation of cell populations. For example, how can designing circuits with precisely controlled biostability or multistability make a lot of sense for synthetic cell differentiation? Which circuit topologies enable efficient signal amplification with high fidelity and robustness in the treatment of disease? How do cellular circuits control the size and type of cell populations in the steady state? We envision that these issues require quantitative and comprehensive consideration of fundamental design principles at the level of cellular circuitry.

### **Engineered living cell therapeutics**

Another major trend in cellular and genetic circuit engineering is the extension of proof-of-concepts, currently in “toy” systems, to disease-related clinical applications. Compared with traditional molecular drug forms, live cell drugs have significant advantages as integrated platforms for deploying payload drugs or performing complex functions (e.g., cell lysis, wound healing) that would be intelligently controlled by integrated genes or cellular circuits. By doing so, cellular drugs can show significantly improved disease-fighting efficacy with minimal side effects. For example, reconfigured cytokine signaling pathways can act as cytokine switches to sense and eliminate pro-tumor cytokines and create a pro-immune cytokine microenvironment (Zhang et al., 2021b). In CAR-T cells, protein circuits with logic gates or hypersensitivity functions are deployed to generate more specific recognition of tumor antigens (Hernandez-Lopez et al., 2021; Williams et al., 2020). CAR-T immunotherapy has demonstrated the power of living cells as a drug form, i.e. cell therapy. In another case, optogenetically controlled gene circuits successfully produced intelligent control of steady-state glucose levels in animals’ blood through a closed-loop control strategy (Yu et al., 2022). These striking examples show that synthetic live-cell medicines have ushered in a new revolution in the treatment of intractable diseases.

Due to the importance to human health, the current success of cell therapy is mainly based on the use of immune cells, especially T cells, as the chassis therapeutic cells. More recently, other immune cells, such as natural killer cells, macrophages, have shown considerable potential not only in cancer treatment but also in treating infectious diseases (Fiscaro and Boni, 2022; Klichinsky et al., 2020). While many cells being engineered as drugs are still at proof-of-concept, we envision that once we can design more accurate and functional gene and cellular circuits, engineering at the primary cell level will be easier. Some recent studies have shown substantial improvement at the clinical level. Notably, multicellular systems will provide additional advantages for cell therapy, which can significantly reduce engineering costs by distributing functional circuit modules into different sub-cell types. A group of cells with well-programmed interacting circuits will work as a whole to enable more efficient, safer and less expensive therapeutic functions.

### **Cell-free synthetic biology**

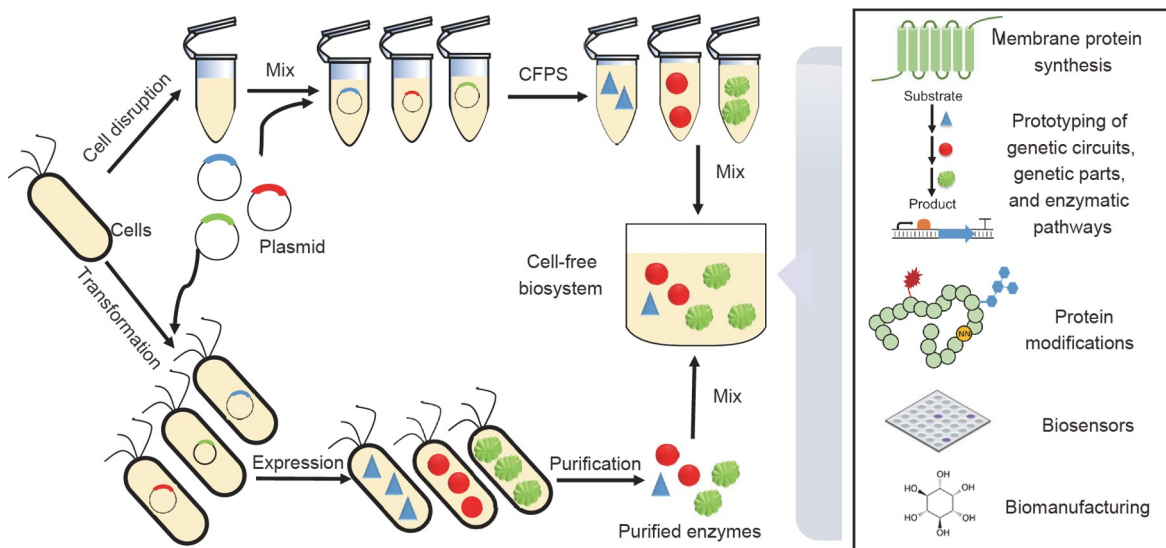
A cell-free synthetic system represents another technical route of synthetic biology in parallel with cell engineering. The target of cell-free synthetic biology is an open system without cell structure, and is focused on the desired meta-

bolic network, using corresponding active components, such as enzymes and coenzymes, to complement complicated biochemistry reactions. Cell-free synthetic biology originated with Eduard Buchner’s paradigm-shifting discovery of “cell-free ethanol fermentation by non-living yeast lysate” (Nobel Chemistry Prize 1907). Another milestone was the discovery of the genetic code and its function in protein synthesis by Nirenberg and Matthaei (Nobel Physiology or Medicine Prize 1968). For the development of cell-free synthetic biology, two types of cell-free systems have been proposed: a cell-extract based system and a purified enzyme-based system (Figure 8). A cell-extract based system has always been used for cell-free protein synthesis (CFPS), to realize the fundamental processes of the central dogma (DNA to RNA, RNA to protein) outside of cells. The purified enzyme-based system consists of numerous purified or partially purified enzymes to implement complicated cascade enzyme reactions, mainly for the biomanufacturing of functional biomolecules and biochemicals. Compared to the system that is carried out inside cells, a cell-free synthetic biosystem features many advantages such as high product yield, fast reaction rate, high engineering flexibility, an accelerated design-build-test-learn cycle, high tolerance to toxic environments, and easy scale-up. These features make cell-free synthetic biology an important enabling technology for many applications.

### **Cell-free biosystem for protein synthesis and applications**

Cell-free biological systems for protein synthesis consist of crude cell extracts, DNA templates, ATP regeneration systems, amino acids, nucleotides, cofactors, and buffers (Silverman et al., 2020). Several cell extracts from *E. coli*, *S. cerevisiae*, wheat germ, rabbit reticulocytes, insect cells, and Chinese hamster ovary cells can be selected according to requirements. The system can be used for the synthesis of toxic or membrane proteins, prototyping of biological functions, protein modification, and biosensors.

It is difficult to overexpress toxic proteins at high yields *in vivo* because toxic proteins may interfere with cellular metabolic pathways and membrane proteins are always expressed in the form of inclusion bodies. Cell-free biological systems for protein synthesis can be used to synthesize toxic proteins such as restriction endonucleases (Goodsell, 2002), cytolethal dilatation toxins (Ceelen et al., 2006), and human microtubule-binding proteins (Betton, 2003) because the *in vitro* system is tolerant of toxic environments. Membrane proteins can be expressed in cell-free biological systems by adding surfactants, liposomes, or nanodiscs (Junge et al., 2011; Matthies et al., 2011; Panganiban et al., 2018; Shelby et al., 2020). Many membrane proteins, such as G protein-coupled receptors (Kaiser et al., 2008; Wang et al., 2011), tetracycline pumps (Wuu and Swartz, 2008), ATP synthases



**Figure 8** Various applications by cell-free synthetic biology system that contained cell-extract based system for cell-free protein synthesis (CFPS) and purified enzyme-based system.

(Matthies et al., 2011), and Hepatitis C virus membrane proteins (Fogeron et al., 2015), are all produced by cell-extract-based cell-free biological systems.

For the prototyping of biological functions, such as genetic components, genetic circuits, and metabolic pathways, cell-free biological systems provide an important platform *in vitro* and allow implementation in cells (Silverman et al., 2020). For a single genetic component (promoter, ribosome binding site, and terminator), a library of variants of linear expression templates can be generated by PCR mutagenesis, and then, with the help of microfluidics, cells containing the single gene variant can be extracted, encapsulated in picoliter droplets (Fallah-Araghi et al., 2012; Zhang et al., 2019b).

In addition to probing individual genetic components, cell-free biological systems can be used to determine how these components work together in synthetic genetic control networks or “circuits” (Siegal-Gaskins et al., 2014). Numerous cell-free genetic circuits have been assembled and prototyped, including cascades driven by the sequential expression of orthogonal polymerases or sigma factors (Garamella et al., 2016; Noireaux et al., 2003; Shin and Noireaux, 2012), as well as feedforward loops and negative autoregulators (Hori et al., 2017; Hu et al., 2018; Takahashi et al., 2015a; Takahashi et al., 2015b). With the goal of engineering cellular metabolism, cell-free biological systems offer enormous possibilities for elucidating these metabolic pathways. It would be of considerable advantage to using cell extracts for protein synthesis in cell-free biological systems in which the expression of enzymes encoding DNA templates can lead to the self-assembly of pathways in a single reaction. To date, several reports have confirmed this approach. For example, two pathways containing three and

six enzymes respectively were re-identified from linearly expressed DNA by cell-free biological systems to produce N-acetylglucosamine and peptidoglycan precursors, respectively (Sheng et al., 2014; Zhou et al., 2010). A five-enzyme pathway that converts tryptophan to purpurin has also been demonstrated (Garamella et al., 2016; Pardee et al., 2016b). Furthermore, a combinatorial strategy was recently used to construct a 17-step enzymatic pathway for n-butanol (Karim and Jewett, 2016). Combined with data-driven design, cell-free biosystems can be used to rapidly evaluate hundreds of pathway combinations in *E. coli* extracts to enhance butanol and 3-hydroxybutyrate production in Gram-positive anaerobic bacteria, demonstrating cell-free and *in vivo* pathway performance (Karim et al., 2020)

For a broad range of protein modifications including glycosylation (Jaroentomechai et al., 2018), phosphorylation (Oza et al., 2015), PEGylation (Shozen et al., 2009) and insertion of unnatural amino acids (uAAs) (Perez et al., 2016), cell-free biological systems offer robust control and versatility, bypassing limitations related to cell-based toxicity and permeability (Jaroentomechai et al., 2018). Studying the modification of intracellular proteins is often challenging because it is difficult to obtain proteins with homogenous modifications within the cell. Cell-free biological systems have been shown to have highly homogeneous protein modification functions. A classic example is the glycosylation at specific sites on proteins. Many therapeutic proteins are highly dependent on efficient and homogeneous glycosylation (Li and d’Anjou, 2009).

Cell-free biological systems using *E. coli* cell extracts represent an ideal test bed for detecting glycosylation, as *E. coli* does not possess a native glycosylation function. Thus,

the ability to accelerate cell-free carbohydrate screening in prokaryotic cells using cell-free techniques could have a transformative impact on the design of glycosylation therapies and vaccines (Valderrama-Rincon et al., 2012; Wacker et al., 2002). Open cell-free biological systems are particularly suitable for the use of orthogonal translation systems consisting of non-native tRNA and aminoacyl tRNA synthetases, adding uAAs at the UAG amber stop codon of mRNAs (Albayrak and Swartz, 2013; Hong et al., 2014; Martin et al., 2018). The incorporation of uAAs into proteins offers unlimited possibilities for the use of modified proteins as therapeutics. Once uAAs are incorporated at precise locations on target proteins, they act as biorthogonal chemical handles that react with functionalized small molecules to generate therapeutic conjugates, such as antibody-drug conjugates (ADCs) (Agarwal and Bertozzi, 2015; Ratner, 2014; Yin et al., 2012).

When we assess the role of cell-free biosystems as biosensors, they provide several practical advantages over whole-cell biosensors. Cell wall-impermeable or cytotoxic analytes can be detected in cell-free biological systems, and they are more reliable because of the potential for mutation and plasmid loss in whole-cell sensors (Silverman et al., 2020). The properties of cell-free biological systems for protein synthesis can be used to host gene circuit-based sensors that can detect nucleic acids and small molecules with extreme sensitivity and specificity (Tinafar et al., 2019). To detect nucleic acids primarily from disease-causing viruses and bacteria, RNA extracted from pathogen-containing samples is added to a cell-free biological system programmed to produce reporter proteins only in the presence of target nucleic acid sequences through a designed toehold switch riboregulator. It could replace reverse transcription PCR (RT-PCR) for more rapid diagnostic testing. Using this strategy, many viruses can be detected quickly, including Ebola (Pardee et al., 2014), Zika (Pardee et al., 2016a), Norwalk virus (Ma et al., 2018a), Cucumber Mosaic virus (Verosloff et al., 2019), SARS-Cov-2 (Hunt et al., 2022; Ma et al., 2022a), and certain gut-colonizing bacteria (Takahashi et al., 2018). This cell-free system for virus detection can be fixed on paper by freeze-dry technology to improve its portability and stability (Hunt et al., 2022), providing an alternative to meet the urgent diagnostic requirement for current Covid-19 and future virus pandemic. Progress in cell-free detection of small molecules (e.g., environmental toxins or cellular metabolites) has been slower than in the detection of nucleic acids, because there are no analogs to synthetic ribose modulation for the construction of sensors for arbitrary small molecules. Most reported cell-free small molecule sensors detect environmental toxins, such as mercury (Salehi et al., 2017) and fluoride (Thavarajah et al., 2019), drugs, such as gamma-hydroxybutyrate (Gräwe et al., 2019), or bacterial quorum-sensing signals

such as N-butyl-L-homoserine lactone (Wen et al., 2017). Studies have shown that cell-free sensors can be freeze-dried and remain active for months even when dried on paper substrates (Pardee et al., 2014), offering an alternative means to address the unmet need for easy distribution and low-cost sensing by cell-free systems (Tinafar et al., 2019).

### ***Cell-free biosystems based on purified enzymes for biomanufacturing***

Cell-free biological systems based on purified enzymes refer to the construction of biocatalytic systems that constitute multiple purified/partially purified enzymes for converting certain substrates to desired compounds through engineered reaction pathways (Wei et al., 2020). Here, we focus on cell-free biological systems for biomanufacturing using sustainable substrates such as starch, glucose, cellulose, and carbon dioxide.

Myo-inositol (hereafter referred to as inositol) and hydrogen are two typical products produced directly from starch by cell-free biological systems. Inositol is widely used in the cosmetic, pharmaceutical and food industries. It is obtained from phytic acid by acid hydrolysis. This method uses expensive raw materials and produces serious phosphorus pollution. Zhang et al. and Atomi et al. both constructed a cell-free biological system containing four enzymatic reactions that can convert starch to inositol with a theoretical product yield of 100% (Fujisawa et al., 2017; You et al., 2017). All enzymes in this biological system are thermophilic, so the enzymes can be easily purified by thermal treatment and high reaction temperature, avoiding microbial contamination. Compared with traditional chemical methods, this new method of producing inositol from starch has great potential for green inositol production. Currently, Bohaoda Biological (China) is building an industrial facility that is scaling up this novel method to produce inositol (You et al., 2017). Many other value-added chemicals, such as glucosamine (Meng et al., 2020), allulose (Li et al., 2021), and (-)-vibo quercitol (Bai et al., 2019), can be synthesized by similar enzymatic treatment of starch. Hydrogen is the transportation fuel of the future, and improved energy efficiency through fuel cells has the potential to reduce greenhouse gas emissions and provide end users with zero pollutants (Armaroli and Balzani, 2010). Natural cellular metabolic pathways can only produce up to 4 moles of H<sub>2</sub> per mole of glucose (Chou et al., 2008; Veit et al., 2008; Wu et al., 2017). Zhang and colleagues conducted a proof-of-concept experiment that produced 12 moles of H<sub>2</sub> per mole of glucose through a cell-free biological system containing 13 purified enzymes. The biological system converts starch into H<sub>2</sub> and CO<sub>2</sub> almost quantitatively with the following total stoichiometry: C<sub>6</sub>H<sub>10</sub>O<sub>5</sub>+7H<sub>2</sub>O=12H<sub>2</sub>+6CO<sub>2</sub>. This biological system can be slightly modified to



develop sugar biobatteries with an energy density that is an order of magnitude higher than lithium-ion batteries (Zhu et al., 2014). This cell-free biological system for hydrogen production lays the foundation for future sugar-hydrogen vehicles.

The production of ethanol, isobutanol, and prenylated natural compounds is described here when glucose is used as a substrate in cell-free biological systems. Ethanol is the most important gasoline additive, and isobutanol is four-carbon liquid alcohol compatible with current internal combustion engines and transportation pipelines (Atsumi et al., 2008; Li et al., 2011a; Welch and Scopes, 1985). Sieber and colleagues designed a cell-free biological system that can produce ethanol and isobutanol from glucose via pyruvate (Guterl et al., 2012). Compared with the 10 enzymes used in the natural glycolytic pathway, this biological system uses only four enzymes to convert glucose to pyruvate, which can be converted to ethanol and isobutanol. This cell-free biosystem produces large amounts of isobutanol even in the presence of 4% (v/v) isobutanol, whereas even low concentrations (e.g., 1%–2% v/v) prevent microbial production of isobutanol (Atsumi et al., 2008). This progress demonstrates that cell-free biological systems are highly tolerant of toxic environments. To produce prenylated natural compounds, Bowie and colleagues designed a cell-free biological system consisting of more than 20 enzymes (Valliere et al., 2019). These enzymes can be divided into 4 main reaction modules: a glycolysis module for the production of pyruvate from glucose; an acetyl-CoA module for the production of acetyl-CoA from pyruvate, and a mevalonate module for the production of geranyl pyrophosphate (GPP) from acetyl-CoA; and a prenylation module for the production of the desired prenylated product. The prenylation module can also be modulated by using alternative enzymes and substrates to produce various prenylated compounds, such as isoprenoids (*repeated word*) and cannabinoids. After system optimization, this cell-free biosystem produced a cannabinoid titer of 1.25 g L<sup>-1</sup>, which was at least two orders of magnitude higher than published results using live cells (Luo et al., 2019).

When cellulose is used as a substrate, a typical example is the production of starch from cellulose by cell-free systems. This biological system contains endoglucanase, cellobiohydrolase, cellobiose phosphorylase, and alpha-glucan phosphorylase for a one-pot enzymatic conversion of pretreated biomass to starch. Up to 30% of the anhydroglucose units in cellulose are converted to starch (You et al., 2013). Because the annual source of cellulose feedstock is ~40 times greater in mass than the starch for food and feed, this cost-effective transformation of non-food cellulose to starch can reshape the bioeconomy and solve the triple dilemma of food, bio-fuels, and the environment (Zhang, 2013).

Researchers from the Tianjin Institute of Industrial Bio-

technology, Chinese Academy of Sciences constructed an artificial starch anabolic pathway (ASAP) to utilize CO<sub>2</sub> and hydrogen in the synthesis of starch (Cai et al., 2021). ASAP is a chemical-biological hybrid system, which includes a chemical system and a cell-free biological system. The chemical system converts CO<sub>2</sub> and hydrogen to methanol (Wang et al., 2017). The cell-free biological system contains 11 core enzymes and three auxiliary enzymes to convert methanol to starch. After optimization of conditions, including modular assembly and substitution and protein engineering of three rate-limiting enzymes, this chemical-biological hybrid system converts carbon dioxide to starch at a rate of 22 nmol min<sup>-1</sup> mg<sup>-1</sup> of total catalyst, which is 8.5 times higher than in a maize substrate system. This approach offers a potential strategy to feed the world, and more importantly, provides a potential solution to the problem of food sources when exploring other planets (Wu and Bornscheuer, 2022).

In conclusion, cell-free synthetic biology offers a game-changing tool to circumvent the limitations inherent in living cells. With a plethora of research across different fields, including gene expression, genetic networks, protein modification, on-demand biosensing, and biomanufacturing using cell-free biological systems, the prospects of cell-free synthetic biology are evident. However, to realize the true potential of cell-free biological systems, several challenges need to be overcome, including the longevity of such biological systems and the regeneration of unstable natural cofactors. After addressing these shortcomings, cell-free synthetic biology will bring biology and biotechnology into a new era with many interesting results. It is exciting to see cell-free synthetic biology combined with other cutting-edge disciplines such as material science, electronics, computer, and artificial intelligence.

## AI and synthetic biology

Obtaining ideal biological components is the basis for building synthetic biological systems. With the recently witnessed increase in computing power, artificial intelligence (AI) has been shown to excel in a variety of challenging tasks such as image generation (Bau et al., 2020), natural language processing (LeCun et al., 2015), and synthetic biology applications (Wang et al., 2020). This section will briefly describe AI-dependent approaches that have shown increasing success in mining complex biological properties and designing optimized synthetic biological components (bioparts), especially gene regulatory sequences. Excellent reviews are available for in-depth discussion on specific applications such as metabolic engineering (Lawson et al., 2021), gene therapy (Huang et al., 2021) and drug discovery (Sanchez-Lengeling and Aspuru-Guzik, 2018).

### A general framework of AI guided bioparts inverse design

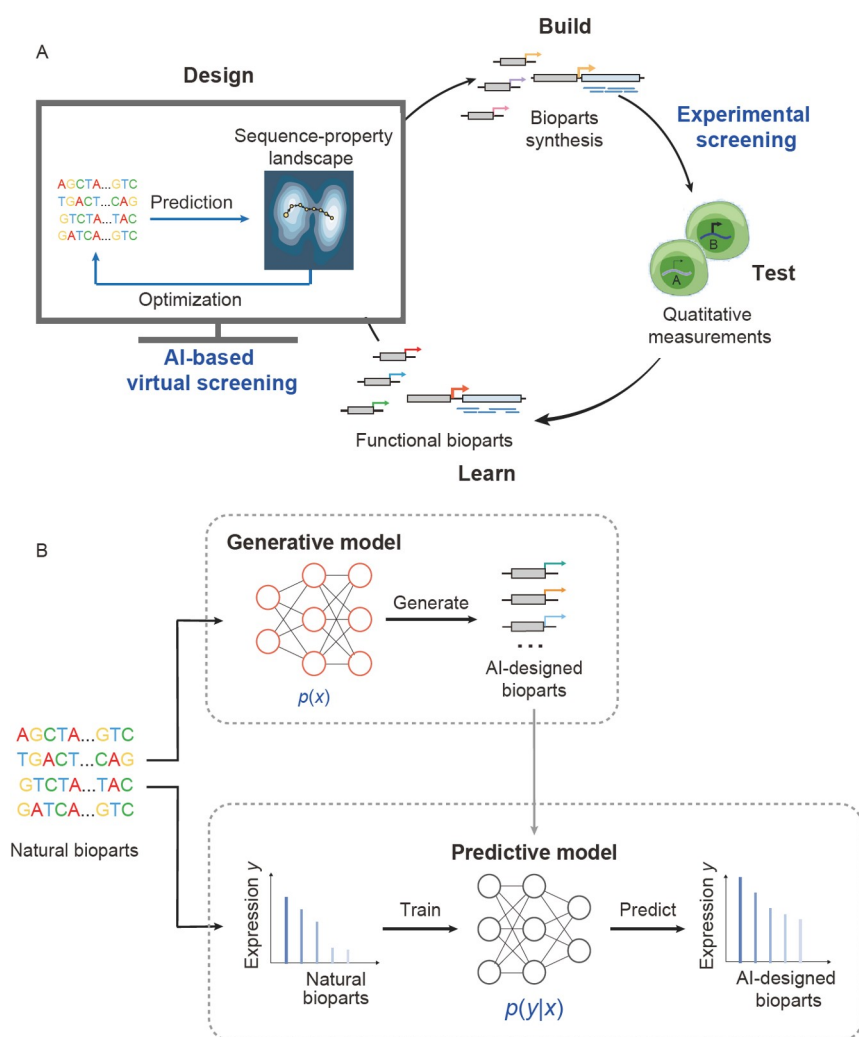
Bioparts design is an important yet complex task, which aims to reverse engineer new biomolecules based on specific target properties. It is experimentally difficult to exhaustively search the potential sequence space to discover new bioparts (e.g., a 100 bp of DNA sequence forms a potential sequence space of  $4^{100}$ ). Therefore, virtual screening offers a promising alternative for exploring this vast space. Armed with a computational model that can estimate the fitness landscape of a sequence space, it is now possible to select candidate designs with high fitness and employ an iterative process to achieve an efficient virtual screen of bioparts (Figure 9A).

From a machine learning perspective, the inverse design problem of bioparts can be abstracted as the mathematical problem of estimating the joint distribution of bioparts functions and sampling the target biopart  $x$  with target function  $y$  from it. Among the target functional  $y$ , the biopart

design problem can be formulated in probabilistic terms as finding mutually compatible sequence-function pairs that maximize the joint probability  $p(x,y)$  (Anishchenko et al., 2021b). Using the probability chain rule, we can get

$$p(x,y) = p(y|x) * p(x), \quad (1)$$

where the first term represents the conditional probability of the function  $y$  for the given sequence  $x$ , and the second term represents the biocompatibility of sequences  $x$ , which is constrained by chemical and biophysical properties. The development of machine learning methods, especially deep learning, has enabled increasingly accurate estimates of adaptive environments (Angermueller et al., 2016) and has considerably improved the efficiency of generating candidate designs that satisfy biological constraints. Combining virtual screening and high-throughput experimental screening in a closed loop facilitates iterative optimization of virtual screening and further accelerates design progress (Figure 9A).



**Figure 9** A, The schematic of Design-Build-Test-Learn (DBTL) process for bioparts design. Bioparts design integrates the AI-guided virtual screening, bioparts synthesis, biological measurements and learning of functional features into a closed-loop framework. B, The framework about the generative models and predictive models.

## Deep learning models for synthetic biology

With the increase in computational power and high-throughput omics data, the use of deep learning has emerged as a promising approach to learn complex patterns and efficiently estimate data distributions implicitly or explicitly. We briefly introduce two main classes of deep learning models widely used in biological component design: predictive models and generative models.

**Predictive model:** To estimate the term  $p(y/x)$ , predictive models were constructed to evaluate the property  $y$  on condition of input bioparts  $x$ . The input to the prediction model is the sequences of bioparts and the output is the predicted properties of those sequences. For example, we provided the model with a promoter sequence and had it predict the expression levels of downstream genes. One of the widely used models is based on convolutional neural networks (CNNs), which can efficiently learn local patterns in sequences, such as transcription factor binding sites and their combinations. For example, Zrimec et al. used a CNN model to predict gene expression of promoter, 5'UTR, 3'UTR and terminator sequences and achieved an R-squared value of 0.822 in predicting *S. cerevisiae* mRNA abundance (Zrimec et al., 2020). Another well-known application of CNN-based models comes from AlphaFold (Senior et al., 2020). In the first stage of protein structure prediction, they used 64 residual convolutional blocks to predict distance and torsion distribution and achieved considerable improvements over a comparative CASP 13 system (AlQuraishi, 2019).

Predictive models based on recurrent neural network (RNN) models and attention-based neural networks are also widely used to capture long-term interactions between different regulatory elements. RNN is an artificial neural network that uses sequence information to extract long-term correlations (Greener et al., 2022). Quang et al. (Quang and Xie, 2016) proposed a hybrid convolutional and recurrent deep neural network to quantify the functionality of DNA sequences. Attention is a technique that mimics cognitive attention, which may be associated with different locations in a single sequence and has achieved impressive results in natural language understanding tasks. The advantage of the attention mechanism is that it can notice regions or patterns of interest regardless of the distance, thus facilitating the understanding of long-range syntax in DNA or protein sequences. The most prominent application of the attention model is AlphaFold2 (Jumper et al., 2021), which significantly outperforms other methods including AlphaFold in protein structure prediction. AlphaFold2 uses an attention mechanism in the predictive model instead of the convolutional layers in AlphaFold. This improvement shows that the attention mechanism has considerable potential for future applications in structure and fitness prediction.

**Generative models:** Generative models aim to understand

the underlying distribution of the data  $p(x)$ . Deep generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), use deep neural networks to implicitly estimate the distribution in the sample space by mapping the sample space to a low-dimensional representation space and generate new samples from it. State-of-the-art generative models have achieved considerable success in generating vivid images in computational vision tasks (Bau et al., 2020). The rationale behind these state-of-the-art performances is that generative models can estimate sample distributions constrained by complex properties, which helps researchers achieve efficient model-guided navigation in the sample space and generate completely new samples that have never been seen before.

Similar methods can also be used to estimate the distribution of functional biomolecular sequences that occupy only a small fraction of the overall sequence space limited by biophysical properties and long-range interactions. Deep generative models can help researchers more efficiently explore candidate sequences that are more likely to be functional.

## Engineering bioparts using AI

AI algorithms have been applied to the design of synthetic biological components, including the design of *cis*-regulatory sequences (Van Brempt et al., 2020), small-molecule drugs (Xu et al., 2019), and small peptides (Cherkasov et al., 2009).

With sufficient numbers of sequence samples ( $x$ ) with functional annotation ( $y$ ), as well as a well-designed AI model, it is now possible to computationally design biological components rationally from scratch. One approach is to combine predictive models with stochastic screening, genetic algorithms or gradient search to virtually screen the fitness landscape to find possible functional biological components. For example, Van Brempt et al. used a training set of over 250,000 synthetic sequences to predict expression, then applied random screening to select *E. coli* sigma 70 promoters with differential expression, and experimentally verify a Spearman's rank correlation factor equal to 0.909 (Van Brempt et al., 2020). Kotopka et al. implemented a genetic algorithm and gradient search optimization to design constitutive and inducible yeast promoters for high expression (Kotopka and Smolke, 2020). Bogard et al. designed surrogate polyadenylation sites through gradient ascent optimization based on a predictive model that maps the position weight matrices (PWM) inputs of surrogate polyadenylation sites to isoform numbers (Bogard et al., 2019). Bryant et al. designed highly diverse adeno-associated virus type 2 (AAV2) capsid protein variants by gradually moving away from natural AAV serotype sequences guided by predictive models (Bryant et al., 2021).

In addition to obtaining better predictive models, another important direction is to obtain better estimates of sample distribution in order to more efficiently generate new biological component candidates. Taking high-expression promoter sequence design as an example (Figure 9B), Wang et al. used generative adversarial networks to survey more than 10,000 natural sequences and applied *de novo* promoter design in *E. coli*. The generative model first learns the promoter distribution from the natural promoter dataset. The generator detects key regulatory patterns, such as transcription factor binding sites (TFBSs), and can generate new samples that match sequence signatures. A predictive model is then trained to estimate the properties of the samples. Combining generative and predictive models, a synthetic high-expression promoter candidate set can be obtained by virtual screening. Finally, an iterative process including artificial intelligence-based virtual screening and experimental validation (Figure 9A) helps the model to efficiently learn the functions of the biological components. As a result, up to 70.8% of AI-designed promoters were *experimentally* demonstrated to be functional, few of which exhibited significant sequence similarity to the *E. coli* genome (Wang et al., 2020). In another example, Repecka et al. proposed a variant of self-attention-based generative adversarial network to learn natural protein sequence diversity, with 24% of the generated sequences being functional including a highly mutated variant of 106 amino-acid substitutions (Repecka et al., 2021). Shin et al. also introduced a deep generative model, which successfully designed and tested a diverse  $10^5$  nanobody library, namely by exploring new sequence spaces (Shin et al., 2021). Biswas et al. applied a deep generative model with long-short memory to capture uniform protein sequence distributions and computationally explore protein landscapes in the range of  $10^7$ – $10^8$  variants. They successfully constructed the fluorescent protein avGFP and the enzyme TEM-1 $\beta$ -lactamase from wild-type sequences, serving as training data in only 24 or 96 characterized sequence variants (Biswas et al., 2021).

With the computational power to process complex data, AI has been successfully applied to a variety of synthetic biology problems and has shown unprecedented efficiency, accelerating inverse design tasks several orders of magnitude faster than traditional experimental methods. AI-based approaches, such as deep predictive models, deep generative models, and reinforcement learning methods, have brought significant improvements in synthetic regulatory sequence design and drug discovery (Sanchez-Lengeling and Aspuru-Guzik, 2018; Faulon and Faure, 2021). Integrating AI models into a closed-loop bioprocess optimization framework will considerably speed up the comprehensive design process.

Despite these recent advances, the application of artificial intelligence in synthetic biology is still in its infancy. One

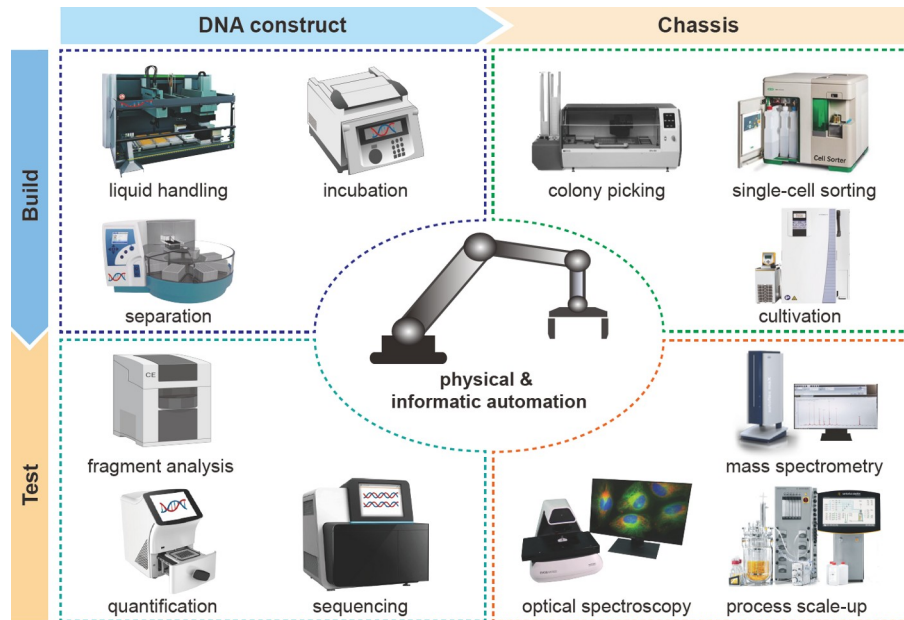
major reason is that the capabilities of existing AI methods are limited by the size of training samples. Compared to computer vision and natural language processing tasks, which often have millions or even billions of training samples, the sample size of biological data is too small to fully unleash the power of these deep learning models. Therefore, it is important for the community to provide standardized samples with sequence to functional annotation pairs to better train AI models. Another point is that the state-of-the-art AI frameworks, such as convolutional neural networks and attention-based models, originate from non-biological domains, such as computer vision or natural language processing tasks. It is now critical to develop new AI frameworks that better incorporate knowledge from the biological domain. For example, the interplay between synthetic gene circuits and complex multilevel regulation within cells remains to be studied. In addition, functional bioparts in cells usually show dynamic changes in time series, so how to measure and capture the dynamic distribution of bioparts is also an important problem to be solved. Overcoming these challenges will bring enormous potential opportunities for synthetic biology and artificial intelligence in the near future.

## **Biofoundries—process automation in synthetic biology**

### *DBTL automation using biofoundries*

As discussed in the first two sections, due to the lack of predictive modeling, a trial-and-error process is often employed to create biological systems with desirable properties. Through physical and information automation (Figure 10), biofoundries promise to implement and accelerate the design-build-test-learn (DBTL) cycle as an engineering framework for synthetic biology (Chao et al., 2017b; Hillson et al., 2019). The integration of computer-aided design (CAD), robotics, and high-throughput instrumentation allows efficient exploration of genetic and process variables and rapid data generation to recommend experimental plans for the next DBTL iteration using active learning algorithms (HamediRad et al., 2019; Zhang et al., 2020a; Zhang et al., 2021a). Additionally, standardization of materials, hardware, protocols, and data reporting eliminates idiosyncratic biases and errors (Beal et al., 2020). The enhanced reproducibility thus enables the collective aggregation of big data across batches, projects and institutions to obtain mechanistic and statistical models of engineering biology (Farzaneh and Freemont, 2021; Hillson et al., 2019). Currently, public institutions and private companies around the world are building biofoundries (Table 1). Numerous robotic workflows have been developed for the automated construction and testing of biologically based synthetic genetic constructs and organisms, with a primary focus on microbial cell fac-





**Figure 10** Biofoundries provide an integrated infrastructure to automate DBTL loops of synthetic biology via physical and informatic automation. Build: programmed by robotic scripts, automation instruments construct gene cassettes, pathways/circuits, and synthetic genomes, followed by cell culture, genetic transformation, and clonal selection. Test: test automation permits rapid and large-scale genotype-phenotype mapping across scales with findable, accessible, interoperable, and reusable (FAIR) metadata for DBTL iterations using adaptive learning algorithms.

ories for chemical/biochemical production (Hillson et al., 2019; Zhang et al., 2021a).

#### *Build automation: DNA construction*

In biofoundries, robotic protocols can be applied to large-scale fabrication of expression cassettes, metabolic pathways, genetic circuits, and even entire genomes, using synthetic or cloned DNA fragments (Zhang et al., 2021a). Type II restriction and homology-directed methods are widely used for automation-compatible DNA assembly, primarily because they allow one-step, scar-free assembly of many fragments using standard procedures (Chao et al., 2017a; Dharmadi et al., 2014; Kanigowska et al., 2016; Walsh III et al., 2019). For the Type II restriction method, the Golden Gate method assembles up to 15 fragments in one step on an academic biofoundry called iBioFAB, which can create 400 structures per day (Chao et al., 2017a). For the homology-directed method, up to 12 DNA fragments can be assembled in *S. cerevisiae* using transformation-associated recombination (TAR), achieving a throughput of over 1500 constructs per (Ip et al., 2020). However, DNA assembly is not error-free, and robotic screening is necessary for the rapid and large-scale identification of correctly assembled structures. For example, structural analysis can automate qPCR analysis for detection of assembled junctions (Shapland et al., 2015), and capillary electrophoresis (CE) analysis for matching restriction patterns (Dharmadi et al., 2014). Furthermore, with the help of multiplex DNA barcodes introduced during robotic NGS library preparation, next-generation sequencing

(NGS) validation can be designed to analyze hundreds to thousands of assembled structures in a single run (Shapland et al., 2015; Suckling et al., 2019).

#### *Build automation: engineered organisms*

Basic procedures for organism engineering using robotic systems include cell culture, genetic transformation, and clonal selection (Hamedirad et al., 2019; Ip et al., 2020; Rajakumar et al., 2019; Si et al., 2017a). For these procedures, commercial instrumentation is available and can be easily integrated to automate the handling of aerobic model microorganisms, as described above, when *E. coli* and *S. cerevisiae* are used as recombinant hosts for robotic DNA assembly (Chao et al., 2017a; Ip et al., 2020; Kanigowska et al., 2016). For non-model organisms, specially designed equipment and laboratories are required to support automated workflows. For example, a casting platform is completely enclosed in an environmental control room for robotic manipulation of strictly anaerobic bacteria. Furthermore, although bulk electroporation can be performed in a 96-well format (Park et al., 2011; Wang and Church, 2011), genetic transformation of certain microalgae requires single-cell electroporation in microfluidics (Im et al., 2015). In addition, customization is necessary to modify commercial colony pickers to purify homokaryotic transformants and pick mycelia of multinucleated filamentous fungi (SunSpiral et al., 2022). For organisms that are challenging for robotic protocols, the automation of corresponding cell-free systems offers a viable alternative, as demonstrated by rapid proto-

**Table 1** A list of selected biofoundries and automation capacities

Biofoundry	Build automation-DNA	Build automation-Chassis	Test automation	DBTL integration	Key references
Shenzhen Biofoundry, China	Golden Gate & Gibson 384 per day	<i>E. coli</i> , <i>S. cerevisiae</i> , anaerobic microbe, <i>Streptomyces</i> , phage	Optical spectroscopy, MS, microfluidics, process scale-up	–	(Tang et al., 2021)
Tianjin Biofoundry, China	Golden Gate >96 per day	<i>E. coli</i> , <i>Streptomyces</i> , <i>Corynebacterium glutamicum</i>	Optical spectroscopy, microfluidics	–	(Tu et al., 2021; Wang et al., 2018)
Agile Biofoundry, US	Golden Gate & Gibson >96 per day	<i>E. coli</i> , <i>S. cerevisiae</i> , <i>Pseudomonas</i> , <i>Rhodospiridium</i>	Optical spectroscopy, MS, microfluidics, process scale-up	ART (Automated Recommendation Tool), DIVA bioCAD	(de Rond et al., 2019; Hillson et al., 2012; Oberortner et al., 2020; Radivojević et al., 2020)
iBioFAB, US	Golden Gate & yeast recombination 400 per day	<i>E. coli</i> , <i>S. cerevisiae</i>	Optical spectroscopy, MS, process scale-up	BioAutomata	(Chao et al., 2017a; Hamedirad et al., 2019; Si et al., 2017a; Si et al., 2017b)
London Biofoundry, UK	BASIC >96 per day	<i>E. coli</i> , <i>S. cerevisiae</i> , cell-free	Optical spectroscopy, MS, SARS-CoV-2 clinical diagnostics	Design of experiments	(Crone et al., 2020; Exley et al., 2019; Gowers et al., 2020)
Earlham DNA Foundry, UK	Golden Gate >96 per day	<i>E. coli</i> , cell-free ( <i>E. coli</i> , plant)	Optical spectroscopy	–	(Dudley et al., 2021)
SYNBIOCHEM, UK	Ligase cycling reaction >96 per day	<i>E. coli</i> , cell-free ( <i>E. coli</i> )	Optical spectroscopy	RetroPath, Design of Experiments	(Borkowski et al., 2020; Robinson et al., 2020)
Edinburgh Genome Foundry, UK	Golden Gate & Gibson 2000 per week	<i>E. coli</i> , <i>S. cerevisiae</i>	Optical spectroscopy	CUBA (Collection of Useful Biological Apps)	(Kanigowska et al., 2016; Martella et al., 2017)
SynCTI Biofoundry, Singapore	N.A.	<i>E. coli</i> , <i>S. cerevisiae</i>	Optical spectroscopy, process scale-up	Model-driven bioprocess optimization	(Yeoh et al., 2021)
Amyris Inc., US	Yeast recombination >1500 per week	<i>E. coli</i> , <i>S. cerevisiae</i>	Optical spectroscopy, MS, process scale-up	–	(Ip et al., 2020; Rienzo et al., 2021)

typing of the *Bacillus megaterium* promoter at the London Biofoundry (Moore et al., 2018b).

#### Test automation beyond optical assays

Test automation not only allows for rapid design validation in DBTL cycles, but also allows for large-scale genotype-phenotype mapping in the absence of rational design rules. Spectroscopy is a common test method in microtiter plate format but is usually limited to optically active analytes (Hamedirad et al., 2019) and growth-coupled phenotypes (Si et al., 2017a). Otherwise, chemical/enzymatic conversion or transcriptional biosensors are necessary to link target phenotypes to spectroscopic signals, but assay development often requires time and labor (Rienzo et al., 2021). To expand testing capabilities, traditionally slow but informative modalities of analysis are increasingly integrated into biofoundries (Rienzo et al., 2021; Zhang et al., 2021a). For example, mass spectrometry (MS) provides label-free analysis with superior molecular specificity and coverage; while well suited for metabolic analysis of cell factories, it is often limited by time-consuming chromatographic separations (5–60 min). RapidFire MS, Echo MS, MALDI-MS systems combined with automated sample preparation eliminate chromatography steps, thus enabling rapid chemical analysis of cell cultures in 1–10 s per sample (Fu et al., 2020; Si et al., 2021). Robotic creation of microbial libraries followed by

high-throughput MS analysis has been successfully used to identify desirable enzymes and strain variants on a biological basis (Gowers et al., 2019; Si et al., 2017b; Xue et al., 2020).

#### DBTL automation: recent examples

In addition to isolated robotic steps in build and test, recently built facilities have demonstrated the progress made in fully automating the DBTL cycle. For example, the BioAutomata platform was built by integrating the iBioFAB foundry, Bayesian algorithms and predictive models (Hamedirad et al., 2019) to investigate the autonomous exploration of variable spaces in promoter strength to optimize biosynthetic flux for lycopene production without investigator intervention. In three automated DBTL loops, BioAutomata outperformed random search by 77% after evaluating only <1% of all possible variants (Hamedirad et al., 2019). Similarly, the Automated Recommendation Tool (ART) developed by the Agile BioFoundry combines machine learning and Bayesian ensemble methods to predict production levels and guide experiments for subsequent DBTL iterations. When combined with a constraint-based mechanistic genome-scale model, ART successfully increased tryptophan production in *S. cerevisiae* (Zhang et al., 2020a). In addition, at the UK Synthetic Biochemistry Foundry, an iterative DBTL cycle of reverse biosynthetic design, enzyme screening and pathway optimization was guided by a Design of Experiments (DoE) ap-

proach and rapidly created production of 17 material monomers within 85 days in *E. coli* strains (Robinson et al., 2020).

### **Challenges and future developments for biofoundries**

Despite these proof-of-concept successes, three major challenges need to be addressed to realize the full potential of the biofoundries, namely human-machine adaptation, application scalability, and information automation for autonomous discovery.

#### *Human-robotics synergy*

Firstly, existing laboratory software, equipment, and protocols in synthetic biology are often oriented toward humans and translating them to robotic execution is not trivial (Engineering Biology Research Consortium, 2019). For example, certain manual steps are difficult to automate (e.g., removing the alcohol supernatant as completely as possible during DNA precipitation); custom hardware (i.e., microfluidics) is not compatible with commercial robotics and related software. However, the learning curve for bench biologists to develop robotic workflows directly using machine-related programming languages can be steep due to the lack of high-level protocol compilers for biofoundries. To facilitate bidirectional adaptation between humans and machines, it is necessary to develop modular and layered software packages with appropriate abstraction and formalism (Wilbraham et al., 2021). By this means, scientists, engineers, programmers, and system integrators can work independently in their respective areas of expertise while collaboratively developing molecular tools, hardware, software, processes, and applications.

#### *Physical automation: flexibility and versatility*

Secondly, the early implementation of biofoundries only addressed a series of core needs of synthetic biology, with areas of limited application. For example, most current workflows focus on DNA assembly or engineering *E. coli* and *S. cerevisiae* as chemical production hosts (Zhang et al., 2021a). Other highly sought-after applications, such as biosensing and bioremediation, are not well supported by biofoundries. New applications and related technologies require constant customization and reconfiguration of software, hardware, and workflows. Conceptually, process flexibility can be improved by decoupling unit operations, allowing new processes to be incorporated without significant impact on upstream or downstream steps (Engineering Biology Research Consortium, 2019). However, the current biofoundry test models are mostly limited to industry-standard microplates. As phenotype is influenced by both genetic design and environmental constraints, it is likely that test results in microplates will not translate well to practical deployment (Rienzo et al., 2021). For example, deep-well plates in vibrating incubators

cannot replicate the dynamically regulated environment in feedback bioreactors. In this case, hardware integration must go beyond robotic sample transport between component devices. For example, by integrating microfluidic sensors and actuators, BioLector systems enable real-time monitoring and feedback control of pH, dissolved oxygen tension, and optical density in microplate culture models (Funke et al., 2010).

#### *Information automation through standardization and community efforts*

Thirdly, information automation is as important as physical automation, so that experimental results and process metadata can be generated in a findable, accessible, interoperable, and reusable (FAIR) manner (Engineering Biology Research Consortium, 2019). In addition, unified laboratory software, robotics, and integration software in biofoundries will help develop protocols and data standards to enhance reproducibility and interoperability (Beal et al., 2020). Concerted development of standards is indeed an essential task of the recently established Global Biofund Alliance (GBA) (Hillson et al., 2019). Using such an approach, automation and standardization in building and testing not only generates high-quality large-scale datasets, but also enable sophisticated experimental designs for iterative feedback between machine learning models and synthetic biology validation (Carbonell et al., 2019). We envision that, despite the lack of understanding of mechanistic principles, the predictive power of data-driven approaches will address challenges in the design of biological systems.

### **Summary**

Through physical and information automation, synthetic biology foundries around the world can envision large-scale production of synthetic DNA structures and engineered organisms on an unprecedented scale. However, the ultimate goal of basic biological research is not to automate the black-box optimization of a specific biological system one at a time, but to accelerate the scientific discovery of general theories and design rules of engineering biology (Beal and Rogers, 2020). Through continuous engineering and innovation, a tight coupling between hypothesis-generating AI agents and automated experiments can be achieved, leading to the development of an autonomously driven foundry for predictive design of biology (Coley et al., 2020a; Coley et al., 2020b; Kitano, 2021).

### **Conclusions and future prospects**

The ability to synthesize DNA at scale opens the door to writing genomes and represents a powerful approach to solving fundamental biological problems enabling the crea-

tion of engineered organisms with useful properties (Boeke et al., 2016; de la Torre, 2020). The cost of sequencing the human genome has decreased a million fold and its time required has been reduced from 15 years to days or even hours. Within the same timeframe, the cost of DNA synthesis has only declined linearly. High-throughput DNA synthesis and assembly technologies will considerably accelerate genome engineering. Importantly, computer-aided protein structure prediction (Baek et al., 2021; Huang et al., 2022); Senior et al., 2020) will allow scientists to design genomes with new functions from scratch.

DNA storage, an ingenious combination of DNA synthesis and DNA sequencing technologies in information technology applications, is a truly state-of-the-art technology. However, for the same reasons mentioned above, the cost of writing and reading is exceedingly high. Driven by the rigid demand for exponential growth in data storage, these issues remain to be urgently resolved.

CRISPR/Cas gene editing has become a routine technique in synthetic biology. In addition to DNA editing, RNA editing can avoid permanent changes in the genome and reduce the risk of off-target mutations. Single-base editing systems can be used for gene therapy of various diseases caused by single-base mutations. Considering that synthetic biology usually requires complex gene manipulation, a more accurate, efficient and multi-target genome editing system is imperative.

Protein-directed molecular evolution by random mutation has achieved great success in enzyme engineering. With the rapid development of bioinformatics, rational design based on in-depth understanding of protein structure and function is becoming more and more realistic. Recent algorithmic breakthroughs in protein structure prediction may drive protein design from fixed backbone design to *de novo* design. Just imagine that a protein of 200 amino acid residues in length has  $20^{200}$  possible sequences, far exceeding the number of sequences formed by natural evolution over billions of years (Huang et al., 2016); the chances of designing a new protein would be unlimited.

Expanding the genetic codon with unnatural amino acids and molecular machinery responsible for replication, transcription, and translation not only opens a window for re-considering life forms on Earth, but also provides an unprecedented means to create semi-synthetic organisms (Malyshev et al., 2014; Zhang et al., 2017b). For example, synthetic viruses containing unnatural base pairs can be produced in large quantities by adding unnatural amino acids to cell culture. However, in the absence of unnatural amino acids synthetic viruses are unable to replicate inside the human body, while retaining full immunogenicity (Si et al., 2016). This strategy ensures safety and effectiveness of vaccine development, and its protocol principle seems potentially universal. However, natural cells do not readily

accept unnatural codes. To improve orthogonality with the host cell's own genetic system, it is necessary to establish an unnatural system that does not interfere with the natural cellular system. Another new concept is mirror biology. The mirror-image synthetic biology systems are composed of D-amino acids (Weidmann et al., 2019) and L-DNA (Fan et al., 2021). Similarly, the mirror biomacromolecules resist degradation by natural enzymatic systems and avoid triggering immune responses, which may form a powerful biorthogonal system that functions in living cells.

Cell engineering is a synthesis of various techniques in synthetic biology. Chassis cells are an attractive concept, assuming genetic circuits and modules are "plug and play" on the chassis. Due to a clearer genetic background, current chassis cells are based on model microbial genome optimization and minimization, or created by partially rewriting genomes. However, so far, there is no truly universal chassis cell. The adaptability of various gene circuit plug-ins or the robustness of chassis cells often requires extensive trial-and-error experiments or computer-aided rational design. In comparison, plant- and animal-derived chassis cells are still under development due to the complexity of their genomes and intracellular environments, but already have some exciting applications (Li et al., 2019b). In addition, artificial coculture systems containing designed cell-to-cell communication and co-metabolism and cell-free systems containing sequence enzyme catalysis modules are alternatives to the mono-culture biomanufacturing. Organ engineering, including organ-on-a-chip, 3D-printed organs, and organoid engineering, is another multicellular, multi-layered task designed to overcome the difficulties of using the human body for medical experiments, providing *in vitro* means for physiology, pathology, regenerative medicine research and drug development (Wang et al., 2021b; Zhang et al., 2018).

Artificial intelligence is playing an increasingly important role in synthetic biology. Its impact spans almost everything from *de novo* protein design, chassis cell design to bio-foundry operations. A recent example is computer-aided design of antibody nanocages, in which up to 30 antibodies are precisely assembled into these structures (Divine et al., 2021). The considerable amount of genetic information carried by DNA molecules enables a protein structure with infinite diversity, which determines that our living system is a giant complex system. The combination of biology and artificial intelligence can help us efficiently explore this system. This, in turn, will accelerate the development of quantitative synthetic biology, thereby improving the rational design capabilities and predictability of synthetic biology, supporting the realization of the "function emergence".

**Compliance and ethics** The authors declare that they have no conflict of interest.



**Acknowledgements** This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29050100, XDB29050500, XDA24020102) to X.E. Zhang, C. Liu and C. Gao, respectively; the National Natural Science Foundation of China (31725002, 31861143017, 32022044, 62050152 and 32071428) to J. Dai, Y. Yuan, C. You, and X. Wang, respectively; the National Key Research and Development Program of China (2020YFA0907700, 2018YFA0901600, 2019YFA09004500) to Y. Feng and P. Wei. We thank Ms Min Li of the Institute of Biophysics, Chinese Academy of Sciences for her assistance in preparation of the manuscript.

## References

- Abudayyeh, O.O., Gootenberg, J.S., Essletzbichler, P., Han, S., Joung, J., Belanto, J.J., Verdine, V., Cox, D.B.T., Kellner, M.J., Regev, A., et al. (2017). RNA targeting with CRISPR-Cas13. *Nature* 550, 280–284.
- Abudayyeh, O.O., Gootenberg, J.S., Franklin, B., Koob, J., Kellner, M.J., Ladha, A., Joung, J., Kirchgatterer, P., Cox, D.B.T., and Zhang, F. (2019). A cytosine deaminase for programmable single-base RNA editing. *Science* 365, 382–386.
- Acharya, N., Kumar, P., and Varshney, U. (2003). Complexes of the uracil-DNA glycosylase inhibitor protein, Ugi, with *Mycobacterium smegmatis* and *Mycobacterium tuberculosis* uracil-DNA glycosylases. *Microbiology* 149, 1647–1658.
- Aderem, A. (2005). Systems biology: its practice and challenges. *Cell* 121, 511–513.
- Afriat-Jurnou, L., Jackson, C.J., and Tawfik, D.S. (2012). Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry* 51, 6047–6055.
- Agarwal, K.L., Büchi, H., Caruthers, M.H., Gupta, N., Khorana, H.G., Kleppe, K., Kumar, A., Ohtsuka, E., Rajbhandary, U.L., Van de Sande, J.H., et al. (1970). Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* 227, 27–34.
- Agarwal, P., and Bertozzi, C.R. (2015). Site-specific antibody-drug conjugates: the nexus bioorthogonal chemistry, protein engineering, and drug development. *Bioconjug Chem* 26, 176–192.
- Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.C., Marquez, M., Klibanov, A.M., Griffiths, A.D., and Weitz, D.A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc Natl Acad Sci USA* 107, 4004–4009.
- Aharoni, A., Thieme, K., Chiu, C.P.C., Buchini, S., Lairson, L.L., Chen, H., Strynadka, N.C.J., Wakarchuk, W.W., and Withers, S.G. (2006). High-throughput screening methodology for the directed evolution of glycosyltransferases. *Nat Methods* 3, 609–614.
- AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics* 35, 4862–4865.
- Albayrak, C., and Swartz, J.R. (2013). Cell-free co-production of an orthogonal transfer RNA activates efficient site-specific non-natural amino acid incorporation. *Nucleic Acids Res* 41, 5949–5963.
- Altan-Bonnet, G., and Mukherjee, R. (2019). Cytokine-mediated communication: a quantitative appraisal of immune complexity. *Nat Rev Immunol* 19, 205–217.
- Amaral, M.M.F., Frigotto, L., and Hine, A.V. (2017). Beyond the natural proteome: nondegenerate saturation mutagenesis—methodologies and advantages. In: *Proteomics in Biology, Part A. Methods in Enzymology*. Volume 585. 111–133.
- Anavy, L., Vaknin, I., Atar, O., Amit, R., and Yakhini, Z. (2019). Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat Biotechnol* 37, 1229–1236.
- Anderson, J.C., Dueber, J.E., Leguia, M., Wu, G.C., Goler, J.A., Arkin, A. P., and Keasling, J.D. (2010). BglBricks: a flexible standard for biological part assembly. *J Biol Eng* 4, 1.
- Anderson, J.C., Voigt, C.A., and Arkin, A.P. (2007). Environmental signal integration by a modular AND gate. *Mol Syst Biol* 3, 133.
- Angermueller, C., Pärnamäa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol Syst Biol* 12, 878.
- Anishchenko, I., Baek, M., Park, H., Hiranuma, N., Kim, D.E., Dauparas, J., Mansoor, S., Humphreys, I.R., and Baker, D. (2021a). Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP<sub>14</sub>. *Proteins* 89, 1722–1733.
- Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S., Hao, J., Bafna, K., Norm, C., Kang, A., Bera, A.K., et al. (2021b). *De novo* protein design by deep network hallucination. *Nature* 600, 547–552.
- Versai, A. (2019). Data startup Catalog Stores 16 GB of Wikipedia Text onto DNA strands. Available online at <https://www.technowize.com/data-startup-catalog-stores-16gb-of-wikipedia-text-onto-dna-strands/>.
- Annaluru, N., Muller, H., Mitchell, L.A., Ramalingam, S., Stracquadanio, G., Richardson, S.M., Dymond, J.S., Kuang, Z., Scheifele, L.Z., Cooper, E.M., et al. (2014). Total synthesis of a functional designer eukaryotic chromosome. *Science* 344, 55–58.
- Antkowiak, P.L., Lietard, J., Darestani, M.Z., Somoza, M.M., Stark, W.J., Heckel, R., and Grass, R.N. (2020). Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nat Commun* 11, 5345.
- Anzalone, A.V., Gao, X.D., Podracky, C.J., Nelson, A.T., Koblan, L.W., Raguram, A., Levy, J.M., Mercer, J.A.M., and Liu, D.R. (2021). Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat Biotechnol* 40, 731–740.
- Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157.
- Armaroli, N., and Balzani, V. (2010). The Hydrogen Issue. *ChemSusChem* 4, 21–36.
- Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38, W529–W533.
- Atsumi, S., Hanai, T., and Liao, J.C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 451, 86–89.
- Au, L.C., Yang, F.Y., Yang, W.J., Lo, S.H., and Kao, C.F. (1998). Gene synthesis by a LCR-based approach: high-level production of leptin-L54 using synthetic gene in *Escherichia coli*. *Biochem Biophys Res Commun* 248, 200–203.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876.
- Bai, X., Meng, D., Wei, X., Zhou, X., Lu, F., and You, C. (2019). Facile synthesis of (–)-vibo-queritol from maltodextrin via an in vitro synthetic enzymatic biosystem. *Biotechnol Bioeng* 116, 2710–2719.
- Bancroft, C., Bowler, T., Bloom, B., and Clelland, C.T. (2001). Long-term storage of information in DNA. *Science* 293, 1763–1765.
- Banal, J.L., Shepherd, T.R., Berleant, J., Huang, H., Reyes, M., Ackerman, C.M., Blainey, P.C., and Bathe, M. (2021). Random access DNA memory using Boolean search in an archival file storage system. *Nat Mater* 20, 1272–1280.
- Bang, D., and Church, G.M. (2008). Gene synthesis by circular assembly amplification. *Nat Methods* 5, 37–39.
- Bano, S., Wu, X., and Zhang, X. (2021). Towards sustainable agriculture: rhizosphere microbiome engineering. *Appl Microbiol Biotechnol* 105, 7141–7160.
- Barthel, S., Palluk, S., Hillson, N.J., Keasling, J.D., and Arlow, D.H. (2020). Enhancing terminal deoxynucleotidyl transferase activity on substrates with 3' terminal structures for enzymatic *de novo* DNA synthesis. *Genes* 11, 102.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712.
- Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H., and Weiss, R. (2005).

- A synthetic multicellular system for programmed pattern formation. *Nature* 434, 1130–1134.
- Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proc Natl Acad Sci USA* 117, 30071–30078.
- Beaucage, S.L., and Caruthers, M.H. (1981). Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett* 22, 1859–1862.
- Bee, C., Chen, Y.J., Queen, M., Ward, D., Liu, X., Organick, L., Seelig, G., Strauss, K., and Ceze, L. (2021). Molecular-level similarity search brings computing to DNA data storage. *Nat Commun* 12, 4764.
- Beal, J., Goñi-Moreno, A., Myers, C., Hecht, A., de Vicente, M.D.C., Parco, M., Schmidt, M., Timmis, K., Baldwin, G., Friedrichs, S., et al. (2020). The long journey towards standards for engineering biosystems. *EMBO Rep* 21, e50521.
- Beal, J., and Rogers, M. (2020). Levels of autonomy in synthetic biology engineering. *Mol Syst Biol* 16, e10019.
- Betton, J.M. (2003). Rapid translation system (RTS): a promising alternative for recombinant protein production. *Curr Protein Pept Sci* 4, 73–80.
- Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., and Church, G.M. (2021). Low-N protein engineering with data-efficient deep learning. *Nat Methods* 18, 389–396.
- Boeke, J.D., Church, G., Hessel, A., Kelley, N.J., Arkin, A., Cai, Y., Carlson, R., Chakravarti, A., Cornish, V.W., Holt, L., et al. (2016). The genome project-write. *Science* 353, 126–127.
- Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A deep neural network for predicting and engineering alternative polyadenylation. *Cell* 178, 91–106.e23.
- Bollum, F.J. (1962). Oligodeoxyribonucleotide-primed reactions catalyzed by calf thymus polymerase. *J Biol Chem* 237, 1945–1949.
- Borkowski, O., Koch, M., Zettor, A., Pandi, A., Batista, A.C., Soudier, P., and Faulon, J.L. (2020). Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun* 11, 1872.
- Bornscheuer, U.T., Altenbuchner, J., and Meyer, H.H. (1999). Directed evolution of an esterase: screening of enzyme libraries based on pH-indicators and a growth assay. *Bioorg Medicinal Chem* 7, 2169–2173.
- Borovkov, A.Y., Loskutov, A.V., Robida, M.D., Day, K.M., Cano, J.A., Le Olson, T., Patel, H., Brown, K., Hunter, P.D., and Sykes, K.F. (2010). High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Res* 38, e180.
- Brands, S., Brass, H.U.C., Klein, A.S., Pietruszka, J., Ruff, A.J., and Schwaneberg, U. (2020). A colourimetric high-throughput screening system for directed evolution of prodigiosin ligase pigc. *Chem Commun* 56, 8631–8634.
- Brini, E., Simmerling, C., and Dill, K. (2020). Protein storytelling through physics. *Science* 370, eaaz3041.
- Brower, K.K., Carswell-Crumpton, C., Klemm, S., Cruz, B., Kim, G., Calhoun, S.G.K., Nichols, L., and Fordyce, P.M. (2020). Double emulsion flow cytometry with high-throughput single droplet isolation and nucleic acid recovery. *Lab Chip* 20, 2062–2074.
- Bryant, D.H., Bashir, A., Sinai, S., Jain, N.K., Ogden, P.J., Riley, P.F., Church, G.M., Colwell, L.J., and Kelsic, E.D. (2021). Deep diversification of an AAV capsid protein by machine learning. *Nat Biotechnol* 39, 691–696.
- Budin, I., and Szostak, J.W. (2011). Physical effects underlying the transition from primitive to modern cell membranes. *Proc Natl Acad Sci USA* 108, 5249–5254.
- Cadet, F., Fontaine, N., Li, G., Sanchis, J., Ng Fuk Chong, M., Pandjaitan, R., Vetrivel, I., Offmann, B., and Reetz, M.T. (2018). A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci Rep* 8, 16757–16771.
- Cai, T., Sun, H., Qiao, J., Zhu, L., Zhang, F., Zhang, J., Tang, Z., Wei, X., Yang, J., Yuan, Q., et al. (2021). Cell-free chemoenzymatic starch synthesis from carbon dioxide. *Science* 373, 1523–1527.
- Callura, J.M., Dwyer, D.J., Isaacs, F.J., Cantor, C.R., and Collins, J.J. (2010). Tracking, tuning, and terminating microbial physiology using synthetic riboregulators. *Proc Natl Acad Sci USA* 107, 15898–15903.
- Cao, L., Goresnik, I., Coventry, B., Case, J.B., Miller, L., Kozodoy, L., Chen, R.E., Carter, L., Walls, A.C., Park, Y.J., et al. (2020). *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* 370, 426–431.
- Carbonell, P., Radivojevic, T., and García Martín, H. (2019). Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth Biol* 8, 1474–1477.
- Ceelen, L.M., Decostere, A., Ducatelle, R., and Haesebrouck, F. (2006). Cytolethal distending toxin generates cell death by inducing a bottleneck in the cell cycle. *Microbiol Res* 161, 109–120.
- Cella, F., Wroblewska, L., Weiss, R., and Siciliano, V. (2018). Engineering protein-protein devices for multilayered regulation of mRNA translation using orthogonal proteases in mammalian cells. *Nat Commun* 9, 4392.
- Cello, J., Paul, A.V., and Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297, 1016–1018.
- Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* 39, e82.
- Ceze, L., Nivala, J., and Strauss, K. (2019). Molecular digital data storage using DNA. *Nat Rev Genet* 20, 456–466.
- Chao, R., Liang, J., Tasan, I., Si, T., Ju, L., and Zhao, H. (2017a). Fully automated one-step synthesis of single-transcript TALEN pairs using a biological foundry. *ACS Synth Biol* 6, 678–685.
- Chao, R., Mishra, S., Si, T., and Zhao, H. (2017b). Engineering biological systems using automated biofoundries. *Metab Eng* 42, 98–108.
- Charbit, A., Boulain, J.C., Ryter, A., and Hofnung, M. (1986). Probing the topology of a bacterial membrane protein by genetic insertion of a foreign epitope; expression at the cell surface. *EMBO J* 5, 3029–3037.
- Chau, A.H., Walter, J.M., Gerardin, J., Tang, C., and Lim, W.A. (2012). Designing synthetic regulatory networks capable of self-organizing cell polarization. *Cell* 151, 320–332.
- Chen, K., Kong, J., Zhu, J., Ermann, N., Predki, P., and Keyser, U.F. (2019a). Digital data storage using DNA nanostructures and solid-state nanopores. *Nano Lett* 19, 1210–1215.
- Chen, K., Zhu, J., Bošković, F., and Keyser, U.F. (2020a). Nanopore-based DNA hard drives for rewritable and secure data storage. *Nano Lett* 20, 3754–3760.
- Chen, P.J., Hussmann, J.A., Yan, J., Knipping, F., Ravisankar, P., Chen, P. F., Chen, C., Nelson, J.W., Newby, G.A., Sahin, M., et al. (2021a). Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* 184, 5635–5652.e29.
- Chen, T., Hongdilokkul, N., Liu, Z., Adhikary, R., Tsuen, S.S., and Romesberg, F.E. (2016). Evolution of thermophilic DNA polymerases for the recognition and amplification of C2'-modified DNA. *Nat Chem* 8, 556–562.
- Chen, W., Han, M., Zhou, J., Ge, Q., Wang, P., Zhang, X., Zhu, S., Song, L., and Yuan, Y. (2021b). An artificial chromosome for data storage. *Natl Sci Rev* 8, nwab028.
- Chen, W.G., Huang, G., Li, B.Z., Yin, Y., and Yuan, Y.J. (2020b). DNA information storage for audio and video files. *Sci Sin-Vitae* 50, 81–85.
- Chen, Y., Zhang, S., Young, E.M., Jones, T.S., Densmore, D., and Voigt, C. A. (2020c). Genetic circuit design automation for yeast. *Nat Microbiol* 5, 1349–1360.
- Chen, Z., Boyken, S.E., Jia, M., Busch, F., Flores-Solis, D., Bick, M.J., Lu, P., VanAernum, Z.L., Sahasrabudhe, A., Langan, R.A., et al. (2019b). Programmable design of orthogonal protein heterodimers. *Nature* 565, 106–111.
- Chen, Z., and Elowitz, M.B. (2021). Programmable protein circuit design. *Cell* 184, 2284–2301.
- Chen, Z., Kibler, R.D., Hunt, A., Busch, F., Pearl, J., Jia, M., VanAernum,

- Z.L., Wicky, B.I.M., Dods, G., Liao, H., et al. (2020d). *De novo* design of protein logic gates. *Science* 368, 78–84.
- Cherkasov, A., Hilpert, K., Jenssen, H., Fjell, C.D., Waldbrook, M., Mullaly, S.C., Volkmer, R., and Hancock, R.E.W. (2009). Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chem Biol* 4, 65–74.
- Chica, R.A., Doucet, N., and Pelletier, J.N. (2005). Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr Opin Biotechnol* 16, 378–384.
- Cho, S.I., Lee, S., Mok, Y.G., Lim, K., Lee, J., Lee, J.M., Chung, E., and Kim, J.S. (2022). Targeted A-to-G base editing in human mitochondrial DNA with programmable deaminases. *Cell* 185, 1764–1776.e12.
- Choi, J., Chen, W., Suiter, C.C., Lee, C., Chardon, F.M., Yang, W., Leith, A., Daza, R.M., Martin, B., and Shendure, J. (2022). Precise genomic deletions using paired prime editing. *Nat Biotechnol* 40, 218–226.
- Chou, C.J., Jenney Jr, F.E., Adams, M.W.W., and Kelly, R.M. (2008). Hydrogenesis in hyperthermophilic microorganisms: implications for biofuels. *Metab Eng* 10, 394–404.
- Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., et al. (2012). CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* 8, e1002708.
- Cravens, A., Jamil, O.K., Kong, D., Sockolosky, J.T., and Smolke, C.D. (2021). Polymerase-guided base editing enables *in vivo* mutagenesis and rapid protein engineering. *Nat Commun* 12, 1–2.
- Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* 186, 757–761.
- Church, G.M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science* 337, 1628.
- Cheng, J.Y. (2002). High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer. *Nucleic Acids Res* 30, 93e–93.
- Channon, K., Bromley, E.H.C., and Woolfson, D.N. (2008). Synthetic biology through biomolecular design and engineering. *Curr Opin Struct Biol* 18, 491–498.
- Coley, C.W., Eyke, N.S., and Jensen, K.F. (2020a). Autonomous discovery in the chemical sciences part I: progress. *Angew Chem Int Ed* 59, 22858–22893.
- Coley, C.W., Eyke, N.S., and Jensen, K.F. (2020b). Autonomous discovery in the chemical sciences part II: outlook. *Angew Chem Int Ed* 59, 23414–23436.
- Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823.
- Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M. J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. *Science* 358, 1019–1027.
- Crone, M.A., Priestman, M., Ciecionska, M., Jensen, K., Sharp, D.J., Anand, A., Randell, P., Storch, M., and Freemont, P.S. (2020). A role for Biofoundries in rapid development and validation of automated SARS-CoV-2 clinical diagnostics. *Nat Commun* 11, 4464.
- Cui, Y., Chen, Y., Liu, X., Dong, S., Tian, Y., Qiao, Y., Mitra, R., Han, J., Li, C., Han, X., et al. (2021a). Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy. *ACS Catal* 11, 1340–1350.
- Cui, Y., Wang, Y., Tian, W., Bu, Y., Li, T., Cui, X., Zhu, T., Li, R., and Wu, B. (2021b). Development of a versatile and efficient C–N lyase platform for asymmetric hydroamination via computational enzyme redesign. *Nat Catal* 4, 364–373.
- Dai, J., Boeke, J.D., Luo, Z., Jiang, S., and Cai, Y. (2020). Sc3.0: revamping and minimizing the yeast genome. *Genome Biol* 21, 205.
- Davis, J. (1996). *Microvenus*. *Art J* 55, 70–74.
- de la Torre, D. (2020). *Rewriting the genome of Escherichia coli*. Dissertation for Doctoral Degree. Cambridge: University of Cambridge.
- Dehouck, Y., Kwasigroch, J.M., Gilis, D., and Rooman, M. (2011). Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinf* 12, 151–162.
- de Kok, S., Stanton, L.H., Slaby, T., Durot, M., Holmes, V.F., Patel, K.G., Platt, D., Shapland, E.B., Serber, Z., Dean, J., et al. (2014). Rapid and reliable DNA assembly *via* ligase cycling reaction. *ACS Synth Biol* 3, 97–106.
- de Rond, T., Gao, J., Zargar, A., de Raad, M., Cunha, J., Northen, T.R., and Keasling, J.D. (2019). A high-throughput mass spectrometric enzyme activity assay enabling the discovery of cytochrome P450 biocatalysts. *Angew Chem Int Ed* 58, 10114–10119.
- Deng, X., Chen, L., Hei, M., Liu, T., Feng, Y., and Yang, G.Y. (2020). Structure-guided reshaping of the acyl binding pocket of ‘TesA thioesterase enhances octanoic acid production in *E. coli*. *Metab Eng* 61, 24–32.
- Deng, Z.X. (2019). Synthetic biology takes advantage of the golden age, building to know, building to use (in Chinese). *Life Sci* 31, 323–324.
- Dharmadi, Y., Patel, K., Shapland, E., Hollis, D., Slaby, T., Klinkner, N., Dean, J., and Chandran, S.S. (2014). High-throughput, cost-effective verification of structural DNA assembly. *Nucleic Acids Res* 42, e22.
- Divine, R., Dang, H.V., Ueda, G., Fallas, J.A., Vulovic, I., Sheffler, W., Saini, S., Zhao, Y.T., Raj, I.X., Morawski, P.A., et al. (2021). Designed proteins assemble antibodies into modular nanocages. *Science* 372, eabd9994.
- Dong, Y., Sun, F., Ping, Z., Ouyang, Q., and Qian, L. (2020). DNA storage: research landscape and future prospects. *Natl Sci Rev* 7, 1092–1107.
- Dormitzer, P.R., Suphaphiphat, P., Gibson, D.G., Wentworth, D.E., Stockwell, T.B., Algire, M.A., Alperovich, N., Barro, M., Brown, D. M., Craig, S., et al. (2013). Synthetic generation of influenza vaccine viruses for rapid response to pandemics. *Sci Transl Med* 5, 185ra68.
- Dudley, Q.M., Cai, Y.M., Kallam, K., Debreyne, H., Carrasco Lopez, J.A., and Patron, N.J. (2021). Biofoundry-assisted expression and characterization of plant proteins. *Synth Biol* 6, ysab029.
- Dymond, J.S., Richardson, S.M., Coombes, C.E., Babatz, T., Muller, H., Annaluru, N., Blake, W.J., Scherzmann, J.W., Dai, J., Lindstrom, D. L., et al. (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* 477, 471–476.
- Eisenstein, M. (2020). Enzymatic DNA synthesis enters new phase. *Nat Biotechnol* 38, 1113–1115.
- Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* 403, 335–338.
- Elowitz, M., and Lim, W.A. (2010). Build life to understand it. *Nature* 468, 889–890.
- Engelowski, E., Schneider, A., Franke, M., Xu, H., Clemen, R., Lang, A., Baran, P., Binsch, C., Knebel, B., Al-Hasani, H., et al. (2018). Synthetic cytokine receptors transmit biological signals using artificial ligands. *Nat Commun* 9, 2034.
- Engineering Biology Research Consortium (2019). *Engineering biology: a Research Roadmap for the Next-Generation Bioeconomy*. Retrieved from <https://roadmap.ebrc.org>. doi: 10.25498/E4159B.
- Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* 3, e3647.
- Epinat, J.C., Arnould, S., Chames, P., Rochaix, P., Desfontaines, D., Puzin, C., Patin, A., Zanghellini, A., Pâques, F., and Lacroix, E. (2003). A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res* 31, 2952–2962.
- Erickson, D.W., Schink, S.J., Patsalo, V., Williamson, J.R., Gerland, U., and Hwa, T. (2017). A global resource allocation strategy governs growth transition kinetics of *Escherichia coli*. *Nature* 551, 119–123.
- Erlach, Y., and Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 950–954.
- Exley, K., Reynolds, C.R., Suckling, L., Chee, S.M., Tsipa, A., Freemont, P. S., McClymont, D., and Kitney, R.I. (2019). Utilising datasheets for the informed automated design and build of a synthetic metabolic pathway. *J Biol Eng* 13, 8.
- Fallah-Araghi, A., Baret, J.C., Ryckelynck, M., and Griffiths, A.D. (2012).



- A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab Chip* 12, 882–891.
- Fan, C., Deng, Q., and Zhu, T.F. (2021). Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image Pfu DNA polymerase. *Nat Biotechnol* 39, 1548–1555.
- Fan, Z., Fang, L., Wu, L., Wang, Z., Wang, Y., Han, C., and Liu, X. (2022). Improved catalytic activity of a novel aspartate kinase by site-directed saturation mutagenesis. *Bioprocess Biosyst Eng* 45, 541–551.
- Farzaneh, T., and Freemont, P.S. (2021). Biofoundries are a nucleating hub for industrial translation. *Synth Biol* 6, 1–6.
- Faulon, J.L., and Faure, L. (2021). In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. *Curr Opin Chem Biol* 65, 85–92.
- Fernandez-Rodriguez, J., and Voigt, C.A. (2016). Post-translational control of genetic circuits using *Potyvirus* proteases. *Nucleic Acids Res* 44, 6493–6502.
- Fiscaro, P., and Boni, C. (2022). T and NK cell-based immunotherapy in chronic viral hepatitis and hepatocellular carcinoma. *Cells* 11, 180.
- Fogeron, M.L., Badillo, A., Jirasko, V., Gouttenoire, J., Paul, D., Lancien, L., Moradpour, D., Bartenschlager, R., Meier, B.H., Penin, F., et al. (2015). Wheat germ cell-free expression: two detergents with a low critical micelle concentration allow for production of soluble HCV membrane proteins. *Protein Expression Purification* 105, 39–46.
- Fredens, J., Wang, K., de la Torre, D., Funke, L.F.H., Robertson, W.E., Christova, Y., Chia, T., Schmied, W.H., Dunkelmann, D.L., Beránek, V., et al. (2019). Total synthesis of *Escherichia coli* with a recoded genome. *Nature* 569, 514–518.
- Fu, L., Zhang, J., and Si, T. (2020). Recent advances in high-throughput mass spectrometry that accelerates enzyme engineering for biofuel research. *BMC Energy* 2, 1.
- Fujisawa, T., Fujinaga, S., and Atomi, H. (2017). An *In Vitro* enzyme system for the production of *myo*-inositol from starch. *Appl Environ Microbiol* 83, e00550.
- Funke, M., Buchenauer, A., Schnakenberg, U., Mokwa, W., Diederichs, S., Mertens, A., Müller, C., Kensy, F., and Büchs, J. (2010). Microfluidic biolector-microfluidic bioprocess control in microtiter plates. *Biotechnol Bioeng* 107, 497–505.
- Gao, C. (2021). Genome engineering for crop improvement and future agriculture. *Cell* 184, 1621–1635.
- Gao, Y., Chen, X., Qiao, H., Ke, Y., and Qi, H. (2020). Low-bias manipulation of DNA oligo pool for robust data storage. *ACS Synth Biol* 9, 3344–3352.
- Gao, X.J., Chong, L.S., Kim, M.S., and Elowitz, M.B. (2018). Programmable protein circuits in living cells. *Science* 361, 1252–1258.
- Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403, 339–342.
- Garamella, J., Marshall, R., Rustad, M., and Noireaux, V. (2016). The all *E. coli* TX-TL Toolbox 2.0: a platform for cell-free synthetic biology. *ACS Synth Biol* 5, 344–355.
- Garst, A.D., Bassalo, M.C., Pines, G., Lynch, S.A., Halweg-Edwards, A.L., Liu, R., Liang, L., Wang, Z., Zeitoun, R., Alexander, W.G., et al. (2017). Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat Biotechnol* 35, 48–55.
- Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci USA* 109, 2579–2586.
- Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551, 464–471.
- Gaudelli, N.M., Lam, D.K., Rees, H.A., Solá-Esteves, N.M., Barrera, L.A., Born, D.A., Edwards, A., Gehrke, J.M., Lee, S.J., Liquori, A.J., et al. (2020). Directed evolution of adenine base editors with increased activity and therapeutic application. *Nat Biotechnol* 38, 892–900.
- Ghosh, E., Dwivedi, H., Baidya, M., Srivastava, A., Kumari, P., Stepniwski, T., Kim, H.R., Lee, M.H., van Gastel, J., Chaturvedi, M., et al. (2019). Conformational sensors and domain swapping reveal structural and functional differences between  $\beta$ -arrestin isoforms. *Cell Rep* 28, 3287–3299.e6.
- Gibson, D.G. (2009a). Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res* 37, 6984–6990.
- Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., et al. (2008a). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215–1220.
- Gibson, D.G., Benders, G.A., Axelrod, K.C., Zaveri, J., Algire, M.A., Moodie, M., Montague, M.G., Venter, J.C., Smith, H.O., and Hutchison Clyde A., I. (2008b). One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. *Proc Natl Acad Sci USA* 105, 20404–20409.
- Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., et al. (2010a). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56.
- Gibson, D.G., Smith, H.O., Hutchison III, C.A., Venter, J.C., and Merryman, C. (2010b). Chemical synthesis of the mouse mitochondrial genome. *Nat Methods* 7, 901–903.
- Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison III, C.A., and Smith, H.O. (2009b). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6, 343–345.
- Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., and Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat Chem Biol* 9, 494–498.
- Gillmore, J.D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M.L., Seitzer, J., O’Connell, D., Walsh, K.R., Wood, K., et al. (2021). CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. *N Engl J Med* 385, 493–502.
- Glasgow, A.A., Huang, Y.M., Mandell, D.J., Thompson, M., Ritterson, R., Loshbaugh, A.L., Pellegrino, J., Krivacic, C., Pache, R.A., Barlow, K. A., et al. (2019). Computational design of a modular protein sense-response system. *Science* 366, 1024–1028.
- Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., and Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494, 77–80.
- Goodsell, D.S. (2002). The molecular perspective: restriction endonucleases. *Stem Cells* 20, 190–191.
- Gowers, G.O.F., Cameron, S.J.S., Perdonés-Montero, A., Bell, D., Chee, S. M., Kern, M., Tew, D., Ellis, T., and Takáts, Z. (2019). Off-colony screening of biosynthetic libraries by rapid laser-enabled mass spectrometry. *ACS Synth Biol* 8, 2566–2575.
- Gowers, G.O.F., Chee, S.M., Bell, D., Suckling, L., Kern, M., Tew, D., McClymont, D.W., and Ellis, T. (2020). Improved betulinic acid biosynthesis using synthetic yeast chromosome recombination and semi-automated rapid LC-MS screening. *Nat Commun* 11, 868.
- Grass, R.N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W.J. (2015). Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew Chem Int Ed* 54, 2552–2555.
- Gray, D.C., Mahrus, S., and Wells, J.A. (2010). Activation of specific apoptotic caspases with an engineered small-molecule-activated protease. *Cell* 142, 637–646.
- Gräwe, A., Dreyer, A., Vornholt, T., Barteczko, U., Buchholz, L., Drews, G., Ho, U.L., Jackowski, M.E., Kracht, M., Lüders, J., et al. (2019). A paper-based, cell-free biosensor system for the detection of heavy metals and date rape drugs. *PLoS ONE* 14, e0210940.
- Greener, J.G., Kandathil, S.M., Moffat, L., and Jones, D.T. (2022). A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 23, 40–55.
- Guo, Y., Dong, J., Zhou, T., Auxillos, J., Li, T., Zhang, W., Wang, L., Shen, Y., Luo, Y., Zheng, Y., et al. (2015). YeastFab: the design and construction of standard biological parts for metabolic engineering in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 43, e88.



- Guterl, J.K., Garbe, D., Carsten, J., Steffler, F., Sommer, B., Reiß, S., Philipp, A., Haack, M., Rühmann, B., Koltermann, A., et al. (2012). Cell-free metabolic engineering: production of chemicals by minimized reaction cascades. *ChemSusChem* 5, 2165–2172.
- Halperin, S.O., Tou, C.J., Wong, E.B., Modavi, C., Schaffer, D.V., and Dueber, J.E. (2018). CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* 560, 248–252.
- Hamedirad, M., Chao, R., Weisberg, S., Lian, J., Sinha, S., and Zhao, H. (2019). Towards a fully automated algorithm driven platform for biosystems design. *Nat Commun* 10, 5150.
- Hasty, J., McMillen, D., and Collins, J.J. (2002). Engineered gene circuits. *Nature* 420, 224–230.
- Han, M.Z., Chen, W.G., Song, L.F., Li, B.Z., and Yuan, Y.J., (2021). DNA information storage: bridging biological and digital world. *Synth Biol J* 2, 309–322.
- Hanes, J., and Plückthun, A. (1997). *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci USA* 94, 4937–4942.
- Harrington, L.B., Burstein, D., Chen, J.S., Paez-Espino, D., Ma, E., Witte, I.P., Cofsky, J.C., Kyrpides, N.C., Banfield, J.F., and Doudna, J.A. (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* 362, 839–842.
- Heckel, R., Mikutis, G., and Grass, R.N. (2019). A characterization of the DNA data storage channel. *Sci Rep* 9, 9663.
- Hernandez-Lopez, R.A., Yu, W., Cabral, K.A., Creasey, O.A., Lopez Pazmino, M.D.P., Tonai, Y., De Guzman, A., Mäkelä, A., Saksela, K., Gartner, Z.J., et al. (2021). T cell circuits that sense antigen density with an ultrasensitive threshold. *Science* 371, 1166–1171.
- Hervé, J.C., and Derangeon, M. (2013). Gap-junction-mediated cell-to-cell communication. *Cell Tissue Res* 352, 21–31.
- Hillson, N., Caddick, M., Cai, Y., Carrasco, J.A., Chang, M.W., Curach, N. C., Bell, D.J., Le Feuvre, R., Friedman, D.C., Fu, X., et al. (2019). Building a global alliance of biofoundries. *Nat Commun* 10, 2040.
- Hillson, N.J., Rosengarten, R.D., and Keasling, J.D. (2012). j5 DNA assembly design automation software. *ACS Synth Biol* 1, 14–21.
- Hong, S.H., Ntai, I., Haimovich, A.D., Kelleher, N.L., Isaacs, F.J., and Jewett, M.C. (2014). Cell-free protein synthesis from a release factor 1 deficient *Escherichia coli* activates efficient and multiple site-specific nonstandard amino acid incorporation. *ACS Synth Biol* 3, 398–409.
- Hoque, M.A., Zhang, Y., Li, Z., Cui, L., and Feng, Y. (2020). Remodeling enzyme active sites by stepwise loop insertion. *Methods Enzymol* 643, 111–127.
- Hori, Y., Katak, C., Murray, R.M., and Abate, A.R. (2017). Cell-free extract based optimization of biomolecular circuits with droplet microfluidics. *Lab Chip* 17, 3037–3042.
- Hu, C.Y., Takahashi, M.K., Zhang, Y., and Lucks, J.B. (2018). Engineering a functional small RNA negative autoregulation network with model-guided design. *ACS Synth Biol* 7, 1507–1518.
- Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J., Huang, C., Hong, J., Chen, Q., and Liu, H. (2022). A backbone-centred energy function of neural networks for protein design. *Nature* 602, 523–528.
- Huang, K., Xiao, C., Glass, L.M., Critchlow, C.W., Gibson, G., and Sun, J. (2021). Machine learning applications for therapeutic tasks with genomics data. *Patterns* 2, 100328.
- Huang, P.S., Boyken, S.E., and Baker, D. (2016). The coming of age of *de novo* protein design. *Nature* 537, 320–327.
- Hughes, R.A., and Ellington, A.D. (2017). Synthetic DNA Synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harb Perspect Biol* 9, a023812.
- Hughes, R.A., Miklos, A.E., and Ellington, A.D. (2011). Gene synthesis: methods and applications. *Methods Enzymol* 498, 277–309.
- Hunt, J.P., Zhao, E.L., Free, T.J., Soltani, M., Warr, C.A., Benedict, A.B., Takahashi, M.K., Griffiths, J.S., Pitt, W.G., and Bundy, B.C. (2022). Towards detection of SARS-CoV-2 RNA in human saliva: a paper-based cell-free toehold switch biosensor with a visual bioluminescent output. *New Biotechnol* 66, 53–60.
- Hutchison III, C.A., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351, aad6253.
- Im, D.J., Jeong, S.N., Yoo, B.S., Kim, B., Kim, D.P., Jeong, W.J., and Kang, I.S. (2015). Digital microfluidic approach for efficient electroporation with high productivity: transgene expression of microalgae without cell wall removal. *Anal Chem* 87, 6592–6599.
- Ip, K., Yadin, R., and George, K.W. (2020). High-throughput DNA assembly using yeast homologous recombination. *Methods Mol Biol* 2205, 79–89.
- Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* 169, 5429–5433.
- Itaya, M., Fujita, K., Kuroki, A., and Tsuge, K. (2008). Bottom-up genome assembly using the *Bacillus subtilis* genome vector. *Nat Methods* 5, 41–43.
- Itaya, M., Tsuge, K., Koizumi, M., and Fujita, K. (2005). Combining two genomes in one cell: stable cloning of the *Synechocystis* PCC6803 genome in the *Bacillus subtilis* 168 genome. *Proc Natl Acad Sci USA* 102, 15971–15976.
- Jackson, H.J., Rafiq, S., and Brentjens, R.J. (2016). Driving CAR T-cells forward. *Nat Rev Clin Oncol* 13, 370–383.
- Jansen, R., Embden, J.D.A., Gastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43, 1565–1575.
- Jaroentomechai, T., Stark, J.C., Natarajan, A., Glasscock, C.J., Yates, L.E., Hsu, K.J., Mrksich, M., Jewett, M.C., and DeLisa, M.P. (2018). Author Correction: Single-pot glycoprotein biosynthesis using a cell-free transcription-translation system enriched with glycosylation machinery. *Nat Commun* 9, 3396.
- Jiang, S., Tang, Y., Xiang, L., Zhu, X., Cai, Z., Li, L., Chen, Y., Chen, P., Feng, Y., Lin, X., et al. (2022). Efficient *de novo* assembly and modification of large DNA fragments. *Sci China Life Sci* 65, 1445–1455.
- Jiang, S., Zhao, S., Cai, Z., Tang, Y., and Dai, J. (2020a). Synthetic yeast genomes for studying chromosomal features. *Curr Opin Syst Biol* 23, 1–7.
- Jiang, S., Si, T., and Dai, J. (2020b). Whole-genome regulation for yeast metabolic engineering. *Small Methods* 4, 1900640.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.
- Jones, T.S., Oliveira, S.M.D., Myers, C.J., Voigt, C.A., and Densmore, D. (2022). Genetic circuit design automation with Cello 2.0. *Nat Protoc* 17, 1097–1113.
- Junge, F., Haberstock, S., Roos, C., Stefer, S., Proverbio, D., Dötsch, V., and Bernhard, F. (2011). Advances in cell-free protein synthesis for the functional and structural analysis of membrane proteins. *New Biotechnol* 28, 262–271.
- Jürgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000). Directed evolution of a ( $\beta$ )<sub>8</sub>-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc Natl Acad Sci USA* 97, 9925–9930.
- Kai, L., and Schille, P. (2019). Cell-free protein synthesis and its perspectives for assembling cells from the bottom-up. *Adv Biosyst* 3, e1800322.
- Kaiser, L., Graveland-Bikker, J., Steuerwald, D., Vanberghem, M., Herlihy, K., and Zhang, S. (2008). Efficient cell-free production of olfactory receptors: detergent optimization, structure, and ligand binding analyses. *Proc Natl Acad Sci USA* 105, 15726–15731.
- Kan, S.B.J., Lewis, R.D., Chen, K., and Arnold, F.H. (2016). Directed

- evolution of cytochrome c for carbon-silicon bond formation: bringing silicon to life. *Science* 354, 1048–1051.
- Kanigowska, P., Shen, Y., Zheng, Y., Rosser, S., and Cai, Y. (2016). Smart DNA fabrication using sound waves: applying acoustic dispensing technologies to synthetic biology. *SLAS Tech* 21, 49–56.
- Karas, B.J., Molparia, B., Jablanovic, J., Hermann, W.J., Lin, Y.C., Dupont, C.L., Tagwerker, C., Yonemoto, I.T., Noskov, V.N., Chuang, R.Y., et al. (2013). Assembly of eukaryotic algal chromosomes in yeast. *J Biol Eng* 7, 30.
- Karim, A.S., Dudley, Q.M., Juminaga, A., Yuan, Y., Crowe, S.A., Heggstad, J.T., Garg, S., Abdalla, T., Grubbe, W.S., Rasor, B.J., et al. (2020). In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat Chem Biol* 16, 912–919.
- Karim, A.S., and Jewett, M.C. (2016). A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab Eng* 36, 116–126.
- Katrekar, D., Yen, J., Xiang, Y., Saha, A., Meluzzi, D., Savva, Y., and Mali, P. (2022). Efficient in vitro and in vivo RNA editing via recruitment of endogenous ADARs using circular guide RNAs. *Nat Biotechnol* 40, 938–945.
- Kather, I., Jakob, R.P., Dobbek, H., and Schmid, F.X. (2008). Increased folding stability of TEM-1  $\beta$ -lactamase by in vitro selection. *J Mol Biol* 383, 238–251.
- Khan, S., Hauptman, R., and Kelly, L. (2021). Engineering the microbiome to prevent adverse events: challenges and opportunities. *Annu Rev Pharmacol Toxicol* 61, 159–179.
- Khorana, H.G., Buuchi, H., Ghosh, H., Gupta, N., Jacob, T.M., Kossel, H., Morgan, R., Narang, S.A., Ohtsuka, E., and Wells, R.D. (1966). Polynucleotide synthesis and the genetic code. *Cold Spring Harbor Symposia Quantit Biol* 31, 39–49.
- Kiel, C., Yus, E., and Serrano, L. (2010). Engineering signal transduction pathways. *Cell* 140, 33–47.
- Kim, D.Y., Lee, J.M., Moon, S.B., Chin, H.J., Park, S., Lim, Y., Kim, D., Koo, T., Ko, J.H., and Kim, Y.S. (2022). Efficient CRISPR editing with a hypercompact Cas2f1 and engineered guide RNAs delivered by adeno-associated virus. *Nat Biotechnol* 40, 94–102.
- Kim, J.Y., Yoo, H.W., Lee, P.G., Lee, S.G., Seo, J.H., and Kim, B.G. (2019). In vivo protein evolution, next generation protein engineering strategy: from random approach to target-specific approach. *Biotechnol Bioproc E* 24, 85–94.
- Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc Natl Acad Sci USA* 93, 1156–1160.
- Kitano, H. (2021). Nobel Turing Challenge: creating the engine for scientific discovery. *npj Syst Biol Appl* 7, 29.
- Klichinsky, M., Ruella, M., Shestova, O., Lu, X.M., Best, A., Zeeman, M., Schmierer, M., Gabrusiewicz, K., Anderson, N.R., Petty, N.E., et al. (2020). Human chimeric antigen receptor macrophages for cancer immunotherapy. *Nat Biotechnol* 38, 947–953.
- Koblan, L.W., Erdos, M.R., Wilson, C., Cabral, W.A., Levy, J.M., Xiong, Z.M., Tavarez, U.L., Davison, L.M., Gete, Y.G., Mao, X., et al. (2021). In vivo base editing rescues Hutchinson-Gilford progeria syndrome in mice. *Nature* 589, 608–614.
- Koch, J., Gantenbein, S., Masania, K., Stark, W.J., Erlich, Y., and Grass, R. N. (2020). A DNA-of-things storage architecture to create materials with embedded memory. *Nat Biotechnol* 38, 39–43.
- Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424.
- Komor, A.C., Zhao, K.T., Packer, M.S., Gaudelli, N.M., Waterbury, A.L., Koblan, L.W., Kim, Y.B., Badran, A.H., and Liu, D.R. (2017). Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci Adv* 3, ea04774.
- Konno, N., Kijima, Y., Watano, K., Ishiguro, S., Ono, K., Tanaka, M., Mori, H., Masuyama, N., Pratt, D., Ideker, T., et al. (2022). Deep distributed computing to reconstruct extremely large lineage trees. *Nat Biotechnol* 40, 566–575.
- Kosuri, S., and Church, G.M. (2014). Large-scale *de novo* DNA synthesis: technologies and applications. *Nat Methods* 11, 499–507.
- Kosuri, S., Eroshenko, N., Leproust, E.M., Super, M., Way, J., Li, J.B., and Church, G.M. (2010). Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28, 1295–1299.
- Kotopka, B.J., and Smolke, C.D. (2020). Model-driven generation of artificial yeast promoters. *Nat Commun* 11, 2113.
- Kotsias, F., Cebrian, I., and Alloati, A. (2019). Antigen processing and presentation. *Int Rev Cell Mol Biol* 348, 69–121.
- Krokan, H.E., Drablos, F., and Slupphaug, G. (2002). Uracil in DNA—occurrence, consequences and repair. *Oncogene* 21, 8935–8948.
- Kung, Y.T., Du, Y.C., Huang, W.T., Chen, C.C., Ke, L.T., Hu, S.C., Jiang, R.Q., Chu, S.Q., Niu, C.I., Hsu, J.Z., et al. (1965). Total synthesis of crystalline bovine insulin. *Sci Sin* 14, 1710–1716.
- Lapidot, Y., De Groot, N., Rappoport, S., and Elat, D. (1969). Peptidyl transfer RNA VIII. The chemical synthesis of glycine and alanine containing oligopeptidyl transfer RNA. *Biochim Biophys Acta (BBA)-Nucleic Acids Protein Synthesis* 190, 304–311.
- Laskowski, R.A., and Swindells, M.B. (2011). LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* 51, 2778–2786.
- Lau, Y.H., Stirling, F., Kuo, J., Karrenbelt, M.A.P., Chan, Y.A., Riesselman, A., Horton, C.A., Schäfer, E., Lips, D., Weinstock, M.T., et al. (2017). Large-scale recoding of a bacterial genome by iterative recombineering of synthetic DNA. *Nucleic Acids Res* 45, 6971–6980.
- Lawson, C.E., Martí, J.M., Radivojevic, T., Jonnalagadda, S.V.R., Gentz, R., Hillson, N.J., Peisert, S., Kim, J., Simmons, B.A., Petzold, C.J., et al. (2021). Machine learning for metabolic engineering: a review. *Metab Eng* 63, 34–60.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.
- Lee, H., Wiegand, D.J., Griswold, K., Punthambaker, S., Chun, H., Kohman, R.E., and Church, G.M. (2020). Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nat Commun* 11, 5246.
- Lee, H.H., Kalthor, R., Goela, N., Bolot, J., and Church, G.M. (2019). Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat Commun* 10, 2383.
- Leman, J.K., Weitzner, B.D., Lewis, S.M., Adolf-Bryfogle, J., Alam, N., Alford, R.F., Aprahamian, M., Baker, D., Barlow, K.A., Barth, P., et al. (2020). Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nat Methods* 17, 665–680.
- Li, G., Dong, Y., and Reetz, M.T. (2019a). Can machine learning revolutionize directed evolution of selective enzymes? *Adv Synth Catal* 361, 2377–2386.
- Li, H., and d’Anjou, M. (2009). Pharmacological significance of glycosylation in therapeutic proteins. *Curr Opin Biotechnol* 20, 678–684.
- Li, J.H., Mutanda, I., Wang, K., Yang, L., Wang, J., and Wang, Y. (2019b). Chloroplastic metabolic engineering coupled with isoprenoid pool enhancement for committed taxanes biosynthesis in *Nicotiana benthamiana*. *Nat Commun* 10, 4850.
- Li, M.V., Shukla, D., Rhodes, B.H., Lall, A., Shu, J., Moriarity, B.S., and Largaespada, D.A. (2014). HomeRun Vector Assembly System: a flexible and standardized cloning system for assembly of multi-modular DNA constructs. *PLoS ONE* 9, e100948.
- Li, M.Z., and Elledge, S.J. (2007). Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* 4, 251–256.
- Li, S., Lin, D., Zhang, Y., Deng, M., Chen, Y., Lv, B., Li, B., Lei, Y., Wang, Y., Zhao, L., et al. (2022). Genome-edited powdery mildew resistance in wheat without growth penalties. *Nature* 602, 455–460.
- Li, S., Wen, J., and Jia, X. (2011a). Engineering *Bacillus subtilis* for isobutanol production by heterologous Ehrlich pathway construction and the biosynthetic 2-ketoisovalerate precursor pathway

- overexpression. *Appl Microbiol Biotechnol* 91, 577–589.
- Li, T., Huang, S., Jiang, W.Z., Wright, D., Spalding, M.H., Weeks, D.P., and Yang, B. (2011b). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res* 39, 359–372.
- Li, Y., Shi, T., Han, P., and You, C. (2021). Thermodynamics-driven production of value-added<sup>d</sup>-allulose from inexpensive starch by an *In Vitro* enzymatic synthetic biosystem. *ACS Catal* 11, 5088–5099.
- Liang, F., Han, M., Romanienko, P.J., and Jasin, M. (1998). Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc Natl Acad Sci USA* 95, 5172–5177.
- Liang, L., Liu, R., Garst, A.D., Lee, T., Nogué, V.S.I., Beckham, G.T., and Gill, R.T. (2017). CRISPR enabled trackable genome engineering for isopropanol production in *Escherichia coli*. *Metab Eng* 41, 1–10.
- Lim, W.A., Lee, C.M., and Tang, C. (2013). Design principles of regulatory networks: searching for the molecular algorithms of the cell. *Mol Cell* 49, 202–212.
- Lindenburg, L.H., Pantelejevs, T., Gielen, F., Zuazua-Villar, P., Butz, M., Rees, E., Kaminski, C.F., Downs, J.A., Hyvönen, M., and Hollfelder, F. (2021). Improved RAD51 binders through motif shuffling based on the modularity of BRC repeats. *Proc Natl Acad Sci USA* 118, e2017708118.
- Lin, Q., Zong, Y., Xue, C., Wang, S., Jin, S., Zhu, Z., Wang, Y., Anzalone, A.V., Raguram, A., Doman, J.L., et al. (2020). Prime genome editing in rice and wheat. *Nat Biotechnol* 38, 582–585.
- Lin, Q., Jin, S., Zong, Y., Yu, H., Zhu, Z., Liu, G., Kou, L., Wang, Y., Qiu, J.L., Li, J., et al. (2021). High-efficiency prime editing with optimized, paired pegRNAs in plants. *Nat Biotechnol* 39, 923–927.
- Liu, C.L., Fu, X.F., Liu, L., Ren, X., Chau, C.K.L., Li, S., Xiang, L., Zeng, H., Chen, G., Tang, L.H., et al. (2011). Sequential establishment of stripe patterns in an expanding cell population. *Science* 334, 238–241.
- Liu, C.L., Fu, X., and Huang, J.D. (2013). Synthetic biology: a new approach to study biological pattern formation. *Quant Biol* 1, 246–252.
- Liu, J.J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019a). CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* 566, 218–223.
- Liu, Y., Qiu, L., Huang, J., Zhao, B., Wang, Z., Zhu, X., Gao, Y., and Shu, Z. (2015). Screening for mutants with thermostable lipase from *Burkholderia* sp. *Acta Microbiol Sin* 55, 748–754.
- Liu, J.K., Chen, W.H., Ren, S.X., Zhao, G.P., and Wang, J. (2014). iBrick: a new standard for iterative assembly of biological parts with homing endonucleases. *PLoS ONE* 9, e110852.
- Liu, W., Cremer, J., Li, D., Hwa, T., and Liu, C.L. (2019b). An evolutionarily stable strategy to colonize spatially extended habitats. *Nature* 575, 664–668.
- Lu, H., Diaz, D.J., Czarniecki, N.J., Zhu, C., Kim, W., Shroff, R., Acosta, D. J., Alexander, B.R., Cole, H.O., Zhang, Y., et al. (2022). Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* 604, 662–667.
- Luo, X., Reiter, M.A., d’Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., Lin, W., et al. (2019). Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* 567, 123–126.
- Luo, Z., Hoffmann, S.A., Jiang, S., Cai, Y., and Dai, J. (2020). Probing eukaryotic genome functions with synthetic chromosomes. *Exp Cell Res* 390, 111936.
- Luo, Z., Yang, Q., Geng, B., Jiang, S., Yang, S., Li, X., Cai, Y., and Dai, J. (2018). Whole genome engineering by synthesis. *Sci China Life Sci* 61, 1515–1527.
- Luo, Z., Yu, K., Xie, S., Monti, M., Schindler, D., Fang, Y., Zhao, S., Liang, Z., Jiang, S., Luan, M., et al. (2021). Compacting a synthetic yeast chromosome arm. *Genome Biol* 22, 5.
- Lutz, S. (2010). Beyond directed evolution—semi-rational protein engineering and design. *Curr Opin Biotechnol* 21, 734–743.
- Ma, D., Li, Y., Wu, K., Yan, Z., Tang, A.A., Chaudhary, S., Ticktin, Z.M., Alcantar-Fernandez, J., Moreno-Camacho, J.L., Campos-Romero, A., et al. (2022a). Multi-arm RNA junctions encoding molecular logic unconstrained by input sequence for versatile cell-free diagnostics. *Nat Biomed Eng* 6, 298–309.
- Ma, D., Shen, L., Wu, K., Diehnelt, C.W., and Green, A.A. (2018a). Low-cost detection of norovirus using paper-based cell-free systems and synbody-based viral enrichment. *Synth Biol* 3, ysy018.
- Ma, F., Chung, M.T., Yao, Y., Nidetz, R., Lee, L.M., Liu, A.P., Feng, Y., Kurabayashi, K., and Yang, G.Y. (2018b). Efficient molecular evolution to generate enantioselective enzymes using a dual-channel microfluidic droplet screening platform. *Nat Commun* 9, 1030–1037.
- Ma, W., Trusina, A., El-Samad, H., Lim, W.A., and Tang, C. (2009). Defining network topologies that can achieve biochemical adaptation. *Cell* 138, 760–773.
- Ma, Y., Budde, M.W., Mayalu, M.N., Zhu, J., Lu, A.C., Murray, R.M., and Elowitz, M.B. (2022b). Synthetic mammalian signaling circuits for robust cell population control. *Cell* 185, 967–979.e12.
- Malhotra, O.P., Ambasht, P.K., Prabhakar, P., Lal, A.K., and Kayastha, A. M. (1996). An assay procedure for determining the rate of an enzyme reaction lacking an optical signal: validity of coupled enzyme assays. *Biochem Education* 24, 56–59.
- Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa, I.R., and Romesberg, F.E. (2014). A semi-synthetic organism with an expanded genetic alphabet. *Nature* 509, 385–388.
- Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. *Science* 339, 823–826.
- Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322, 1843–1845.
- Martella, A., Matjusaitis, M., Auxillos, J., Pollard, S.M., and Cai, Y. (2017). EMMA: an extensible mammalian modular assembly toolkit for the rapid design and production of diverse expression vectors. *ACS Synth Biol* 6, 1380–1392.
- Martin, R.W., Des Soye, B.J., Kwon, Y.C., Kay, J., Davis, R.G., Thomas, P. M., Majewska, N.I., Chen, C.X., Marcum, R.D., Weiss, M.G., et al. (2018). Cell-free protein synthesis from genomically recoded bacteria enables multisite incorporation of noncanonical amino acids. *Nat Commun* 9, 1203.
- Matthies, D., Haberstock, S., Joos, F., Dötsch, V., Vonck, J., Bernhard, F., and Meier, T. (2011). Cell-free expression and assembly of ATP synthase. *J Mol Biol* 413, 593–603.
- Meiser, L.C., Antkowiak, P.L., Koch, J., Chen, W.D., Kohll, A.X., Stark, W.J., Heckel, R., and Grass, R.N. (2020). Reading and writing digital data in DNA. *Nat Protoc* 15, 86–101.
- Meiser, L.C., Nguyen, B.H., Chen, Y.J., Nivala, J., Strauss, K., Ceze, L., and Grass, R.N. (2022). Synthetic DNA applications in information technology. *Nat Commun* 13, 352.
- Meng, D., Wei, X., Bai, X., Zhou, W., and You, C. (2020). Artificial in vitro synthetic enzymatic biosystem for the one-pot sustainable biomanufacturing of glucosamine from starch and inorganic ammonia. *ACS Catal* 10, 13809–13819.
- Merkle, T., Merz, S., Reautschnig, P., Blaha, A., Li, Q., Vogel, P., Wettengel, J., Li, J.B., and Stafforst, T. (2019). Precise RNA editing by recruiting endogenous ADARs with antisense oligonucleotides. *Nat Biotechnol* 37, 133–138.
- Miller, M.B., and Bassler, B.L. (2001). Quorum sensing in bacteria. *Annu Rev Microbiol* 55, 165–199.
- Miller, T.E., Beneyton, T., Schwander, T., Diehl, C., Girault, M., McLean, R., Chotel, T., Claus, P., Cortina, N.S., Baret, J.C., et al. (2020a). Light-powered CO<sub>2</sub> fixation in a chloroplast mimic with natural and synthetic parts. *Science* 368, 649–654.
- Miller, S.M., Wang, T., and Liu, D.R. (2020b). Phage-assisted continuous and non-continuous evolution. *Nat Protoc* 15, 4101–4127.
- Miller, S.M., Wang, T., Randolph, P.B., Arbab, M., Shen, M.W., Huang, T. P., Matuszek, Z., Newby, G.A., Rees, H.A., and Liu, D.R. (2020c). Continuous evolution of SpCas9 variants compatible with non-G PAMs. *Nat Biotechnol* 38, 471–481.



- Mok, B.Y., de Moraes, M.H., Zeng, J., Bosch, D.E., Kotrys, A.V., Raguram, A., Hsu, F.S., Radey, M.C., Peterson, S.B., Mootha, V.K., et al. (2020). A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* 583, 631–637.
- Moore, C.L., Papa Iii, L.J., and Shoulders, M.D. (2018a). A processive protein chimera introduces mutations across defined DNA regions *in vivo*. *J Am Chem Soc* 140, 11560–11564.
- Moore, J.K., and Haber, J.E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol Cell Biol* 16, 2164–2173.
- Moore, S.J., MacDonald, J.T., Wienecke, S., Ishwarbhai, A., Tsipa, A., Aw, R., Kylilis, N., Bell, D.J., McClymont, D.W., Jensen, K., et al. (2018b). Rapid acquisition and model-based analysis of cell-free transcription-translation reactions from nonmodel bacteria. *Proc Natl Acad Sci USA* 115, E4340–E4349.
- Moris-Varas, F., Shah, A., Aikens, J., Nadkarni, N.P., Rozzell, J.D., and Demirjian, D.C. (1999). Visualization of enzyme-catalyzed reactions using pH indicators: rapid screening of hydrolase libraries and estimation of the enantioselectivity. *Bioorg Medicinal Chem* 7, 2183–2188.
- Morsut, L., Roybal, K.T., Xiong, X., Gordley, R.M., Coyle, S.M., Thomson, M., and Lim, W.A. (2016). Engineering customized cell sensing and response behaviors using synthetic notch receptors. *Cell* 164, 780–791.
- Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., and Kondo, T. (2005). Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation *in vitro*. *Science* 308, 414–415.
- Newby, G.A., Yen, J.S., Woodard, K.J., Mayuranathan, T., Lazzarotto, C. R., Li, Y., Sheppard-Tillman, H., Porter, S.N., Yao, Y., Mayberry, K., et al. (2021). Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature* 595, 295–302.
- Nielsen, A.A.K., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D., and Voigt, C.A. (2016). Genetic circuit design automation. *Science* 352, aac7341.
- Nelson, J.W., Randolph, P.B., Shen, S.P., Everette, K.A., Chen, P.J., Anzalone, A.V., An, M., Newby, G.A., Chen, J.C., Hsu, A., et al. (2022). Engineered pegRNAs improve prime editing efficiency. *Nat Biotechnol* 40, 402–410.
- Neugebauer, M.E., Kissman, E.N., Marchand, J.A., Pelton, J.G., Sambold, N.A., Millar, D.C., and Chang, M.C.Y. (2021). Reaction pathway engineering converts a radical hydroxylase into a halogenase. *Nat Chem Biol* 18, 171–179.
- Nirantar, S.R. (2021). Directed evolution methods for enzyme engineering. *Molecules* 26, 5599–5612.
- Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O.O., Gootenberg, J.S., Mori, H., et al. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* 361, 1259–1262.
- Niu, C., Zhu, L., Xu, X., and Li, Q. (2016). Rational design of thermostability in bacterial 1,3-1,4- $\beta$ -glucanases through spatial compartmentalization of mutational hotspots. *Appl Microbiol Biotechnol* 101, 1085–1097.
- Noireaux, V., Bar-Ziv, R., and Libchaber, A. (2003). Principles of cell-free genetic circuit assembly. *Proc Natl Acad Sci USA* 100, 12672–12677.
- Oberortner, E., Evans, R., Meng, X., Nath, S., Plahar, H., Simirenko, L., Tarver, A., Deutsch, S., Hillson, N.J., and Cheng, J.F. (2020). An integrated computer-aided design and manufacturing workflow for synthetic biology. *Methods Mol Biol* 2205, 3–18.
- Ofran, Y., and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol* 3, e119.
- Organick, L., Ang, S.D., Chen, Y.J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M.Z., Kamath, G., Gopalan, P., Nguyen, B., et al. (2018). Random access in large-scale DNA data storage. *Nat Biotechnol* 36, 242–248.
- Ostrov, N., Landon, M., Guell, M., Kuznetsov, G., Teramoto, J., Cervantes, N., Zhou, M., Singh, K., Napolitano, M.G., Moosburner, M., et al. (2016). Design, synthesis, and testing toward a 57-codon genome. *Science* 353, 819–822.
- Oza, J.P., Aerni, H.R., Pirman, N.L., Barber, K.W., Ter Haar, C.M., Rogulina, S., Amroffell, M.B., Isaacs, F.J., Rinehart, J., and Jewett, M.C. (2015). Robust production of recombinant phosphoproteins using cell-free protein synthesis. *Nat Commun* 6, 8168.
- Özcan, A., Krajewski, R., Ioannidi, E., Lee, B., Gardner, A., Makarova, K.S., Koonin, E.V., Abudayyeh, O.O., and Gootenberg, J.S. (2021). Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. *Nature* 597, 720–725.
- Palluk, S., Arlow, D.H., de Rond, T., Barthel, S., Kang, J.S., Bector, R., Baghdassarian, H.M., Truong, A.N., Kim, P.W., Singh, A.K., et al. (2018). *De novo* DNA synthesis using polymerase-nucleotide conjugates. *Nat Biotechnol* 36, 645–650.
- Panganiban, B., Qiao, B., Jiang, T., DelRe, C., Obadia, M.M., Nguyen, T. D., Smith, A.A.A., Hall, A., Sit, I., Crosby, M.G., et al. (2018). Random heteropolymers preserve protein function in foreign environments. *Science* 359, 1239–1243.
- Panwar, H.S., Ojha, H., Ghosh, P., Raut, S., and Sahu, A. (2016). Domain swapping reveals functional modularity present in the decay-accelerating factor (CD55). *Immunobiology* 221, 1181–1182.
- Pardee, K., Green, A.A., Ferrante, T., Cameron, D.E., DaleyKeyser, A., Yin, P., and Collins, J.J. (2014). Paper-based synthetic gene networks. *Cell* 159, 940–954.
- Pardee, K., Green, A.A., Takahashi, M.K., Braff, D., Lambert, G., Lee, J. W., Ferrante, T., Ma, D., Donghia, N., Fan, M., et al. (2016a). Rapid, low-cost detection of zika virus using programmable biomolecular components. *Cell* 165, 1255–1266.
- Pardee, K., Slomovic, S., Nguyen, P.Q., Lee, J.W., Donghia, N., Burrill, D., Ferrante, T., McSorley, F.R., Furuta, Y., Vernet, A., et al. (2016b). Portable, on-demand biomolecular manufacturing. *Cell* 167, 248–259. e12.
- Park, G., Colot, H.V., Collopy, P.D., Krystofova, S., Crew, C., Ringelberg, C., Litvinkova, L., Altamirano, L., Li, L., Curilla, S., et al. (2011). High-throughput production of gene replacement mutants in *Neurospora crassa*. *Methods Mol Biol* 722, 179–189.
- Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. (2020). CRISPR-Cas $\Phi$  from huge phages is a hypercompact genome editor. *Science* 369, 333–337.
- Perez, J.G., Stark, J.C., and Jewett, M.C. (2016). Cell-free synthetic biology: engineering beyond the cell. *Cold Spring Harb Perspect Biol* 8, a023853.
- Ping, Z., Chen, S., Zhou, G., Huang, X., Zhu, S.J., Zhang, H., Lee, H.H., Lan, Z., Cui, J., Chen, T., et al. (2022). Towards practical and robust DNA-based data archiving using the yin-yang codec system. *Nat Comput Sci* 2, 234–242.
- Ping, Z., Ma, D., Huang, X., Chen, S., Liu, L., Guo, F., Zhu, S.J., and Shen, Y. (2019). Carbon-based archiving: current progress and future prospects of DNA-based data storage. *GigaScience* 8, giz075.
- Qian, S., Clomburg, J.M., and Gonzalez, R. (2019). Engineering *Escherichia coli* as a platform for the *in vivo* synthesis of prenylated aromatics. *Biotechnol Bioeng* 116, 1116–1127.
- Qu, G., Li, A., Acevedo-Rocha, C.G., Sun, Z., and Reetz, M.T. (2020). The crucial role of methodology development in directed evolution of selective enzymes. *Angew Chem Int Ed* 59, 13204–13231.
- Qu, L., Yi, Z., Zhu, S., Wang, C., Cao, Z., Zhou, Z., Yuan, P., Yu, Y., Tian, F., Liu, Z., et al. (2019). Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. *Nat Biotechnol* 37, 1059–1069.
- Quan, J., Saaem, I., Tang, N., Ma, S., Negre, N., Gong, H., White, K.P., and Tian, J. (2011). Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat Biotechnol* 29, 449–452.
- Quan, J., and Tian, J. (2009). Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE* 4, e6441.
- Quan, J., and Tian, J. (2011). Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries.



- Nat Protoc* 6, 242–251.
- Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 44, e107.
- Quijano-Rubio, A., Yeh, H.W., Park, J., Lee, H., Langan, R.A., Boyken, S. E., Lajoie, M.J., Cao, L., Chow, C.M., Miranda, M.C., et al. (2021). *De novo* design of modular and tunable protein biosensors. *Nature* 591, 482–487.
- Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* 520, 186–191.
- Radivojević, T., Costello, Z., Workman, K., and Garcia Martin, H. (2020). A machine learning Automated Recommendation Tool for synthetic biology. *Nat Commun* 11, 4879.
- Rajakumar, P.D., Gowers, G.O.F., Suckling, L., Foster, A., Ellis, T., Kitney, R.L., McClymont, D.W., and Freemont, P.S. (2019). Rapid prototyping platform for *Saccharomyces cerevisiae* using computer-aided genetic design enabled by parallel software and workcell platform development. *SLAS Tech* 24, 291–297.
- Ratner, M. (2014). Celgene wagers on Sutro's cell-free platform to ramp up bispecifics. *Nat Biotechnol* 32, 1175.
- Rayner, S., Brignac, S., Bumeister, R., Belosludtsev, Y., Ward, T., Grant, O., O'Brien, K., Evans, G.A., and Garner, H.R. (1998). MerMade: an oligodeoxyribonucleotide synthesizer for high throughput oligonucleotide production in dual 96-well plates. *Genome Res* 8, 741–747.
- Reautschnig, P., Wahn, N., Wettengel, J., Schulz, A.E., Latifi, N., Vogel, P., Kang, T.W., Pfeiffer, L.S., Zarges, C., Naumann, U., et al. (2022). CLUSTER guide RNAs enable precise and efficient RNA editing with endogenous ADAR enzymes in vivo. *Nat Biotechnol* 40, 759–768.
- Reetz, M.T., Bocola, M., Carballeira, J.D., Zha, D., and Vogel, A. (2005). Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew Chem Int Ed* 44, 4192–4196.
- Reetz, M.T., and Carballeira, J.D. (2007). Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc* 2, 891–903.
- Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., et al. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell* 3, 324–333.
- Reynolds, T.S., Courtney, C.M., Erickson, K.E., Wolfe, L.M., Chatterjee, A., Nagpal, P., and Gill, R.T. (2017). ROS mediated selection for increased NADPH availability in *Escherichia coli*. *Biotechnol Bioeng* 114, 2685–2689.
- Richardson, S.M., Mitchell, L.A., Stracquadiano, G., Yang, K., Dymond, J. S., DiCarlo, J.E., Lee, D., Huang, C.L.V., Chandrasegaran, S., Cai, Y., et al. (2017). Design of a synthetic yeast genome. *Science* 355, 1040–1044.
- Richter, F., and Baker, D. (2013). Computational protein design for synthetic biology. In: *Synthetic Biology*. 101–122.
- Richter, M.F., Zhao, K.T., Eton, E., Lapinaite, A., Newby, G.A., Thuronyi, B.W., Wilson, C., Koblan, L.W., Zeng, J., Bauer, D.E., et al. (2020). Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat Biotechnol* 38, 883–891.
- Rienzo, M., Jackson, S.J., Chao, L.K., Leaf, T., Schmidt, T.J., Navidi, A.H., Nadler, D.C., Ohler, M., and Leavell, M.D. (2021). High-throughput screening for high-efficiency small-molecule biosynthesis. *Metab Eng* 63, 102–125.
- Robinson, C.J., Carbonell, P., Jervis, A.J., Yan, C., Hollywood, K.A., Dunstan, M.S., Currin, A., Swainston, N., Spiess, R., Taylor, S., et al. (2020). Rapid prototyping of microbial production strains for the biomanufacture of potential materials monomers. *Metab Eng* 60, 168–182.
- Ruan, Y., Zhang, R., and Xu, Y. (2022). Directed evolution of maltogenic amylase from bacillus licheniformis R-53: enhancing activity and thermostability improves bread quality and extends shelf life. *Food Chem* 381, 132222–132229.
- Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. (2020). An evolution-based model for designing chorismate mutase enzymes. *Science* 369, 440–445.
- Salehi, A.S.M., Shakalli Tang, M.J., Smith, M.T., Hunt, J.M., Law, R.A., Wood, D.W., and Bundy, B.C. (2017). Cell-free protein synthesis approach to biosensing hTRβ-specific endocrine disruptors. *Anal Chem* 89, 3395–3401.
- Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365.
- Schreuder, M.P., Brekelmans, S., van den Ende, H., and Klis, F.M. (1993). Targeting of a heterologous protein to the cell wall of *Saccharomyces cerevisiae*. *Yeast* 9, 399–409.
- Schwarz, K.A., Daringer, N.M., Dolberg, T.B., and Leonard, J.N. (2017). Rewiring human cellular input-output using modular extracellular sensors. *Nat Chem Biol* 13, 202–209.
- Scott, M., and Hwa, T. (2011). Bacterial growth laws and their applications. *Curr Opin Biotechnol* 22, 559–565.
- Sekiya, T., Takeya, T., Brown, E.L., Belagaje, R., Contreras, R., Fritz, H.J., Gait, M.J., Lees, R.G., Ryan, M.J., Khorana, H.G., et al. (1979). Total synthesis of a tyrosine suppressor transfer RNA gene. XVI. Enzymatic joinings to form the total 207-base pair-long DNA. *J Biol Chem* 254, 5787–5800.
- Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.
- Senoussi, A., Galas, J.C., and Estevez-Torres, A. (2021). Programmed mechano-chemical coupling in reaction-diffusion active matter. *Sci Adv* 7, eabi9865.
- Sesterhenn, F., Yang, C., Bonet, J., Cramer, J.T., Wen, X., Wang, Y., Chiang, C.I., Abriata, L.A., Kucharska, I., Castoro, G., et al. (2020). *De novo* protein design enables the precise induction of rsv-neutralizing antibodies. *Science* 368, eaay5051.
- Shao, Y., Lu, N., Cai, C., Zhou, F., Wang, S., Zhao, Z., Zhao, G., Zhou, J. Q., Xue, X., and Qin, Z. (2019). A single circular chromosome yeast. *Cell Res* 29, 87–89.
- Shao, Y., Wang, Q., Yu, Y., Wang, P., Yuan, J., and Fan, X. (2019). Modification of protease via click chemistry and its application. *J Biol* 36, 100–103.
- Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L.L., Zhou, F., Xiao, S., Liu, L., Zeng, X., et al. (2018). Creating a functional single-chromosome yeast. *Nature* 560, 331–335.
- Shao, Z., Zhao, H., and Zhao, H. (2009). DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* 37, e16.
- Shapland, E.B., Holmes, V., Reeves, C.D., Sorokin, E., Durot, M., Platt, D., Allen, C., Dean, J., Serber, Z., Newman, J., et al. (2015). Low-cost, high-throughput sequencing of DNA assemblies using a highly multiplexed Nextera process. *ACS Synth Biol* 4, 860–866.
- Shelby, M.L., Gilbille, D., Grant, T.D., Bauer, W.J., Segelke, B., He, W., Evans, A.C., Crespo, N., Fischer, P., Pakendorf, T., et al. (2020). Crystallization of ApoA1 and ApoE4 nanolipoprotein particles and initial XFEL-based structural studies. *Crystals* 10, 886.
- Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31, 64–68.
- Shen, Y., Wang, Y., Chen, T., Gao, F., Gong, J., Abramczyk, D., Walker, R., Zhao, H., Chen, S., Liu, W., et al. (2017). Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. *Science* 355, eaaf4791.
- Sheng, J., Huang, L., Zhu, X., Cai, J., and Xu, Z. (2014). Reconstitution of the peptidoglycan cytoplasmic precursor biosynthetic pathway in cell-free system and rapid screening of antisense oligonucleotides for Mur enzymes. *Appl Microbiol Biotechnol* 98, 1785–1794.

- Shetty, R.P., Endy, D., and Knight Jr, T.F. (2008). Engineering BioBrick vectors from BioBrick parts. *J Biol Eng* 2, 5.
- Shin, J., and Noireaux, V. (2012). An *E. coli* cell-free expression toolbox: application to synthetic gene circuits and artificial cells. *ACS Synth Biol* 1, 29–41.
- Shin, J.E., Riesselman, A.J., Kollasch, A.W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A.C., and Marks, D.S. (2021). Protein design and variant prediction using autoregressive generative models. *Nat Commun* 12, 2403.
- Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547, 345–349.
- Shozen, N., Iijima, I., and Hohsaka, T. (2009). Site-specific incorporation of PEGylated amino acids into proteins using nonnatural amino acid mutagenesis. *Bioorg Medicinal Chem Lett* 19, 4909–4911.
- Siegal-Gaskins, D., Tuza, Z.A., Kim, J., Noireaux, V., and Murray, R.M. (2014). Gene circuit performance characterization and resource usage in a cell-free “breadboard”. *ACS Synth Biol* 3, 416–425.
- Si, L., Xu, H., Zhou, X., Zhang, Z., Tian, Z., Wang, Y., Wu, Y., Zhang, B., Niu, Z., Zhang, C., et al. (2016). Generation of influenza A viruses as live but replication-incompetent virus vaccines. *Science* 354, 1170–1173.
- Si, T., Chao, R., Min, Y., Wu, Y., Ren, W., and Zhao, H. (2017a). Automated multiplex genome-scale engineering in yeast. *Nat Commun* 8, 15187.
- Si, T., Li, B., Comi, T.J., Wu, Y., Hu, P., Wu, Y., Min, Y., Mitchell, D.A., Zhao, H., and Sweedler, J.V. (2017b). Profiling of microbial colonies for high-throughput engineering of multistep enzymatic reactions via optically guided matrix-assisted laser desorption/ionization mass spectrometry. *J Am Chem Soc* 139, 12466–12473.
- Si, T., Xue, P., Choe, K., Zhao, H., and Sweedler, J. (2021). High-throughput mass spectrometry complements protein engineering. In: *Protein Engineering: Tools and Applications*, H. Zhao, S.Y. Lee, J. Nielsen and G. Stephanopoulos, ed. (Hoboken: Wiley), pp. 57–79.
- Silva, D.A., Yu, S., Ulge, U.Y., Spangler, J.B., Jude, K.M., Labão-Almeida, C., Ali, L.R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., et al. (2019). *De novo* design of potent and selective mimics of IL-2 and IL-15. *Nature* 565, 186–191.
- Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., and Pâques, F. (2011). Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *CGT* 11, 11–27.
- Silverman, A.D., Karim, A.S., and Jewett, M.C. (2020). Cell-free gene expression: an expanded repertoire of applications. *Nat Rev Genet* 21, 151–170.
- Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228, 1315–1317.
- Smith, H.O., Hutchison Clyde A., I., Pfannkoch, C., and Venter, J.C. (2003). Generating a synthetic genome by whole genome assembly:  $\phi$ X174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci USA* 100, 15440–15445.
- Sokolosky, J.T., Trotta, E., Parisi, G., Picton, L., Su, L.L., Le, A.C., Chhabra, A., Silveria, S.L., George, B.M., King, I.C., et al. (2018). Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. *Science* 359, 1037–1042.
- Stemmer, W.P.C., Cramer, A., Ha, K.D., Brennan, T.M., and Heyneker, H. L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164, 49–53.
- Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S., and Hasty, J. (2008). A fast, robust and tunable synthetic gene oscillator. *Nature* 456, 516–519.
- Suckling, L., McFarlane, C., Sawyer, C., Chambers, S.P., Kitney, R.I., McClymont, D.W., and Freemont, P.S. (2019). Miniaturisation of high-throughput plasmid DNA library preparation for next-generation sequencing using multifactorial optimisation. *Synth Syst Biotechnol* 4, 57–66.
- SunSpiral, V., Fredlund, J., Abdulla, H., Boccazzi, P., Poust, S., Cleto, S.d. L.A., Chaikind, B., Vaughan, D., Bruno, K.S., Westfall, P., et al. (2022). HTP genomic engineering platform for improving fungal strains. US Patent, 11242524.
- Tabatabaei, S.K., Wang, B., Athreya, N.B.M., Enghiad, B., Hernandez, A. G., Fields, C.J., Leburton, J.P., Soloveichik, D., Zhao, H., and Milenkovic, O. (2020). DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nat Commun* 11, 1742.
- Takahashi, C.N., Nguyen, B.H., Strauss, K., and Ceze, L. (2019). Demonstration of end-to-end automation of DNA data storage. *Sci Rep* 9, 4998.
- Takahashi, M.K., Chappell, J., Hayes, C.A., Sun, Z.Z., Kim, J., Singhal, V., Spring, K.J., Al-Khabouri, S., Fall, C.P., Noireaux, V., et al. (2015a). Rapidly characterizing the fast dynamics of RNA genetic circuitry with cell-free transcription-translation (TX-TL) systems. *ACS Synth Biol* 4, 503–515.
- Takahashi, M.K., Hayes, C.A., Chappell, J., Sun, Z.Z., Murray, R.M., Noireaux, V., and Lucks, J.B. (2015b). Characterizing and prototyping genetic networks with cell-free transcription-translation reactions. *Methods* 86, 60–72.
- Takahashi, M.K., Tan, X., Dy, A.J., Braff, D., Akana, R.T., Furuta, Y., Donghia, N., Ananthakrishnan, A., and Collins, J.J. (2018). A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers. *Nat Commun* 9, 3347.
- Tan, Y., Zhang, Y., Han, Y., Liu, H., Chen, H., Ma, F., Withers, S.G., Feng, Y., and Yang, G. (2019). Directed evolution of an  $\alpha$ 1,3-fucosyltransferase using a single-cell ultrahigh-throughput screening method. *Sci Adv* 5, eaaw8451.
- Tan, Z., Li, X., Shi, H., Yin, X., Zhu, X., Bilal, M., and Onchari, M.M. (2022). Enhancing the methanol tolerance of candida antarctica lipase b by saturation mutagenesis for biodiesel preparation. *3 Biotech* 12, 22.
- Tang, T., Fu, L., Guo, E., Zhang, Z., Wang, Z., Ma, C., Zhang, Z., Zhang, J., Huang, J., and Si, T. (2021). Automation in synthetic biology using biological foundries (in Chinese). *Chin Sci Bull* 66, 300–309.
- Tang, W., and Liu, D.R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* 360, eaap8992.
- Thavarajah, W., Silverman, A.D., Verosloff, M.S., Kelley-Loughnane, N., Jewett, M.C., and Lucks, J.B. (2019). Point-of-use detection of environmental fluoride via a cell-free riboswitch-based biosensor. *ACS Synth Biol* 9, 10–18.
- Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. (2004). Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432, 1050–1054.
- Tian, L., Li, M., Patil, A.J., Drinkwater, B.W., and Mann, S. (2019). Artificial morphogen-mediated differentiation in synthetic protocells. *Nat Commun* 10, 3321.
- Tinafar, A., Jaenes, K., and Pardee, K. (2019). Synthetic biology goes cell-free. *BMC Biol* 17, 64.
- Toda, S., McKeithan, W.L., Hakkinen, T.J., Lopez, P., Klein, O.D., and Lim, W.A. (2020). Engineering synthetic morphogen systems that can program multicellular patterning. *Science* 370, 327–331.
- Tong, F., Qin, Z., Wang, H., Jiang, Y., Li, J., Ming, H., Qu, G., Xiao, Y., and Sun, Z. (2022). Biosynthesis of chiral amino alcohols via an engineered amine dehydrogenase in *E. coli*. *Front Bioeng Biotechnol* 9, 778584–778594.
- Tu, R., Zhang, Y., Hua, E., Bai, L., Huang, H., Yun, K., and Wang, M. (2021). Droplet-based microfluidic platform for high-throughput screening of Streptomyces. *Commun Biol* 4, 647.
- Turing, A.M. (1952). The chemical basis of morphogenesis. *Phil Trans R Soc Lond B* 237, 37–72.
- Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 435, 646–651.
- Valderrama-Rincon, J.D., Fisher, A.C., Merritt, J.H., Fan, Y.Y., Reading, C. A., Chhiba, K., Heiss, C., Azadi, P., Aebi, M., and DeLisa, M.P. (2012). An engineered eukaryotic protein glycosylation pathway in *Escherichia*

- coli*. *Nat Chem Biol* 8, 434–436.
- Valliere, M.A., Korman, T.P., Woodall, N.B., Khitrov, G.A., Taylor, R.E., Baker, D., and Bowie, J.U. (2019). A cell-free platform for the prenylation of natural products and application to cannabinoid production. *Nat Commun* 10, 565.
- Van Brempt, M., Clauwaert, J., Mey, F., Stock, M., Maertens, J., Waegeman, W., and De Mey, M. (2020). Predictive design of sigma factor-specific promoters. *Nat Commun* 11, 5822.
- Veit, A., Akhtar, M.K., Mizutani, T., and Jones, P.R. (2008). Constructing and testing the thermodynamic limits of synthetic NAD(P)H:H<sub>2</sub> pathways. *Microb Biotechnol* 1, 382–394.
- Venetz, J.E., Del Medico, L., Wöflle, A., Schächle, P., Bucher, Y., Appert, D., Tschan, F., Flores-Tinoco, C.E., van Kooten, M., Guennoun, R., et al. (2019). Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proc Natl Acad Sci USA* 116, 8070–8079.
- Verhoeven, K.D., Altstadt, O.C., and Savinov, S.N. (2012). Intracellular detection and evolution of site-specific proteases using a genetic selection system. *Appl Biochem Biotechnol* 166, 1340–1354.
- Venter, J.C., Glass, J.I., Hutchison III, C.A., and Vashee, S. (2022). Synthetic chromosomes, genomes, viruses, and cells. *Cell* 185, 2708–2724.
- Verosloff, M., Chappell, J., Perry, K.L., Thompson, J.R., and Lucks, J.B. (2019). PLANT-Dx: a molecular diagnostic for point-of-use detection of plant pathogens. *ACS Synth Biol* 8, 902–905.
- Voigt, C.A. (2020). Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nat Commun* 11, 6379.
- Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., Panico, M., Morris, H.R., Dell, A., Wren, B.W., et al. (2002). N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* 298, 1790–1793.
- Walsh III, D.I., Pavan, M., Ortiz, L., Wick, S., Bobrow, J., Guido, N.J., Leinicke, S., Fu, D., Pandit, S., Qin, L., et al. (2019). Standardizing automated DNA assembly: best practices, metrics, and protocols using robots. *SLAS Tech* 24, 282–290.
- Walton, R.T., Christie, K.A., Whittaker, M.N., and Kleinstiver, B.P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* 368, 290–296.
- Wang, D.B., Zheng, K.Q., Qiu, M.S., Liang, Z.H., Wu, R.L., Chen, C.Q., Wang, E.B., Zhu, Y.S., Shen, Q.X., Yu, Y.H., et al. (1983). Total synthesis of yeast alanine transfer ribonucleic acid. *Sci Sin Ser B-Chem Biol Agric Med Earth Scis* 26, 464–481.
- Wang, H.H., and Church, G.M. (2011). Multiplexed genome engineering and genotyping methods: applications for synthetic biology and metabolic engineering. *Methods Enzymol* 498, 409–426.
- Wang, J., He, Z., Wang, G., Zhang, R., Duan, J., Gao, P., Lei, X., Qiu, H., Zhang, C., Zhang, Y., et al. (2022). Efficient targeted insertion of large DNA fragments without DNA donors. *Nat Methods* 19, 331–340.
- Wang, J., Li, G., Li, Z., Tang, C., Feng, Z., An, H., Liu, H., Liu, T., and Li, C. (2017). A highly selective and stable ZnO-ZrO<sub>2</sub> solid solution catalyst for CO<sub>2</sub> hydrogenation to methanol. *Sci Adv* 3, e1701290.
- Wang, K., Fredens, J., Brunner, S.F., Kim, S.H., Chia, T., and Chin, J.W. (2016). Defining synonymous codon compression schemes by genome recoding. *Nature* 539, 59–64.
- Wang, L., Cao, K., Pedrosa, M.M., Wu, B., Gao, Z., He, B., and Schenk, G. (2021a). Sequence- and structure-guided improvement of the catalytic performance of a gh11 family xylanase from *Bacillus subtilis*. *J Biol Chem* 297, 101262.
- Wang, Y., Liu, Y., Liu, J., Guo, Y., Fan, L., Ni, X., Zheng, X., Wang, M., Zheng, P., Sun, J., et al. (2018). MACBETH: Multiplex automated *Corynebacterium glutamicum* base editing method. *Metab Eng* 47, 200–210.
- Wang, X., Corin, K., Baaske, P., Wienken, C.J., Jerabek-Willemsen, M., Duhr, S., Braun, D., and Zhang, S. (2011). Peptide surfactants for cell-free production of functional G protein-coupled receptors. *Proc Natl Acad Sci USA* 108, 9049–9054.
- Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., and Wang, X. (2020). Synthetic promoter design in *Escherichia coli* based on a deep generative network. *Nucleic Acids Res* 48, 6403–6412.
- Wang, Y., Wang, P., and Qin, J.H. (2021b). Microfluidic organs-on-a-chip for modeling human infectious diseases. *Acc Chem Res* 54, 3550–3562.
- Weber, H., and Khorana, H.G. (1972). CIV. Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast. Chemical synthesis of an icosadeoxynucleotide corresponding to the nucleotide sequence 21 to 40. *J Mol Biol* 72, 219–249.
- Wei, X., Meng, D., and You, C. (2020). *In vitro* metabolic engineering: current status and recent progress. In: *Systems and Synthetic Metabolic Engineering Systems and Synthetic Metabolic Engineering 2020*. 183–206.
- Weidmann, J., Schnölzer, M., Dawson, P.E., and Hoheisel, J.D. (2019). Copying life: synthesis of an enzymatically active mirror-image DNA-ligase made of D-amino acids. *Cell Chem Biol* 26, 645–651.e3.
- Welch, P., and Scopes, R.K. (1985). Studies on cell-free metabolism: ethanol production by a yeast glycolytic system reconstituted from purified enzymes. *J Biotechnol* 2, 257–273.
- Wen, K.Y., Cameron, L., Chappell, J., Jensen, K., Bell, D.J., Kelwick, R., Kopniczky, M., Davies, J.C., Filloux, A., and Freemont, P.S. (2017). A cell-free biosensor for detecting quorum sensing molecules in *P. aeruginosa*-infected respiratory samples. *ACS Synth Biol* 6, 2293–2301.
- Wilbraham, L., Mehr, S.H.M., and Cronin, L. (2021). Digitizing chemistry using the chemical processing unit: from synthesis to discovery. *Acc Chem Res* 54, 253–262.
- Williams, J.Z., Allen, G.M., Shah, D., Sterin, I.S., Kim, K.H., Garcia, V.P., Shavey, G.E., Yu, W., Puig-Saus, C., Tsoi, J., et al. (2020). Precise T cell recognition programs designed by transcriptionally linking multiple receptors. *Science* 370, 1099–1104.
- Wilson, D.S., Keefe, A.D., and Szostak, J.W. (2001). The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci USA* 98, 3750–3755.
- Win, M.N., and Smolke, C.D. (2008). Higher-order cellular information processing with synthetic RNA devices. *Science* 322, 456–460.
- Wu, I.L., Patterson, M.A., Carpenter Desai, H.E., Mehl, R.A., Giorgi, G., and Conticello, V.P. (2013). Multiple site-selective insertions of noncanonical amino acids into sequence-repetitive polypeptides. *ChemBioChem* 14, 968–978.
- Wu, S., and Bornscheuer, U.T. (2022). A chemoenzymatic cascade with the potential to feed the world and allow humans to live in space. *Eng Microbiol* 2, 100006.
- Wu, Y.R., Zhang, M., Zhong, M., and Hu, Z. (2017). Synergistic enzymatic saccharification and fermentation of agar for biohydrogen production. *Bioresource Tech* 241, 369–373.
- Wu, Z., Zhang, Y., Yu, H., Pan, D., Wang, Y., Wang, Y., Li, F., Liu, C., Nan, H., Chen, W., et al. (2021). Programmed genome editing by a miniature CRISPR-Cas12f nuclease. *Nat Chem Biol* 17, 1132–1138.
- Wuu, J.J., and Swartz, J.R. (2008). High yield cell-free production of integral membrane proteins without refolding or detergents. *Biochim Biophys Acta (BBA)-Biomembranes* 1778, 1237–1250.
- Xu, Y., Lin, K., Wang, S., Wang, L., Cai, C., Song, C., Lai, L., and Pei, J. (2019). Deep learning for molecular generation. *Future Medicinal Chem* 11, 567–597.
- Xu, X., Chemparathy, A., Zeng, L., Kempton, H.R., Shang, S., Nakamura, M., and Qi, L.S. (2021a). Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol Cell* 81, 4333–4345.e4.
- Xu, C., Zhou, Y., Xiao, Q., He, B., Geng, G., Wang, Z., Cao, B., Dong, X., Bai, W., Wang, Y., et al. (2021b). Programmable RNA editing with compact CRISPR-Cas13 systems from uncultivated microbes. *Nat Methods* 18, 499–506.
- Xue, P., Si, T., Mishra, S., Zhang, L., Choe, K., Sweedler, J.V., and Zhao, H. (2020). A mass spectrometry-based high-throughput screening method for engineering fatty acid synthases with improved production of medium-chain fatty acids. *Biotechnol Bioeng* 117, 2131–2138.
- Xie, Y., An, J., Yang, G., Wu, G., Zhang, Y., Cui, L., and Feng, Y. (2014). Enhanced enzyme kinetic stability by increasing rigidity within the



- active site. *J Biol Chem* 289, 7994–8006.
- Xie, Z.X., Li, B.Z., Mitchell, L.A., Wu, Y., Qi, X., Jin, Z., Jia, B., Wang, X., Zeng, B.X., Liu, H.M., et al. (2017). “Perfect” designer chromosome V and behavior of a ring derivative. *Science* 355, eaaf4704.
- Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., and Liu, H. (2014). Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat Commun* 5, 5330.
- Yang, D., Park, S.Y., Park, Y.S., Eun, H., and Lee, S.Y. (2020). Metabolic engineering of *Escherichia coli* for natural product biosynthesis. *Trends Biotechnol* 38, 745–765.
- Yang, G., and Withers, S.G. (2009). Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *ChemBioChem* 10, 2704–2715.
- Yeoh, J.W., Jayaraman, S.S.O., Tan, S.G.D., Jayaraman, P., Holowko, M. B., Zhang, J., Kang, C.W., Leo, H.L., and Poh, C.L. (2021). A model-driven approach towards rational microbial bioprocess optimization. *Biotechnol Bioeng* 118, 305–318.
- Yi, Z., Qu, L., Tang, H., Liu, Z., Liu, Y., Tian, F., Wang, C., Zhang, X., Feng, Z., Yu, Y., et al. (2022). Engineered circular ADAR-recruiting RNAs increase the efficiency and fidelity of RNA editing in vitro and in vivo. *Nat Biotechnol* 40, 946–955.
- Yin, G., Garces, E.D., Yang, J., Zhang, J., Tran, C., Steiner, A.R., Roos, C., Bajad, S., Hudak, S., Penta, K., et al. (2012). Aglycosylated antibodies and antibody fragments produced in a scalable in vitro transcription-translation system. *mAbs* 4, 217–225.
- Yoshida, R. (2010). Self-oscillating gels driven by the Belousov-Zhabotinsky reaction as novel smart materials. *Adv Mater* 22, 3463–3483.
- You, C., Chen, H., Myung, S., Sathitsuksanoh, N., Ma, H., Zhang, X.Z., Li, J., and Zhang, Y.H.P. (2013). Enzymatic transformation of nonfood biomass to starch. *Proc Natl Acad Sci USA* 110, 7182–7187.
- You, C., Shi, T., Li, Y., Han, P., Zhou, X., and Zhang, Y.H.P. (2017). An in vitro synthetic biology platform for the industrial biomanufacturing of myo-inositol from starch. *Biotechnol Bioeng* 114, 1855–1864.
- Yu, G., Zhang, M., Gao, L., Zhou, Y., Qiao, L., Yin, J., Wang, Y., Zhou, J., and Ye, H. (2022). Far-red light-activated human islet-like designer cells enable sustained fine-tuned secretion of insulin for glucose control. *Mol Ther* 30, 341–354.
- Zaccolo, M., and Gherardi, E. (1999). The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1  $\beta$ -lactamase 1 Edited by A. R. Fersht. *J Mol Biol* 285, 775–783.
- Zhang, B., Korolj, A., Lai, B.F.L., and Radisic, M. (2018). Advances in organ-on-a-chip engineering. *Nat Rev Mater* 3, 257–278.
- Zhang, J., Petersen, S.D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., Sánchez, B.J., Costello, Z., Chen, Y., Fero, M.J., et al. (2020a). Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat Commun* 11, 4880.
- Zhang, J., Chen, Y., Fu, L., Guo, E., Wang, B., Dai, L., and Si, T. (2021a). Accelerating strain engineering in biofuel research via build and test automation of synthetic biology. *Curr Opin Biotechnol* 67, 88–98.
- Zhang, Q., Hresko, M.E., Picton, L.K., Su, L., Hollander, M.J., Nunez-Cruz, S., Zhang, Z., Assenmacher, C.A., Sockolosky, J.T., Garcia, K.C., et al. (2021b). A human orthogonal IL-2 and IL-2R $\beta$  system enhances CAR T cell expansion and antitumor activity in a murine model of leukemia. *Sci Transl Med* 13, eabg6986.
- Zhang, R., Chen, S., Meng, X., Chai, Z., Wang, D., Yuan, Y., Chen, K., Jiang, L., Li, J., and Gao, C. (2021c). Generating broad-spectrum tolerance to ALS-inhibiting herbicides in rice by base editing. *Sci China Life Sci* 64, 1624–1633.
- Zhang, R., Liu, J., Chai, Z., Chen, S., Bai, Y., Zong, Y., Chen, K., Li, J., Jiang, L., and Gao, C. (2019a). Generation of herbicide tolerance traits and a new selectable marker in wheat using base editing. *Nat Plants* 5, 480–485.
- Zhang, W., Mitchell, L.A., Bader, J.S., and Boeke, J.D. (2020b). Synthetic genomes. *Annu Rev Biochem* 89, 77–101.
- Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., Wang, A., Jiang, S., Jiang, Q., Gong, J., et al. (2017a). Engineering the ribosomal DNA in a megabase synthetic chromosome. *Science* 355, eaaf3981.
- Zhang, X.E. (2018). Preface to special issue of synthetic biology (in Chinese). *Bull Chin Acad Sci* 33, 1132–1134.
- Zhang, X.E. (2019). Synthetic biology in China: review and prospects (in Chinese). *Sci Sin-Vitae* 49, 1543–1572.
- Zhang, Y.-P. (2013). Next generation biorefineries will solve the food, biofuels, and environmental trilemma in the energy-food-water nexus. *Energy Sci Eng* 1, 27–41.
- Zhang, Y., Minagawa, Y., Kizoe, H., Miyazaki, K., Iino, R., Ueno, H., Tabata, K.V., Shimane, Y., and Noji, H. (2019b). Accurate high-throughput screening based on digital protein synthesis in a massively parallel femtoliter droplet array. *Sci Adv* 5, eaav8185.
- Zhang, Y., Ptacin, J.L., Fischer, E.C., Aerni, H.R., Caffaro, C.E., San Jose, K., Feldman, A.W., Turner, C.R., and Romesberg, F.E. (2017b). A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* 551, 644–647.
- Zhang, Z.B., Wang, Q.Y., Ke, Y.X., Liu, S.Y., Ju, J.Q., Lim, W.A., Tang, C., and Wei, P. (2017c). Design of tunable oscillatory dynamics in a synthetic NF- $\kappa$ B signaling circuit. *Cell Syst* 5, 460–470.e5.
- Zhao, G.P. (2018). Synthetic biology: unsealing the convergence era of life science research (in Chinese). *Bull Chin Acad Sci* 33, 1135–1149.
- Zhao, H., Ding, W., Zang, J., Yang, Y., Liu, C., Hu, L., Chen, Y., Liu, G., Fang, Y., Yuan, Y., et al. (2021). Directed-evolution of translation system for efficient unnatural amino acids incorporation and generalizable synthetic auxotroph construction. *Nat Commun* 12, 7039–7050.
- Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 163, 759–771.
- Zheng, H., Bai, Y., Jiang, M., Tokuyasu, T.A., Huang, X., Zhong, F., Wu, Y., Fu, X., Kleckner, N., Hwa, T., et al. (2020). General quantitative relations linking cell growth and the cell cycle in *Escherichia coli*. *Nat Microbiol* 5, 995–1001.
- Zhirnov, V., Zadegan, R.M., Sandhu, G.S., Church, G.M., and Hughes, W. L. (2016). Nucleic acid memory. *Nat Mater* 15, 366–370.
- Zhou, F., Mu, X., Nie, Y., and Xu, Y. (2021). Enhanced catalytic efficiency and coenzyme affinity of leucine dehydrogenase by comprehensive screening strategy for L-tert-leucine synthesis. *Appl Microbiol Biotechnol* 105, 3625–3634.
- Zhou, J., Huang, L., Lian, J., Sheng, J., Cai, J., and Xu, Z. (2010). Reconstruction of the UDP-N-acetylglucosamine biosynthetic pathway in cell-free system. *Biotechnol Lett* 32, 1481–1486.
- Zhu, Y., Yin, X., Liu, H., Li, H., Chen, Y., Li, L., Xiao, A., and Ni, H. (2019). Substitution of his260 residue alters the thermostability of pseudoalteromonas carrageenovora arylsulfatase. *Acta Oceanol Sin* 38, 75–82.
- Zhu, Z., Kin Tam, T., Sun, F., You, C., and Percival Zhang, Y.H. (2014). A high-energy-density sugar biobattery based on a synthetic enzymatic pathway. *Nat Commun* 5, 3026.
- Zong, Y., Liu, Y., Xue, C., Li, B., Li, X., Wang, Y., Li, J., Liu, G., Huang, X., Cao, X., et al. (2022). An engineered prime editor with enhanced editing efficiency in plants. *Nat Biotechnol* 40, 1394–1402.
- Zrimec, J., Börlin, C.S., Buric, F., Muhammad, A.S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., and Zeleznik, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun* 11, 6141.