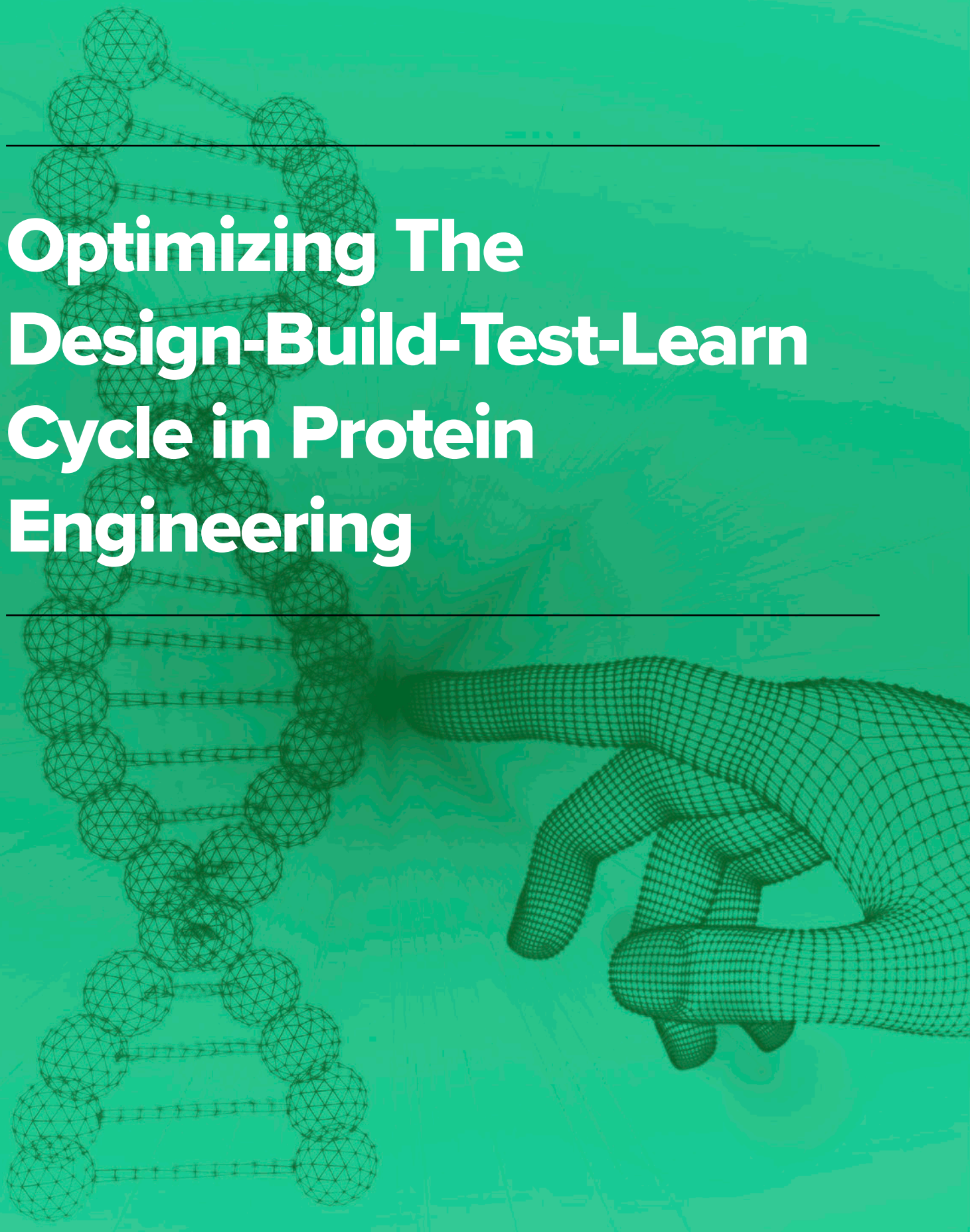
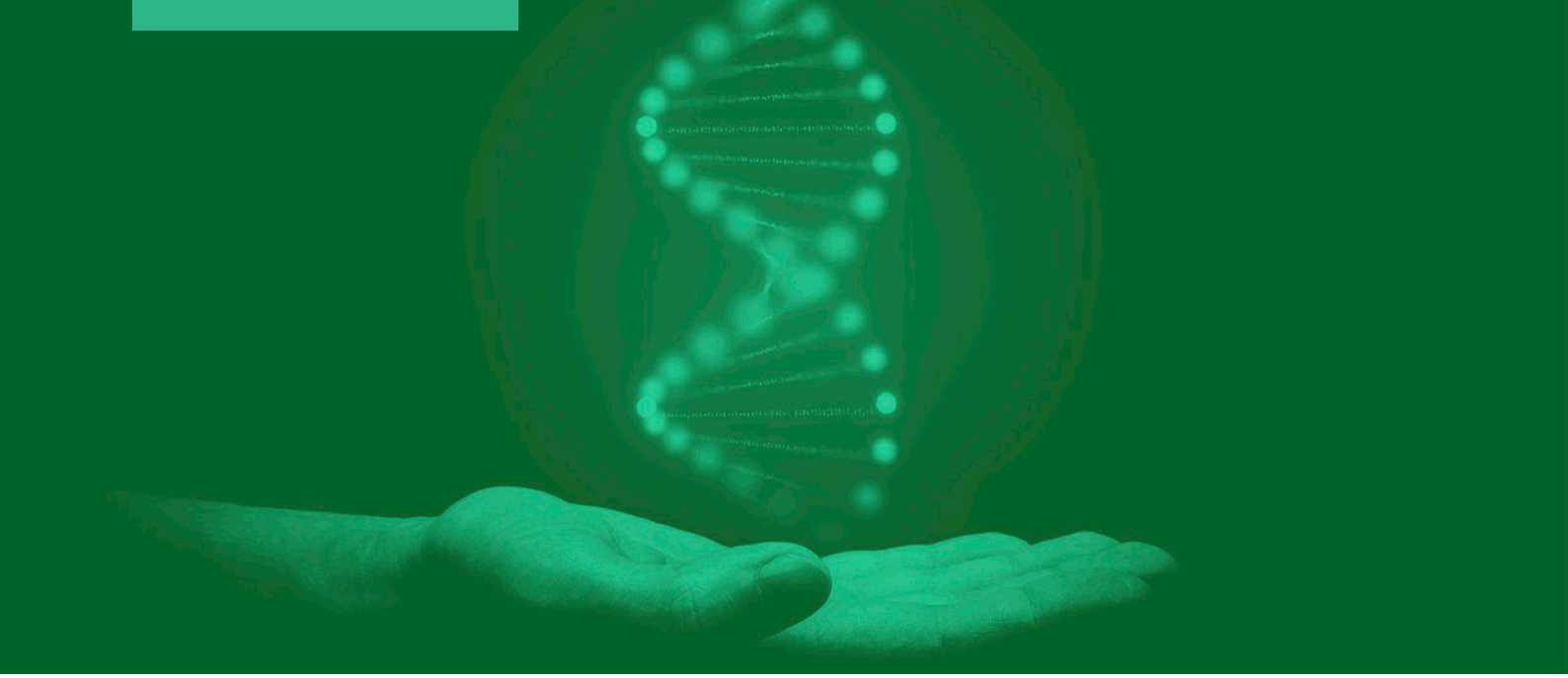
TWIST BIOSCIENCE

Optimizing The Design-Build-Test-Learn Cycle in Protein Engineering



Content Page



Foreword

Proteins are nature's workhorses; they are critically involved in defining and maintaining cellular structure and function in all living organisms. During the last 70 years, researchers have been developing tools that allow them to carefully rework the structure of proteins.

The emergence of new computational methods and high-throughput synthetic DNA has transformed the field of protein engineering, enabling biotechnology applications to scale-up and spread across diverse industries. By harnessing the power of synthetic biology, protein engineering has emerged as a powerful tool for the development of important protein-based tools, including biocatalysts, biosensors, therapeutics, bioremediation and biofuels.

The design-build-test-learn (DBTL) cycle, a cornerstone of synthetic biology, is an iterative process that enables scientists to test different protein sequences and optimize protein discovery. A successful and fast DBTL cycle depends on how quickly and how broadly DNA can be accessed for the test phase. Thus, access to affordable, high-throughput synthetic DNA has allowed researchers to artificially build custom DNA sequences and accelerate their discoveries.

This eBook explores the evolution and applications of protein engineering including the impact of affordable, high-throughput synthetic DNA on DBTL cycle optimization and protein discovery.

CHAPTER 1

Introduction to Protein Engineering

Proteins may not be the heritable units of life, but they are undoubtedly nature's workhorses. Across life's kingdoms, proteins are critically involved in defining and maintaining cellular structure. They manifest and regulate cell signaling as well as carry out the mechanics of life's central dogma. Stated plainly, proteins are powerful biomolecules with the capacity for a wide range of functions.

Researchers have long been interested in manipulating proteins to serve human needs, such as the synthesis of complex chemicals, the metabolism of pollutants, or the formation of advanced therapeutics. But doing so is exceedingly difficult as proteins have evolved over billions of years to serve specific functions under specific conditions. Removing the protein from those conditions, or asking it to perform a different function, requires a careful rework of its structure and underlying DNA sequence. This process is known as protein engineering.

Through the calculated alteration of DNA sequences, protein engineers can create novel proteins capable of carrying out complex and new chemical reactions. Such proteins may be applied to the synthesis of advanced therapeutics, the metabolism of pollutants, or the large-scale production of natural chemicals to name just a few.

However, the process of protein engineering is far from simple. The complex interplay between proteins' many structural elements and their function is still poorly understood. As a result, researchers often must progress through an iterative cycle wherein proteins are designed, tested, and redesigned in order to obtain an optimal protein. Fortunately, recent advances in synthetic biology and artificial intelligence are easing the challenges of protein engineering and opening new possibilities. Here, we briefly introduce protein engineering and some of its many applications.

Why is protein engineering needed?

Proteins have evolved to carry out a myriad of functions that help increase an organism's fitness. This process has generated a wide range of protein types, each with unique abilities. Enzymes, for example, carry out spatially, temporally, and physically controlled reactions that enable the formation of complex chemicals. Such chemicals can be valuable for therapeutic and industrial applications. Yet, they are also exceedingly difficult to replicate through synthetic means. Therefore, much of modern drug development and industrial chemical synthesis relies on harvesting chemicals from natural sources or else blending natural processes with synthetic ones.

However, natural enzymes have evolved to enhance organismal fitness, not research pipelines. Whether in their native host or in a heterologous setting, natural proteins are unlikely to behave optimally, which may mean slow reaction times, low output, or promiscuity (if the proteins work at all). Protein engineering gives researchers a way to evolve proteins for their own needs.

For example, the biosynthesis of various classes of natural products, such as fatty acids, isoprenoids, alkaloids, and flavonoids has been traditionally performed using natural enzymes in both microbial and cell-free systems.[1,2] However, this approach is usually hampered by limited enzyme activity, narrow substrate ranges, poor stabilities and loss of function in heterologous hosts. Protein engineering is used to overcome these limitations by enhancing enzymatic activity of the proteins involved in the biosynthesis of these products.[3] Similarly, engineered enzymes are increasingly being used in industrial processes as a "greener" alternative to chemical catalysts.[3,4]

Protein engineering is particularly useful in the discovery and optimization of biologics such as

antibodies and T cell receptors. The development of these proteins requires multiple steps of optimization to improve safety, efficacy, manufacturability and, ultimately, clinical success. Thanks to recent developments in protein engineering methods, the number of therapeutic antibodies approved for preclinical and clinical use has increased exponentially in the last few years.[5,6]

The power of rational design and synthetic biology

There are three major approaches for creating new proteins: directed evolution, rational design, and *de novo* design.[4] Directed evolution does not require previous knowledge of a protein's 3D structure or mechanism of action. It involves random mutagenesis of the DNA sequence in naturally occurring proteins as well as posterior screening to find proteins with new and desirable properties. On the other hand, rational (or data-driven) design uses available information on the protein structure to modify amino acid residues that are relevant for protein function.[4] Thus, rational design takes advantage of biochemical data, protein structure and molecular modeling data to predict beneficial mutations that can be incorporated by site-specific mutagenesis.[7] The third approach, by contrast, is not based on naturally occurring proteins. Instead, *de novo* design uses computational algorithms to tailor synthetic proteins with sequences unrelated to those found in nature.[8,9]

The increasing availability of protein structures, biochemical data and computational methods has favored the implementation of rational design approaches. In effect, rational protein design accounted for the largest share of the global protein engineering market in recent years.[10] By using information on a protein's 3D structure and function, rational design allows researchers to predict which mutations will be beneficial to incorporate precise and specific properties to the protein (e.g., substrate specificity, pH, or temperature stability). In this context, artificial intelligence and computational modeling can help researchers to identify the sequences of interest that can be targeted.

Both directed evolution and rational design require the creation of DNA libraries – a collection of DNA fragments containing different variants of the protein of interest. Yet, the rational design approach significantly reduces the library size which, in turn reduces the time and effort invested into library screening.[7] Traditional methods for constructing DNA libraries can be expensive and time consuming. However, this has changed with the advent of synthetic DNA technologies. For example, the development of an advanced silicon-based synthesis platform enables researchers to create complex custom libraries using large-scale, highly accurate DNA synthesis.

Together, rational design and synthetic DNA libraries bring success across a broad range of applications. They allow researchers to 1) efficiently build genetic engineering tools (such as different variants of the genome-editing Cas9 protein),[11] 2) create synthetic genetic circuits (i.e., networks of genes that have been engineered to co-opt a host cell's machinery) to produce plant-derived therapeutics on an industrial scale,[12,13] and 3) discover, optimize and produce novel industrial enzymes and biologics.[14,15] Finally, rational design and synthetic biology are being applied to develop potential solutions to environmental challenges.[16]

For example, they are being increasingly used for bioremediation – the production of enzymes that degrade different types of contaminant waste.[16,17] Furthermore, the US Advanced Research Projects Agency (ARPA) is funding different projects aiming to create electrofuels. These are a more efficient type of biofuel manufactured using captured carbon dioxide (CO_2) and hydrogen obtained from sustainable electricity sources.[18] Synthetic biologists, metabolic engineers, and microbiologists are working together to optimize the cellular machinery of microorganisms so that they can efficiently fix CO_2 to produce electrofuels. Ideally, the widespread use of electrofuels would help to move us away from fossil fuels, limit greenhouse gas emissions and reduce demands for land, water and fertilizer traditionally required to produce biofuels.

Conclusion

By harnessing the power of rational design and synthetic biology, protein engineering has emerged as a powerful tool for the development of important protein-based tools, including biocatalysts, biosensors, therapeutics, bioremediation, and biofuels. The next chapter of this eBook will briefly discuss the evolution of protein engineering and present some of the applications of synthetic DNA in more detail. Subsequent chapters will discuss the impact of artificial intelligence in this area of research as well as how synthetic DNA and protein engineering is driving modern drug discovery.

References

1. Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. *Nat Chem*. 2016;8:531-541. doi:10.1038/nchem.2479
2. O'Connor SE. Engineering of secondary metabolism. *Annu Rev Genet*. 2015;49:71-94. doi:10.1146/annurev-genet-120213-092053
3. Li C, Zhang R, Wang J, Wilson LM, Yan Y. Protein engineering for improving and diversifying natural product biosynthesis. *Trends Biotechnol*. 2020;38:729-744. doi:10.1016/j.tibtech.2019.12.008
4. Singh RK, Lee JK, Selvaraj C, *et al*. Protein engineering approaches in the post-genomic era. *Curr Protein Pept Sci*. 2018;19(1):5-15. doi:10.2174/1389203718666161117114244
5. Kaplon H, Reichert JM. Antibodies to watch in 2018. *MAbs*. 2018;10:183-203. doi:10.1080/19420862.2018.1415671
6. Gil, M. Therapeutic antibody engineering: past, present and future. *Technology Networks*. https://www.technologynetworks.com/tn/lists/therapeutic-antibody-engineering-past-present-and-future-354173. Published: September 28, 2021. Accessed: November 21, 2022
7. Steiner K, Schwab H. Recent advances in rational approaches for enzyme engineering. *Comput Struct Biotechnol J*. 2012;2:e201209010. doi:10.5936/csbj.20120901
8. Pan X, Kortemme T. Recent advances in *de novo* protein design: principles, methods, and applications. *J Biol Chem*. 2021;296:100558. doi:10.1016/j.jbc.2021.100558
9. Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. *Nature*. 2016;537:320-327. doi:10.1038/nature19946
10. Protein engineering market report. *Markets and Markets*. https://www.marketsandmarkets.com/Market-Reports/protein-antibody-engineering-market-898.html?gclid=Cj0KCQjwn-4qWBhCvARIsAFNAMijq1N9ZFsHN3mFmlILyWQweSWPNGk-kk9eRQD2fKT1B0oywkTKTIC1oaAlJ9EALw_wcB. Published: January 2020. Accessed: November 16 2022
11. Thean DGL, Chu HY, Fong JHC, *et al*. Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. *Nat Commun*. 2022;13:2219. doi:10.1038/s41467-022-29874-5
12. Srinivasan P, Smolke CD. Biosynthesis of medicinal tropane alkaloids in yeast. *Nature*. 2020;585(7826):614-619. doi:10.1038/s41586-020-2650-9
13. Wu Y, Chen MN, Li S. *De novo* biosynthesis of diverse plant-derived styrylpyrones in *Saccharomyces cerevisiae*. *Metab Eng Commun*. 2022;14:e00195. doi:10.1016/j.mec.2022.e00195
14. Quinn D. Enzyme identification and engineering for novel industrial enzymes. *Twist Bioscience*. https://www.twistbioscience.com/resources/webinar/enzyme-identification-and-engineering-novel-industrial-enzymes. Accessed: November 16 2022
15. The Baker lab at the University of Washington: designing proteins to revolutionize biotechnology. *Twist Bioscience*. https://www.twistbioscience.com/resources/case-study/baker-lab-university-washington-designing-proteins-revolutionize-biotechnology. Accessed: November 16 2022
16. Bell, E.L., Smithson, R., Kilbride, S. *et al.* Directed evolution of an efficient and thermostable PET depolymerase. *Nat Catal*. 2022;5:673–681 (2022). doi:10.1038/s41929-022-00821-3
17. Dutta K, Shityakov S, Khalifa I. New trends in bioremediation technologies toward environment-friendly society: a mini-review. *Front Bioeng Biotechnol*. 2021;9:666858. doi:10.3389/fbioe.2021.666858
18. Hawkins AS, McTernan PM, Lian H, Kelly RM, Adams MW. Biological conversion of carbon dioxide and hydrogen into liquid fuels and industrial chemicals. *Curr Opin Biotechnol*. 2013;24:376-384. doi:10.1016/j.copbio.2013.02.017

CHAPTER 2

From Site Mutagenesis to Synthetic Biology

One could say that the history of protein engineering started almost 70 years ago with the discovery of the structure of DNA in 1953, and Francis Crick's seminal lecture on gene function in 1957.[1,2] The ideas presented in this lecture, widely known as 'the central dogma', still frame how we understand life today, and marked the beginning of a shift towards a better understanding of protein synthesis. Later, between 1961 and 1964, the genetic code was deciphered allowing researchers to understand the relationship between the information contained in DNA and the structure of proteins.[3,4] This discovery marked a milestone in the history of modern biology, enabling the advent of recombinant DNA technologies and the generation of the first recombinant DNA molecules in 1972.[5] Since then, the field of protein engineering has grown immensely, enabling researchers to alter the structure of proteins in a systematic and versatile manner. From site-directed mutagenesis and directed evolution, to contemporary protein engineering rooted in computational tools and synthetic DNA, this chapter will present a brief history of the evolution of protein engineering approaches.

Random vs site directed mutagenesis

Random mutagenesis is a technique wherein mutations are randomly introduced into the DNA of an organism through the use of error-prone PCR, radiation, transposons and mutagenic chemicals. However, scientists can also introduce mutations into a particular location of the DNA molecule using a technique known as site-directed mutagenesis.[6] The development of this technique can be attributed to the work of many researchers in the early 1970s such as Clyde Hutchison, Marshall Edgell, Charles Weissmann and Michael Smith.[7,8] Using site-directed mutagenesis, Michael Smith and his team were able to produce the first mutant DNA in 1978 and large quantities of a mutated enzyme in 1982.[9,10] For this work, Michael Smith was awarded the Nobel Prize in Chemistry 1993, shared with Kary Mullis who developed the polymerase chain reaction (PCR) technique. Together, these two methods laid the foundation for modern protein engineering.

Protein engineering via site-directed mutagenesis is called "rational design" as it relies on information on protein structure to rationally modify selected amino acids that are known to be relevant for protein function (Figure 1). This approach proved to be highly effective to study the role of individual amino acids. However, a limited understanding of protein structure and function during the 1980s and 1990s made it unsuitable for efficiently synthesizing proteins with new properties. Advances in the following decades would prove critical to enabling large-scale protein engineering.

Directed evolution

In the 1990s, the field was transformed by a new approach, termed directed evolution, which does not require previous knowledge of protein structure or function. Instead, this approach attempts to recreate the natural evolutionary processes of variation and selection. Using this technique, researchers can generate diverse combinations of mutations in the genes coding for a protein of interest (the size of the generated gene libraries varies from a few to many million variants). These variants are used to transform the microorganism hosts that will produce the mutant proteins. Finally, the mutant proteins are screened (or selected) for the desired function and the improved proteins are used as the parents for another round of mutations.[6,11] As a result, beneficial mutations are accumulated until the goal is reached or no further improvements are found (Figure 1).

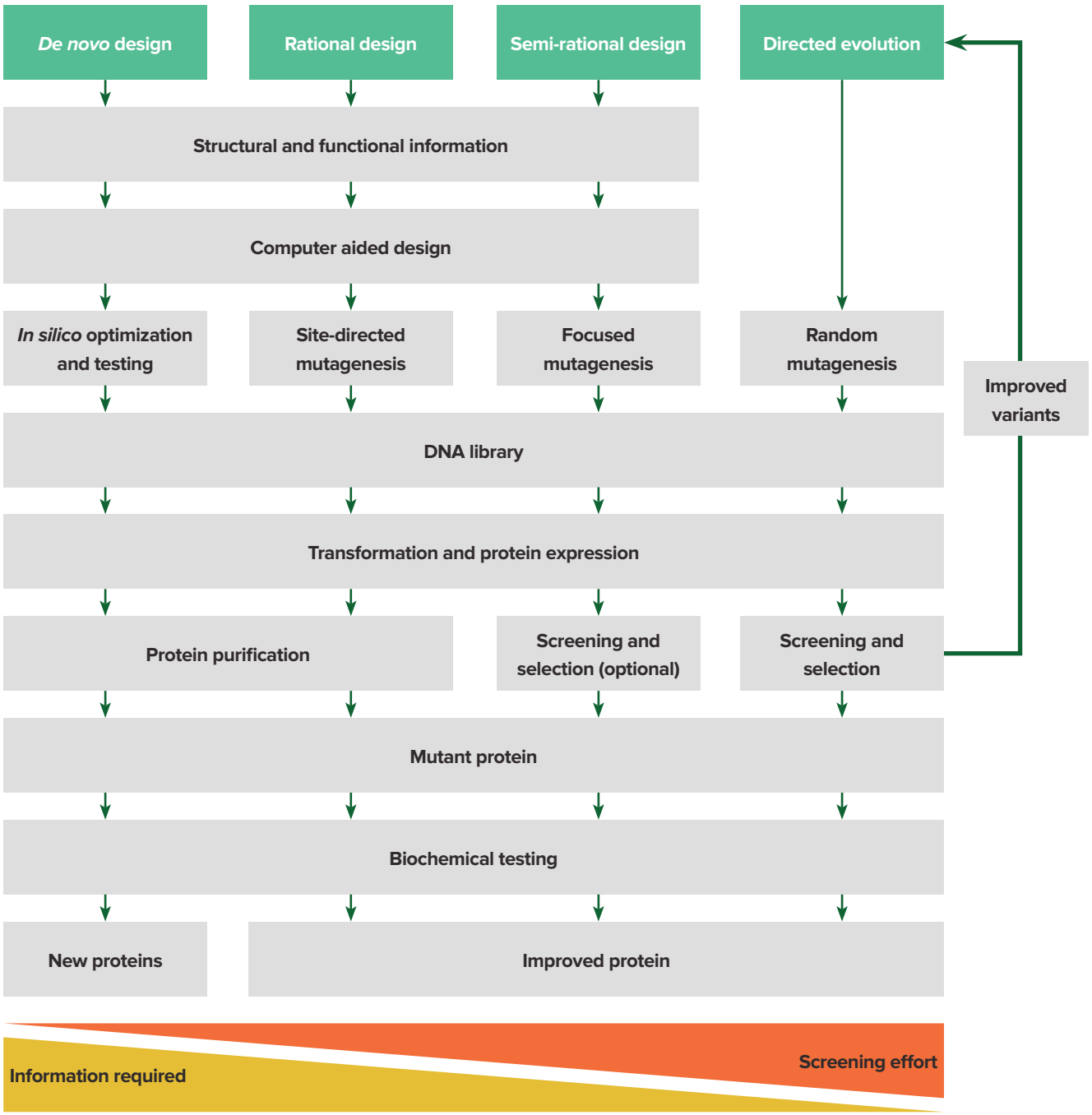


Figure 1: Different approaches to protein engineering (rational design, directed evolution, semi-rational design and *de novo* design)

Directed evolution is based on the pioneering research of many scientists, including Frances Arnold, George Smith and Greg Winter who received the Nobel Prize in Chemistry 2018 for their work in this field.[12,13,14,15] This approach made it possible to engineer new capabilities in proteins of which relatively little is known. For example, it has been used to develop biocatalysts with high specificity, limited side reactions and tolerance for diverse reaction conditions.[16,17,18] These new biocatalysts have applications in various industries including chemical products, biofuels, plastics and biopesticides.[19]

However, directed evolution has some limitations. For example, success requires a significant investment in specific high-throughput assays to screen the large number of mutant proteins generated.[20] Moreover, not all protein functions are readily amenable to development of a high-throughput screening method, nor are all screening methodologies easy to implement at the required scale. In some cases, time poses another limitation as the evolution of certain phenotypes, while theoretically feasible, may occur on time-scales that are not practically feasible.[20] Finally, this approach is unable to thoroughly search the vast

sequence space of a protein; for a typical protein of 300 amino acids, the number of variants containing three simultaneous mutations exceeds 10^{10} – which is often too many to be screened experimentally.[21]

In order to expand the boundaries of protein engineering, researchers needed a way to rationally design protein libraries for directed evolution. Rather than randomly mutating proteins and greatly inflating the number of mutants to be screened, mutagenesis could — in theory — be limited to key portions of the protein that are likely to affect the desired protein function.

Protein engineering in the genomics era

An enormous amount of genomic, structural and functional data has been accrued in recent decades. This, together with the development of computational and artificial intelligence (AI) methods (reviewed in chapter 3), led to a renaissance in the rational design approach. However, its general application has often been hampered by the complexity of a protein's structure/function relationship. Thus, a semi-rational approach combining rational design and directed evolution has been increasingly used to create novel protein functions (Figure 1).[22] In this approach, structural information is used to identify the protein's active site (i.e., amino acids relevant to protein function). Random mutagenesis is then applied only to that particular site, which produces smaller and "smarter" libraries, making the process more efficient.[22]

Interestingly, there are 20^{200} possible amino-acid sequences for a protein with 200 amino acids, yet the number of known naturally occurring proteins is only in the order of 10^{12}.[23] Therefore, sampling natural proteins alone will not allow us to explore the full sequence space accessible to proteins. Guided by the physical principles underlying protein folding, computational modeling and AI tools, the *de novo* approach is able to generate new proteins with sequences unrelated to those found in nature (Figure 1).[23,24] *De novo* design could be considered a category within rational design, as it relies on structural and functional data to model and design novel functions. Moreover, because the proteins being designed do not exist in nature, synthetic genes that encode the novel amino-acid sequences must first be synthesized in the lab.[23] Thus, de *novo* protein design relies heavily on synthetic DNA.

Synthetic DNA

Synthesizing DNA libraries has traditionally been a costly, slow, and error-prone process. The core technology behind current methods, known as phosphoramidite synthesis, was developed in the early 1980s.[25] It creates a DNA molecule by sequentially attaching together nucleotides in a solid support (Figure 2). The first DNA synthesizers were marketed in the late 1980s and could synthesize a single oligonucleotide in a day.[26] Since then, DNA synthesis has improved dramatically, from DNA synthesis on columns – where reactions occur in large separate compartments – to DNA microarray technologies which use glass or silicon supports.[27] The emergence

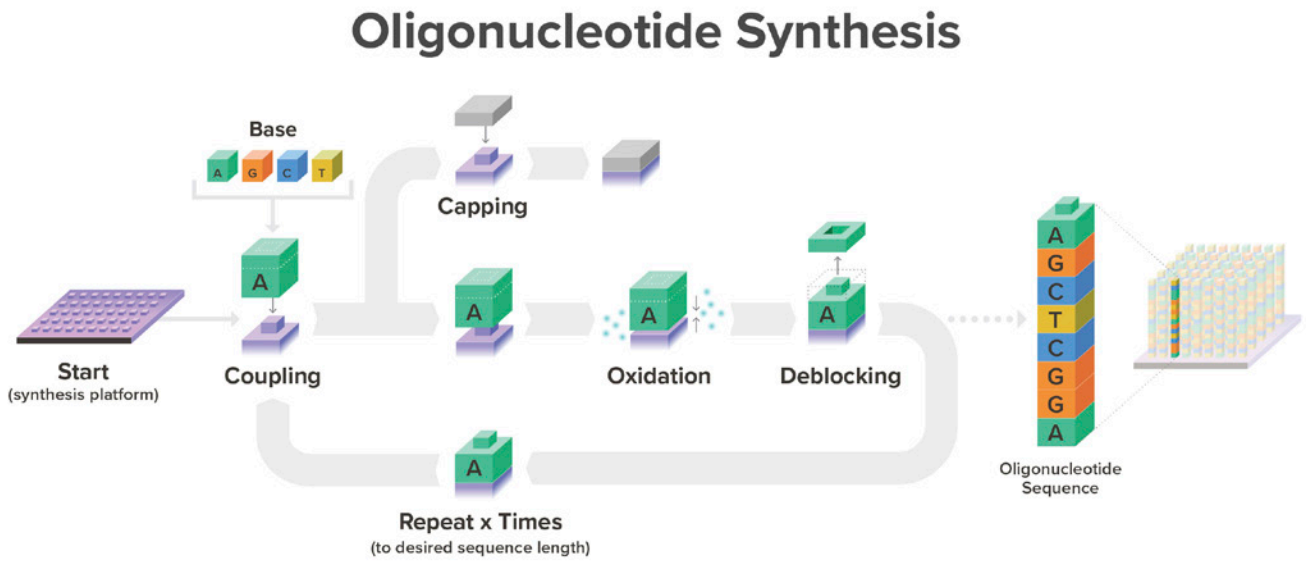


Figure 2: DNA synthesis using cyclic phosphoramidite chemistry.[26]

of DNA microarrays in the 1990s and 2000s allowed researchers to produce oligonucleotides confined to specific spots or volumes using a variety of mechanisms, including photolithography, inject printing, electrochemical array and microfluidics.[28]

Silicon has several properties allowing for the miniaturization of phosphoramidite chemistry, making it an excellent solid support for DNA synthesis. New silicon-powered DNA synthesis harnesses the highly scalable production and processing infrastructure of the semiconductor industry to achieve precision in manufacturing DNA at scale. This technology reduces the reaction volumes by a factor of 1,000,000 while increasing throughput by a factor of 1,000. This enables the synthesis of 9,600 genes on a single silicon chip at full scale compared to traditional synthesis methods which produce a single gene in the same physical space using a 96-well plate. As a result, more DNA can be synthesized on each chip, fewer quantities of reagents are used, and the cost per gene is two to three times lower.

Scaling up protein discovery

The Design-Build-Test-Learn (DBTL) cycle is a general iterative engineering framework that has been adopted by synthetic biology to increase the general efficacy of the protein discovery process (Figure 3).[29] This methodology has, however, several bottlenecks, including the diversity and quality of the DNA produced, the throughput, the turnaround time and the costs to scale up production. High-throughput technologies, such as silicon-based DNA synthesis methods, overcome these hurdles by enabling more precise synthesis of thousands of DNA sequences in parallel. This in turn allows researchers to produce and test a much wider range of variants, greatly speeding up the DBTL cycle.

Access to high volumes of synthetic DNA has transformed the DBTL cycle of biotech companies allowing them to rapidly scale-up production. This is the case of Ginkgo Bioworks, a company that uses engineered microorganisms to manufacture different products for a variety of applications from therapeutics and food to industrial chemicals and biofuels.[30] The company success is possible thanks to the implementation of AI, automatization and access to large amount of high-quality synthetic DNA enabling the acceleration of the DBTL cycle and subsequent process scaling.

Conclusion

The field of protein engineering has witnessed incredible advances since the 1880s. Over the last few decades, the advent of computational methods and affordable, high-throughput synthetic DNA has transformed the field, enabling biotechnology applications to scale-up and spread across diverse industries.

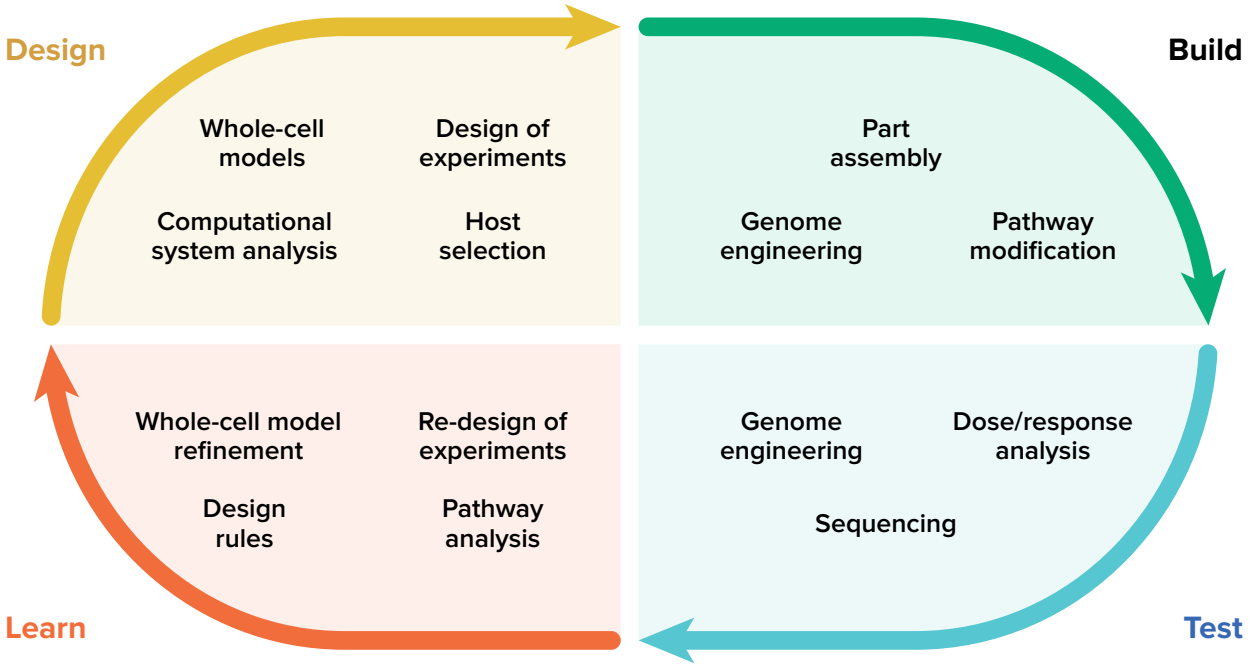


Figure 3: The Design-Build-Test-Learn (DBTL) cycle.

References

1. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 1953;171:737-738. doi:10.1038/171737a0
2. Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol*. 2017;15:e2003243. doi:10.1371/journal.pbio.2003243
3. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. *Nature*. 1961;192:1227-1232. doi:10.1038/1921227a0
4. Nirenberg M, Leder P. RNA codewords and proteins synthesis: the effect of trinucleotides upon the binding of sRNA to ribisomes. *Science*. 1964;145:1399-1407. doi:10.1126/science.145.3639.1399
5. Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli. P Natl A Sci USA*. 1972;69:2904-2909. doi:10.1073/pnas.69.10.2904
6. Lutz S, Iamurri SM. Protein engineering: past, present, and future. *Methods Mol Biol*. 2018;1685:1-12. doi:10.1007/978-1-4939-7366-8_1
7. Hutchison CA 3rd, Edgell MH. Genetic assay for small fragments of bacteriophage phi X174 deoxyribonucleic acid. *J Virol*. 1971;8:181-189. doi:10.1128/JVI.8.2.181-189.1971
8. Weissmann C. The end of the road. *Prion*. 2012;6:97-104. doi:10.4161/pri.19778
9. Hutchison CA 3rd, Phillips S, Edgell MH, Gillam S, Jahnke P, Smith M. Mutagenesis at a specific position in a DNA sequence. *J Biol Chem*. 1978;253:6551-6560.
10. Winter G, Fersht AR, Wilkinson AJ, Zoller M, Smith M. Redesigning enzyme structure by site-directed mutagenesis: tyrosyl tRNA synthetase and ATP binding. *Nature*. 1982;299:756-758. doi:10.1038/299756a0
11. Bloom JD, Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. *P Natl Acad Sci USA* 2009;106:9995-10000. doi:10.1073/pnas.0901522106
12. Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. *Nature*. 1994;370:389-391. doi:10.1038/370389a0
13. Chen K, Arnold FH. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *P Natl Acad Sci USA* 1993;90:5618-5622. doi:10.1073/pnas.90.12.5618
14. Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*. 1985;228:1315-1317. doi:10.1126/science.4001944
15. Ward ES, Güssow D, Griffiths AD, Jones PT, Winter G. Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli*. *Nature*. 1989;341:544-546. doi:10.1038/341544a0
16. Moore JC, Arnold FH. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nat Biotechnol*. 1996;14:458-467. doi:10.1038/nbt0496-458
17. You L, Arnold FH. Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. *Protein Eng*. 1996;9:77-83. doi:10.1093/protein/9.1.77
18. Wittmann BJ, Knight AM, Hofstra JL, Reisman SE, Kan SBJ, Arnold FH. Diversity-oriented enzymatic synthesis of cyclopropane building blocks. *ACS Catal.* 2020;10:7112–7116.doi:10.1021/acscatal.0c01888
19. Porter JL, Rusli RA, Ollis DL. Directed evolution of enzymes for industrial biocatalysis. *Chembiochem*. 2016;17:197-203. doi:10.1002/cbic.201500280
20. Arnold FH, Georgiou G, eds. Directed evolution library creation: methods and protocols. Vol 230 Methods in molecular biology. Totowa, NJ: Humana Press; 2003. https://books-library.net/files/books-library.online-01231634Ko2I8.pdf. Accessed November 21 2022.
21. Zhao H. Directed evolution of novel protein functions. *Biotechnol Bioeng*. 2007;9:313-317. doi:10.1002/bit.2145
22. Chica RA, Doucet N, Pelletier JN. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr Opin Biotechnol*. 2005;16:378-384. doi:10.1016/j.copbio.2005.06.004
23. Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. *Nature*. 2016;537:320-327. doi:10.1038/nature19946
24. Regan L, DeGrado WF. Characterization of a helical protein designed from first principles. *Science*. 1988;241:976-978. doi:10.1126/science.3043666
25. Matteucci MD, Caruthers MH. Synthesis of deoxyoligonucleotides on a polymer support. *Biotechnology*. 1981;103: 3185-3191. doi:10.1021/ja00401a041
26. Leproust E. DNA synthesis – an integral force in the founding and future of precision medicine. *J Prec Med.* 2021;7:33-38. https://www.thejournalofprecisionmedicine.com/wp-content/uploads/dna-synthesis.pdf
27. Ma S, Tang N, Tian J. DNA synthesis, assembly and applications in synthetic biology. *Curr Opin Chem Biol*. 2012;16:260-267. doi:10.1016/j.cbpa.2012.05.001
28. Tian J, Ma K, Saaem I. Advancing high-throughput gene synthesis technology. *Mol Biosyst*. 2009;5:714-722. doi:10.1039/b822268c
29. Opgenorth P, Costello Z, Okada T, *et al*. Lessons from two design-build-test-learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. *ACS Synth Biol*. 2019;8:1337-1351. doi:10.1021/acssynbio.9b00020
30. Ginkgo Bioworks, biology by design: Applying gigabases of DNA to bioengineering. *Twist Bioscience.* https://www.twistbioscience.com/resources/webinar/ginkgo-bioworks-biology-design-applying-gigabases-dna-bioengineering. Accessed November 21, 2022.

CHAPTER 3

Applications of Synthetic DNA in Protein Engineering

The cornerstone of synthetic biology is the design-build-test-learn (DBTL) cycle; an iterative process requiring a large pool of highly diverse DNA sequences for rapid and affordable generation and optimization of proteins. A successful DBTL cycle relies heavily on the ability to iterate on protein design and its underlying sequences. The cycle time depends on how quickly and how broadly this DNA can be accessed for the test phase. However, until recently, researchers producing DNA sequences had to painstakingly clone an organism with the selected DNA fragment and then insert or remove genes using splicing techniques. Today, access to synthetic DNA allows researchers to artificially build custom DNA sequences and accelerate their discoveries. This chapter explores some of the many applications that synthetic DNA has in the field of protein engineering.

High-throughput recruitment assays

Transcription factors (TFs) are proteins that control gene transcription. They contain at least one DNA-binding domain (DBD), which attaches to a specific sequence of DNA, and an effector domain, which recruits cofactors to either promote or repress transcription. DBDs are generally well-characterized and conserved, however, much less is known about effector domains. Recruitment assays are usually used to study the functions of effector domains. In these assays, a candidate effector protein is fused onto a synthetic DBD and recruited to a reporter gene promoter, resulting in changes in reporter expression.[1,2] Traditional recruitment assays have limited throughput, as each effector protein needs to be individually cloned, delivered into cells, and measured. However, researchers recently developed a high-throughput recruitment method (the HT-recruit assay) that measures the function of tens of thousands of candidate effector domains in parallel.[3] For the HR-recruit assay, researchers used silicon-based DNA synthesis to produce pooled oligonucleotide libraries (also known as oligo pools) encoding variants of the effector domain. They combined this with a synthetic surface marker reporter and next-generation sequencing of the protein domains as a readout. This scalable strategy allows researchers to quantify the effector activity of thousands of domains in nuclear-localized human proteins, providing a comprehensive functional assessment of large domain families.[3] The knowledge gained about the molecular mechanisms driving gene transcription and regulation can be used to tackle dysregulated gene expression in the context of disease.

Synthetic genetic circuits

Synthetic genetic circuits are networks of genes that have been engineered to reprogram a cell function. This allows for the creation of microorganisms with distinct and programmable functionalities.[4] There are three steps to this process: 1) the genetic sequences responsible for the required functions are characterized, 2) these sequences are combined to attain more complex functions and 3) these combinations are inserted into cells. This approach can be used to improve the production of chemicals by timing gene expression at different stages of fermentation, or turning on enzymes under particular conditions (e.g., low oxygen).[5] The applications of synthetic genetic circuits range from the production of plant-derived therapeutics on an industrial scale, to the formation of low-cost, point-of-care diagnostic reporters and biofuels.[5,6,7,8]

Researchers generating synthetic genetic circuits face challenges at every step in the process.[9] For example, many of the elements of the circuit (e.g.,

a sequence encoding specific proteins, promoters, enhancers, etc.) are not well characterized; even if the function of each part is known, they may not work as expected when put together. Thus, researchers must often deal with laborious processes of trial-and-error to optimize a pathway's function. Researchers must also ensure that circuits function reliably and do not cause deleterious effects once placed into cells. However, as circuits get larger, the process of building and testing becomes even more arduous. For example, it is estimated that the synthetic genetic circuit developed to produce a precursor of the antimalarial compound artemisinin, took approximately 150 man-years of work.[9,10] Researchers had to uncover the genes involved in the pathway, develop components to control their expression and test different variants to optimize the circuit.

Synthetic DNA can support the building, testing, and optimization of genetic circuitry.[8] For example, silicon-based DNA synthesis platforms enable large-scale and highly precise synthesis of genetic components, helping researchers to test and learn more efficiently (Table 1). Using this approach, it is possible to multiplex the DBTL cycle and generate complex genetic circuits with iteration cycles as short as three weeks.[7,11]

Table 1: Tools offered by Twist Bioscience to support the efficient building, testing, and optimization of genetic circuitry.

Synthetic genes	Oligo pools	Combinatorial assembly
Rapidly synthesize all pathway components and assemble them into vectors ready for testing.	Precise synthesis of over a million unique oligonucleotides up to 300 nucleotides in length. Facilitation of parallel building and testing of short sequences during genetic circuit optimization.	Generation of libraries containing every possible combination of genes in a pathway. This allows researchers to achieve highly parallelized testing of the genetic circuitry, reducing testing time and costs.

Optimizing biologics

Another exciting application of synthetic biology is the development and optimization of biologics. Thus, in this case the DBTL cycle is used to optimize a protein with high affinity for its therapeutic target.

As with other applications, access to large amounts of DNA to build mutagenesis libraries (highly diverse collections of gene variants) remains a major bottleneck in this process. The combinatorial variant library (CVL) technology offers a robust and efficient solution to this challenge. This technology involves soft mutagenesis (i.e., a process that inserts variants according to predetermined criteria), to optimize a protein's properties. In this method each variant is printed base-by-base and screened prior to synthesis, eliminating stop codons, liability motifs, unwanted mutations and any undesirable biases. CVL libraries reduce the screening burden because, unlike complete mutagenesis libraries – in which the number of mutations in each sequence is maximized – they are enriched for specific functional variants.

This technology was recently used to optimize an antibody for the interleukin-21 receptor (IL-21R) –an important player in the immune system's cytokine response. By leveraging CVL technology in combination with structure-guided library design and model-based selection, it was possible to enrich functional antibody variants with desirable characteristics and generate a high-affinity anti-IL-21R antibody with improved biophysical properties. Such antibodies may one day be used to help combat inflammatory conditions like rheumatoid arthritis.[12,13]

Likewise, the technology can be used to optimize affibodies. For example, a CD69 targeting affibody had a rapid clearance but a weak binding affinity. Due to limitations of current medical imaging technology, researchers were interested in developing a CD69 targeting affibody with better binding affinity so that it may serve as an improved imaging tool.[14] Using CVL technology, researchers were able to methodically test which affibody mutations affect binding affinity to CD69. As a result, they engineered a rapidly cleared, high-affinity CD69 targeting affibody that could detect *in vivo* immune responses with remarkably low background noise.[14,15]

De novo protein design

As demonstrated in the previous section, protein engineering can be used to improve existing proteins, manipulating them to serve a range of therapeutic and industrial purposes. However, protein engineering is not limited to existing proteins as today researchers can build bespoke proteins. Such a task can now be done using the power of computational modeling, artificial intelligence (AI) tools and synthetic DNA to produce novel proteins.[16]

De novo protein design promises to provide a near unlimited variety of chemical reactions, biological interactions, and receptor cascades. The first stage in this process is to decide the desired 3D structure of the protein. Next, researchers design tens of thousands of potential backbone conformations to match that structure and test if the calculated sequences fold into the desired form. It is in this stage that deep learning methods (a type of AI that mimic the learning process of the human brain) become essential.[17] For example, tools such as Rosetta and Alphafold can be used to predict protein structures from their sequences.[18,19]

Other methods, such as lowN and ProGen, use language models to generate new sequences.[17,20] Researchers may search through millions of possibilities before selecting the right candidate, which can take several weeks on hundreds of computer processor cores.[21] Notably, these computational tools can also be used to modify existing proteins (e.g., to increase antibody specificity or enzyme activity). Once selected, the desired sequence must be synthesized and tested.

Despite all these technological advances, progress in novel protein design still requires numerous trial and error attempts since structure and function prediction can have accuracy limitations in vivo. The computational tools used for protein design can produce "false positives" and it is difficult to understand the reasons behind these failures. Access to large quantities of high-quality synthetic DNA enables researchers to test thousands of proteins simultaneously and obtain enough data and experimental feedback to improve computational tools (for more information see pages 20-22). [22,23]

Conclusion

Silicon-based DNA synthesis addresses the challenges of throughput and speed with a revolutionary platform that combines miniaturization, parallelization and vertical integration of the end-to-end process from oligo synthesis to gene assembly. This scalable approach to gene synthesis gives unprecedented access to high quality DNA, enabling sufficient sampling of sequence space, faster screening times, lower costs, shorter cycles and fewer iterations. Hence, synthetic DNA is here to fuel advances in protein engineering and beyond.

References

1. Sadowski I, Ma J, Triezenberg S, Ptashne M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature*. 1988;335:563-564. doi:10.1038/335563a0
2. Tycko J, Van MV, Elowitz MB, Bintu L. Advancing towards a global mammalian gene regulation model through single-cell analysis and synthetic biology. *Curr Op Biomed Engin*. 2017;4:174-193. doi:10.1016/j.cobme.2017.10.01
3. Tycko J, DelRosso N, Hess GT, *et al*. High-throughput discovery and characterization of human transcriptional effectors. *Cell*. 2020;183:2020-2035.e16. doi:10.1016/j.cell.2020.11.024
4. Beitz AM, Oakes CG, Galloway KE. Synthetic gene circuits as tools for drug discovery. *Trends Biotechnol*. 2022;40:210-225. doi:10.1016/j.tibtech.2021.06.007
5. Brophy JA, Voigt CA. Principles of genetic circuit design. *Nat Methods*. 2014;11:508-520. doi:10.1038/nmeth.2926
6. Mansouri M, Fussenegger M. Therapeutic cell engineering: designing programmable synthetic genetic circuits in mammalian cells. *Protein Cell*. 2022;13:476-489. doi:10.1007/s13238-021-00876-1
7. Amalfitano E, Karlikow M, Norouzi M, *et al.* A glucose meter interface for point-of-care gene circuit-based diagnostics. *Nat Commun*. 2021;12:724. doi:10.1038/s41467-020-20639-6
8. Rapid, whole-cell engineering of plant alkaloid biosynthesis in yeast using Twist gene fragments. *Twist Bioscience*. https://www.twistbioscience.com/resources/application-note/rapid-whole-cell-engineering-plant-alkaloid-biosynthesis-yeast-using. Accessed November 21, 2022
9. Kwok R. Five hard truths for synthetic biology. *Nature*. 2010;463:288-290. doi:10.1038/463288a
10. Ro DK, Paradise EM, Ouellet M, *et al*. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*. 2006;440:940-943. doi:10.1038/nature04640
11. Srinivasan P, Smolke CD. Biosynthesis of medicinal tropane alkaloids in yeast. *Nature*. 2020;585:614-619. doi:10.1038/s41586-020-2650-9
12. Optimizing a candidate therapeutic antibody to optimize pharmacokinetics using informed library design. *Twist Bioscience*. https://www.twistbioscience.com/resources/application-note/optimizing-candidate-therapeutic-antibody-optimize-pharmacokinetics-0. Accessed November 21, 2022
13. Campbell SM, DeBartolo J, Apgar JR, *et al*. Combining random mutagenesis, structure-guided design and next-generation sequencing to mitigate polyreactivity of an anti-IL-21R antibody. *MAbs*. 2021;13(1):1883239. doi:10.1080/19420862.2021.1883239
14. Andersson KG, Persson J, Ståhl S, Löfblom J. Autotransporter-mediated display of a naïve affibody library on the outer membrane of *Escherichia coli*. *Biotechnol J*. 2019;14:e1800359. doi:10.1002/biot.201800359
15. Optimizing biologics for in vivo imaging. *Twist Bioscience*. https://www.twistbioscience.com/blog/science/Affibodies-Soft-Mutagenesis?topic=601. Published: March 10, 2022. Accessed: November 21, 2022

16. Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. *Nature*. 2016;537:320-327. doi:10.1038/nature19946
17. Ovchinnikov S, Huang PS. Structure-based protein design with deep learning. *Curr Opin Chem Biol*. 2021;65:136-144. doi:10.1016/j.cbpa.2021.08.004
18. Leman JK, Weitzner BD, Lewis SM, *et al*. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020;17:665-680. doi:10.1038/s41592-020-0848-2
19. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589. doi:10.1038/s41586-021-03819-2
20. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-*N* protein engineering with data-efficient deep learning. *Nat Methods*. 2021;18:389-396. doi:10.1038/s41592-021-01100-y
21. Perkel JM. The computational protein designers. *Nature*. 2019;571:585-587. doi:10.1038/d41586-019-02251-x
22. Rocklin GJ, Chidyausiku TM, Goreshnik I, *et al*. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*. 2017;357:168-175. doi:10.1126/science.aan0693
23. Chevalier A, Silva DA, Rocklin GJ, *et al*. Massively parallel de novo protein design for targeted therapeutics. *Nature*. 2017;550:74-79. doi:10.1038/nature23912

CHAPTER 3

Along Came a Computer

Proteins are the backbone of all life. Therefore, the ability to design novel functional proteins holds the potential to revolutionize research within the biological, medical and broader life sciences field. Protein engineering is a relatively new field that seeks to use computer modeling to improve existing proteins, or create new, synthetic polypeptides – with new or enhanced functions that are not seen in nature.[1]

Dr. Possu Huang received his Ph.D. from Caltech with the first demonstration of a computationally designed novel protein-protein interface. His group at Stanford focuses on advancing the understanding of proteins for the engineering of novel therapeutics and other protein-based nanotechnology. He has contributed to a large number of *de novo* designed proteins, most notably to the unlocking of the design principles behind the TIM barrel fold and the invention of eOD, an HIV immunogen design. His group uses machine learning, computational modeling, structural biology and experimental library optimization to continue the expansion of protein-based molecular platforms.

Dr. Ali Madani completed his Ph.D. at UC Berkeley in machine learning where he built computer vision models that outperformed board-certified cardiologists and used natural language processing for electronic health record prediction tasks. As a senior research scientist at Salesforce Research, he was the architect of the ProGen moonshot using large-scale neural language models for functional protein sequence generation. Currently, Ali is the founder and CEO of Profluent Bio – an artificial intelligence startup in protein design and engineering.

This chapter explores the use of artificial intelligence to drive protein engineering, with reference to the computational methods used by both experts in their research.

Q: Could you please give us a brief overview of your research?

Possu (P): My research uses computational modeling to solve cutting-edge molecular design problems. We develop tools that use proteins as programmable polymers to create new therapeutics or build new materials. We also study the fundamental science of proteins: how they carry out functions and how they are used in biology.

Ali (A): Inspired by advances in deep learning to understand and generate natural language (i.e. spoken language), I develop large-scale machine learning models. These models learn meaningful representations of proteins and are used to controllably generate sequences with a variety of bespoke functions that work well in the real world.

Q: What is Artificial Intelligence (AI)? How has it been used in your research and in the wider landscape of protein engineering?

P: AI in the context of protein science is a process that integrate different data sources – for example, sequence and structure of a protein – and help us to derive a solution supported by those data. With AI, computer algorithms can recognize patterns, make inferences from data and perform predictive tasks. In our lab, we build a system to leverage the data distribution from which we can create novel samples. In the case of protein engineering, we model the structural dynamics of proteins and find new solutions with novel functions. Designing flexible and dynamic proteins to carry out new functions has been a fundamental challenge in protein design. We hope to address this problem with the help of machine learning methods.

A: I think AI is a bit of a loaded term. A famous machine learning researcher Michael I. Jordan, coined the term intelligence automation (IA), the idea being that you're creating models that can augment our abilities rather than replace them. I think a more accurate depiction of the work that I do is 'machine learning'. My research asks if a machine can learn what defines a functional protein, using advanced generative models. For example, we can use these models to predict the next amino acid in a sequence – like a natural language model would predict the next word in a sentence – and then extrapolate this to predict sequences in novel proteins, such as the proteins of an emerging virus. More concretely, machine learning is starting to play a broad role in protein engineering by enabling guided directed evolution campaigns to introduce strategic mutations with the aim of improving the function of a natural protein.

Q: What are the typical challenges that you would face when working in this field?

P: Typical challenges surround modeling the behavior of all atoms within a protein. When building protein models, we include every single atom and chemical bond in the system. This requires sophisticated models to describe the physics and specific environment of proteins. Whilst this remains a challenge, new systems have been developed to capture the behavior of amino acids in a structure with a level of sophistication and accuracy suitable for design. Additionally, amino acids in a protein can interact with each other to create lots of conformations. Hence, another challenge is how to sort protein states. Flexibilities within each atom must also be accounted for, making modeling very complicated.

Whilst patterns can be extracted from protein structure databases and used to mimic the behavior of proteins in nature, when you are creating a new protein, the answer is not in the database. Systems are required to 'learn' the underlying physics, consistent with how proteins behave. Once you have a notion of how they should behave, then new permutations can be created.

A: I spend a lot of my time explaining biology to machine learning scientists and explaining machine learning to biologists! They are two distinct disciplines that have been separately developed, but now more people are getting trained at the intersection of the two.

There are tons of challenges – which translate to a plethora of exciting questions and tasks to tackle. A significant challenge is centered around data - its curation, acquisition, and evaluation. To start, effectively curating data, so it is informative and relevant for model learning, is imperative in machine learning. For functional protein engineering, whether through machine learning and/or directed evolution, it is important to determine how to measure and characterize what you've designed and engineered in the lab. This can be difficult as machine learning needs can be quite data-intensive whereas protein synthesis and characterization is typically low-throughput. Lastly, proper evaluation is required throughout the design cycle. How do we devise proper tests, evaluation criteria, and relevant metrics to assess model performance?

Q: Could you provide a brief overview of the theory behind your approach to protein design and how that compares to other approaches?

P: Today, sequence data are some of the easiest data to acquire due to large-scale sequencing efforts. However, proteins have three-dimensional structures and the Protein Data Bank only contains a few hundred thousand structures. This makes a structure-based approach difficult.

AI can learn from sequence data. The limitation is that the learning of sequence patterns alone may not fully satisfy the level of detail required to achieve function. Some sophisticated design methods can operate on sequences, but with 3D structural data underpinning to be successful. Without the structure data artificially generated sequences almost always look like native proteins and their use is limited. A good pattern recognition software could potentially learn from distant, unrelated, sequences to establish new viable sequence patterns unseen in nature. But so far, that capability hasn't yet been achieved.

Nonetheless, a sequence-based approach is incredibly powerful for improving proteins, such as antibodies, when some experimental data are available. Instead of modeling an antibody structure, drug discovery teams can create large experimental sequence datasets from which an AI system can learn. Antibodies share a lot of structural and sequence characteristics therefore, variable features can be identified from experimental data and optimized. For example, data regarding solubility or immunogenicity can be used to improve the desirability of variable features; however, creating a

totally new protein is much more limited.

A protein's function is determined by its structure. Antibodies bind to target antigens because they share complementary shapes. My lab uses a structure-based approach to model flexible regions of antibody so they can be adapted to different targets. Should a new pathogen emerge to which humans have no natural immunity, antibodies could be created using an AI-driven structure-based approach. AI systems can 'learn' the patterns in antibody structures and create new structures with only new variable regions. Designing against a unique structural epitope is then possible to achieve high binding specificity.

There are other methods that utilize both structure- and sequence-based approaches. Software such as Alpha Fold – an AI program that predicts protein structure based on sequence data – can be used to provide information about how new predicted sequences could translate into new structures. A scheme that iterates sequence and predicts the resulting structure can be used as a design strategy.

A: In natural language, there are structural concepts (subject, verb, syntax, grammar) and there are associations and links from one word to another in a sentence. We can extrapolate this to proteins; as proteins fold, they form secondary and tertiary structures. In a protein sequence, amino acids will interact with and be influenced by other residues that are not immediately neighboring to each other. The linguistics-based model can learn facts about protein science, biophysics, and family ontologies all from the sequence itself! As we scale the capacity of large protein language models, we can achieve greater performance for a variety of property prediction and sequence generation tasks.

The traditional approach to protein design has been structure-first. However prior lessons in machine learning (a la Unreasonable Effectiveness of Data) have indicated we should "follow the data" - meaning it may be better to look at where the largest and most information-rich source of data is. I would argue that protein sequences provide that; there are orders of magnitude more sequences than structures. Language models have been incredibly versatile as universal computation engines to capture the distribution of those observed sequences. From a practical side, a sequence-first approach can get you quite far in capabilities for most protein engineering applications.

Like with all distinctions drawn in science, the motivations and implications of sequence-first vs structure-first modeling for protein design and engineering is complex and depend on what you want to achieve. For example, if you want to interpolate between known functions, a sequence-based perspective might be best. However, if you want to extrapolate to functions that are very different from those found in nature, returning to first principles and governing equations with a structure-based approach might be better.

All in all, I think there's a lot of interplay and we're increasingly seeing the fields merge. For example, many protein structure models are leveraging the coevolutionary signals found in sequence databases and many protein sequence models are implicitly learning three-dimensional structure. There is a lot to be done to bridge the two paradigms between sequence and structure.

Q: How might synthetic DNA tools impact advances in your research field?

P: We produce our designed proteins in bacteria by giving them the DNA sequences that encode the protein. With the computing revolution behind us, synthetic DNA technology is arguably the most important factor driving protein design technology forward.

Taking binder design as an example, while it is possible to create several protein structures that will bind to a target, it's difficult to predict which one will bind most effectively. To find the best solution, the best way is simply to test lots of designs experimentally. The ability to create lots of custom molecules (using synthetic DNA approaches) and test their binding affinity in parallel enables rapid protein engineering processes.

Synthetic DNA technology has dramatically reduced its cost over the last decade. It allows us to test new hypotheses experimentally, which in turn feed back to computer algorithms and our understanding.

A: One of the biggest challenges around design is being able to synthesize and test large numbers of sequences or proteins. Advancing the tools and methodologies to enable high throughput experimentation is critical to the success of this field, particularly the ability to synthesize DNA in a high throughput, reliable manner, at low cost. I am absolutely certain synthetic DNA is going to be huge and will fuel advances in this field.

Q: Finally, where do you see the future of protein engineering heading?

P: We have very good results to suggest that future response times to infection could be shortened with protein engineering. By understanding how specific antibodies bind to an antigen and designing new antibodies on a computer, new biologics could be created in record time – not counting FDA approval. As a result, future pandemic responses may look very different. But the technology on a broader scale is not limited to just infectious diseases and biologics development. Any process that involves proteins can potentially benefit from the ability to tweak or repurpose these proteins, or to replace their functions with new ones. Biology uses proteins for everything, from interacting with the environment to generating energy. We are just at the beginning of understanding and hacking these processes with protein engineering.

A: One goal is to have a versatile model with controllable levers to optimize multiple protein attributes simultaneously. Imagine if you could direct a model with a command, "I want a protein that works for this particular function, resides within this family, and exhibits this level of thermal stability". It would be amazing to toggle these parameters, feeding them easily into the model. Then, you could output a library of proteins that have a high chance of working once you synthesize them in the real world. That would have a huge transformative impact on the whole field of biology: human health, disease, the environment, etc. I think critical to its success will be better capturing the sequence-to-function relationship– a difficult, unscoped research direction for the future.

Taking a step back, proteins enable everything critical to life; they are the molecular workhorses for almost all biological processes. There is a multitude of problems that could be solved by designing new or better proteins, in therapeutics, or in food and agriculture (e.g. alternative meats), or in climate protection (e.g. plastic degrading enzymes). I continue to be excited about the opportunity to solve the world's most pressing needs through protein design and engineering.

Dr. Possu Huang

Assistant Professor of Bioengineering, Stanford University

Dr. Ali Madani

Founder, Profluent Bio

Reference

1. Designing proteins to improve the world. Twist Bioscience. https://www.twistbioscience.com/blog/company-news-updates/designing-proteins-improve-world. Published July 18 2019. Accessed December 15 2022.

CASE STUDY

TWIST BIOSCIENCE

THE BAKER LAB AT THE UNIVERSITY OF WASHINGTON

Designing Proteins to Revolutionize Biotechnology

MAY 2019

The Baker Lab is part of the Institute for Protein Design at the University of Washington. Instead of engineering an existing protein sequence, the lab focuses on the design of entirely new proteins to address challenges in health and technology. Led by David Baker, the lab has registered a series of achievements, including an experimental RSV vaccine, the development of anti-cancer proteins with fewer side effects, and new enzymes that help microbes convert carbon dioxide into useful chemicals.

DAVID BAKER

Recently, Baker Lab received a $45 million grant from the Audacious Project, a funding consortium, to pursue several computational design projects, including a flu vaccine capable of providing lifetime immunization, advanced protein containers for targeted gene delivery, and smart proteins capable of identifying cancerous or otherwise unhealthy cells.

Despite these advances in protein design, progress has been difficult because the majority of designed protein sequences fail to fold or don't function as designed. The computer models used for protein design produce many "false positives" (sequences that look good on the computer but fail in the lab), and understanding why these failures happen is very challenging. A protein might fail for many hundreds of reasons, and a failed design can provide little feedback on what went wrong. Recently, the Baker Lab sought to overcome this challenge through a combination of computation and large-scale experimentation. Testing designs at a much larger scale makes it possible to look for patterns in both the successes and failures.

Evaluating Protein Folding with Oligo Pools

In two recent reports, Baker Lab researchers markedly advanced their protein design methodology using Twist Bioscience Oligo Pools. Oligo pools consist of between thousands and hundreds-of-thousands of synthetic oligonucleotide sequences, pooled together. In papers published in *Science* and *Nature*, the researchers used oligo pools to encode large numbers of potentially viable protein structures. Gabriel J. Rocklin, PhD, a former fellow at the Department of Biochemistry of the University of Washington and a member of Baker Lab, who was a lead author of both studies, said that by designing very small proteins that could be encoded by a DNA sequence the size of a single oligo, new protein designs could be tested in high throughput. Rocklin is currently an assistant professor of Pharmacology at Northwestern University in Chicago, and a member of Northwestern's Center for Synthetic Biology.

Achieving a Rate of Success Not Previously Possible

In one study, Rocklin and the team designed and screened thousands of proteins to determine whether each formed a stable, folded structure. Proteins that don't fold properly are extremely sensitive to proteolytic enzymes, enabling the team to use resistance to proteolysis as a measure of folding. The designed proteins were expressed in yeast, with each design being presented on the surface of the yeast cell that synthesized it. Cells expressing stable proteins were labeled with a fluorescent dye and collected in a cell sorter, so that DNA sequencing could reveal the molecular identities of all the stable designs.

"There are so many things that can go wrong with a protein," Rocklin said. "In the old design approach, you make 10 or 15 proteins, and three of them work while the others don't. That's great to get some success, but it doesn't allow us to improve in the future. Now, by testing 4,000 or 10,000 proteins at a time, we get enough data and experimental feedback to be able to improve our computational tools."

Encoding the larger number of proteins in affordable oligo pools gave Rocklin the information he needed.

"That was super useful feedback that went into improving the computational model and improving the design process," he said. "Large scale oligo library synthesis that's affordable enables

Our ability to get huge numbers of oligos affordably...made the experiments possible.

Gabriel J. Rocklin, PhD
ASSISTANT PROFESSOR, NORTHWESTERN DEPARTMENT OF PHARMACOLOGY AND CENTER FOR SYNTHETIC BIOLOGY

us to do something that I think protein designers had been wanting to do for a long time but had not had a way to do, which is test thousands of completely customized, computer designed protein sequences.

"The very first time we tested any of the designs, the overall success rate was about five percent," Rocklin said. "By the end after four cycles with the oligo pools, the overall success rate was about fifty percent. It was made possible by data that was impossible to collect beforehand."

We believe proteins designed completely from scratch will revolutionize biotechnology.

David Baker

"Large scale testing also enabled a really incredible thing, which was to test thousands of control sequences that were similar to the designs, but were expected *not* to fold," Rocklin added. "To generate similar sequences we expected to fail, we made mutations at key residues, or shuffled the order of the amino acids, which keeps the key physical properties of the protein molecule identical. One would never test controls like this when testing designs one at a time — it's too much work. But it's easy in a large-scale experiment. These controls showed us that many of our designs were much, much more stable than closely-related control sequences, which gave us confidence that the designs were stable because they folded *as designed*, rather than folding into some serendipitous, non-designed structure."

Large-Scale Design for Therapeutic Candidates

In the other study, a team led by Rocklin and Aaron Chevalier and Daniel-Adriano Silva, postdoctoral fellows at Baker Lab and the University of Washington, again tested the protein folding process and also designed proteins to bind to two different targets: virulence factors causing influenza and botulism. Since the proteins were designed to bind to disease-causing targets, they are considered potential therapeutics. By binding to the targets, the molecules could block the virulence factor's disease causing activity. The team produced proteins that target influenza haemagglutinin and botulinum neurotoxin B, along with control sequences to probe contributions to folding and binding, and identified high-affinity binders. By comparing the binding and non-binding design sets, which are two orders of magnitude larger than any previously investigated, the team was able to evaluate and improve their computational model, leading to an increase in design success.

Designing a protein that folds is arguably a simpler challenge than designing a protein that folds into a structure that can bind to a specific target molecule. However, cracking this problem opens up the potential to treat serious global disease threats. Again, the ability to use large amounts of oligonucleotides at competitive prices allowed the researchers to use the power of scale to circumvent this challenge. "Because of the difficulty of that task, if we were testing tens of designs at a time, we might not have gotten anything," Rocklin said.

"In both papers, we took some of the successful designs — whether they were designs that folded properly, or designs that bound to the target — and solved structures of a few of them," Rocklin explained. "In both papers we solved the structures of a small subset of the protein panel. Those structures, along with the thousands of control sequences, gave us confidence that the thousands of other successes that we got were also likely folded as designed, even though we didn't structurally characterize every one of them."

Creating Computationally Designed Scaffolds

"There is a lot of interest in computationally designed non-antibody scaffolds, as they can be much more stable than scaffolds that people find from natural proteins," Rocklin noted. "Because of the stability of computationally designed proteins, you might ultimately have more robust drug candidates that can be stored more easily. That paper and others have shown that you can get similar binding activity to an antibody with computationally designed proteins."

In both experiments, Rocklin and his colleagues were able to create new, folded protein designs that could direct future therapeutic and synthetic protein development. Twist Oligo Pools enabled their high-throughput experiments with the flexibility needed to move forward.

"Our ability to get huge numbers of oligos affordably and also test them all in pools, made the experiments possible," Rocklin said. ■

WHAT CAN TWIST DO FOR YOU?

sales@twistbioscience.com
twistbioscience.com

#WeMakeDNA

Powering Modern Drug Discovery with DNA Synthesis

Over the past century, remarkable progress has been made in the treatment of different diseases, with protein discovery being an important research avenue. With applications ranging from the synthesis of chemical compounds to personalized T-cell therapies, synthetic proteins represent key pillars of modern drug discovery.

With the right tools, engineered proteins can be designed to interact with hard-to-drug targets, to treat emerging diseases, and to overcome drug resistance. However, the efficiency of the protein engineering process is often hampered by time-consuming and labor-intensive methods. This infographic will explore how modern DNA synthesis and parallelization can solve this problem by accelerating the throughput of protein engineering and drug discovery.

Protein-Based Drug Discovery

Drug Biosynthesis

Heterologous enzymes can be used to recreate the biosynthesis of plant-derived medicines (e.g., anti-malarian drug artemisin and anti-nausea tropane alkaloids).[1, 2]

Further, compared to traditional inorganic catalysts, enzymes are:

- more cost-effective
- faster
- more specific

Therapeutic Antibodies

Therapeutic antibodies are customized biopharmaceuticals used in the treatment of different diseases (e.g., cancer, inflammatory diseases, and autoimmunity).

They currently represent a predominant class of novel drugs due to their:[3]

- high affinity and specificity
- low immunogenicity
- amenability to genetic engineering (enabling precise targeting of biomolecules)

Adoptive Cell Therapies

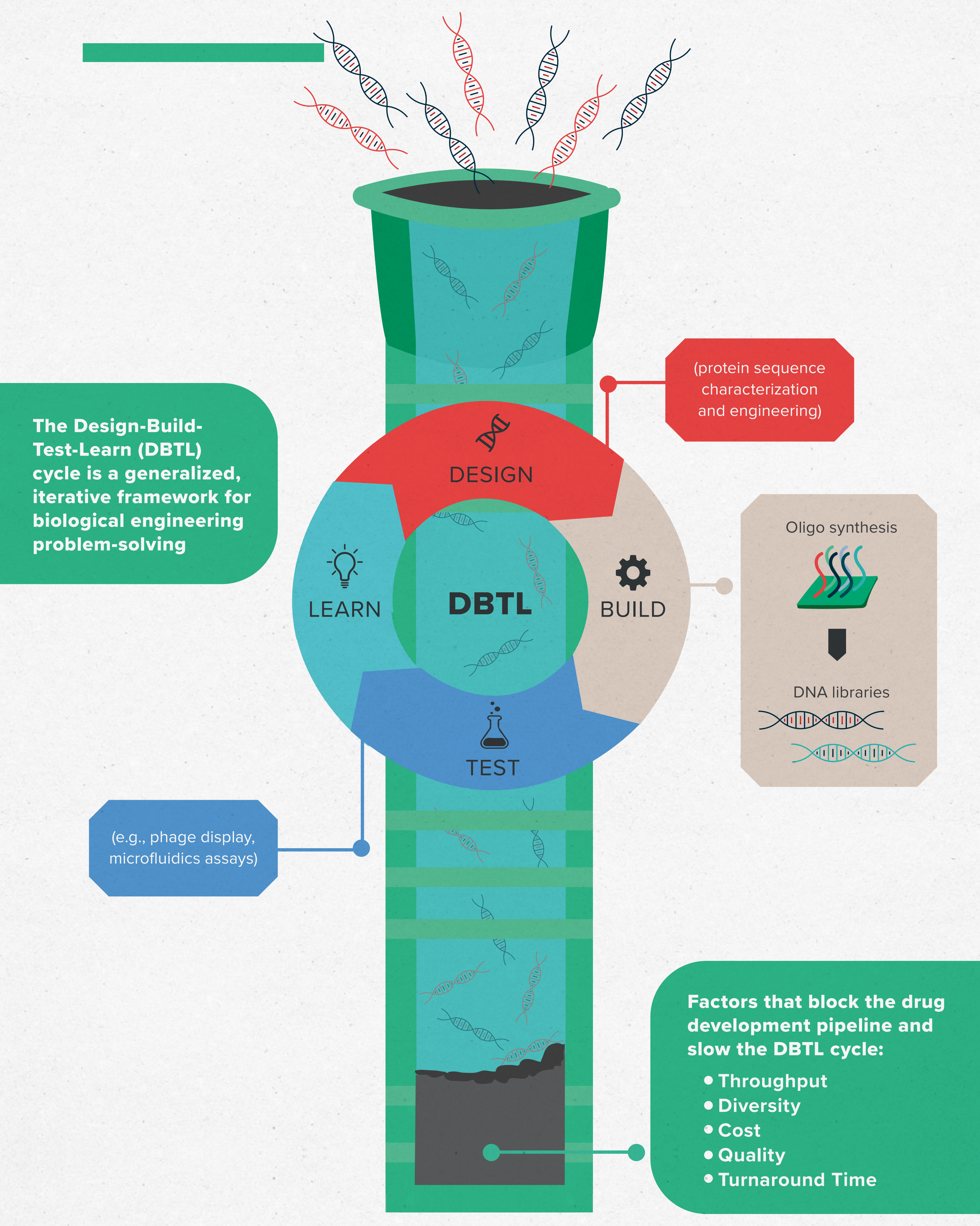
Adoptive cell therapy (ACT) is a highly personalized cancer treatment that uses immune cells to attack tumors.

ACTs use a novel T cell receptor (TCR-T cells) or a chimeric antigen receptor (CAR-T cells) to: [4]

- target tumor-specific antigens
- elicit a potent and specific anti-tumor immune response

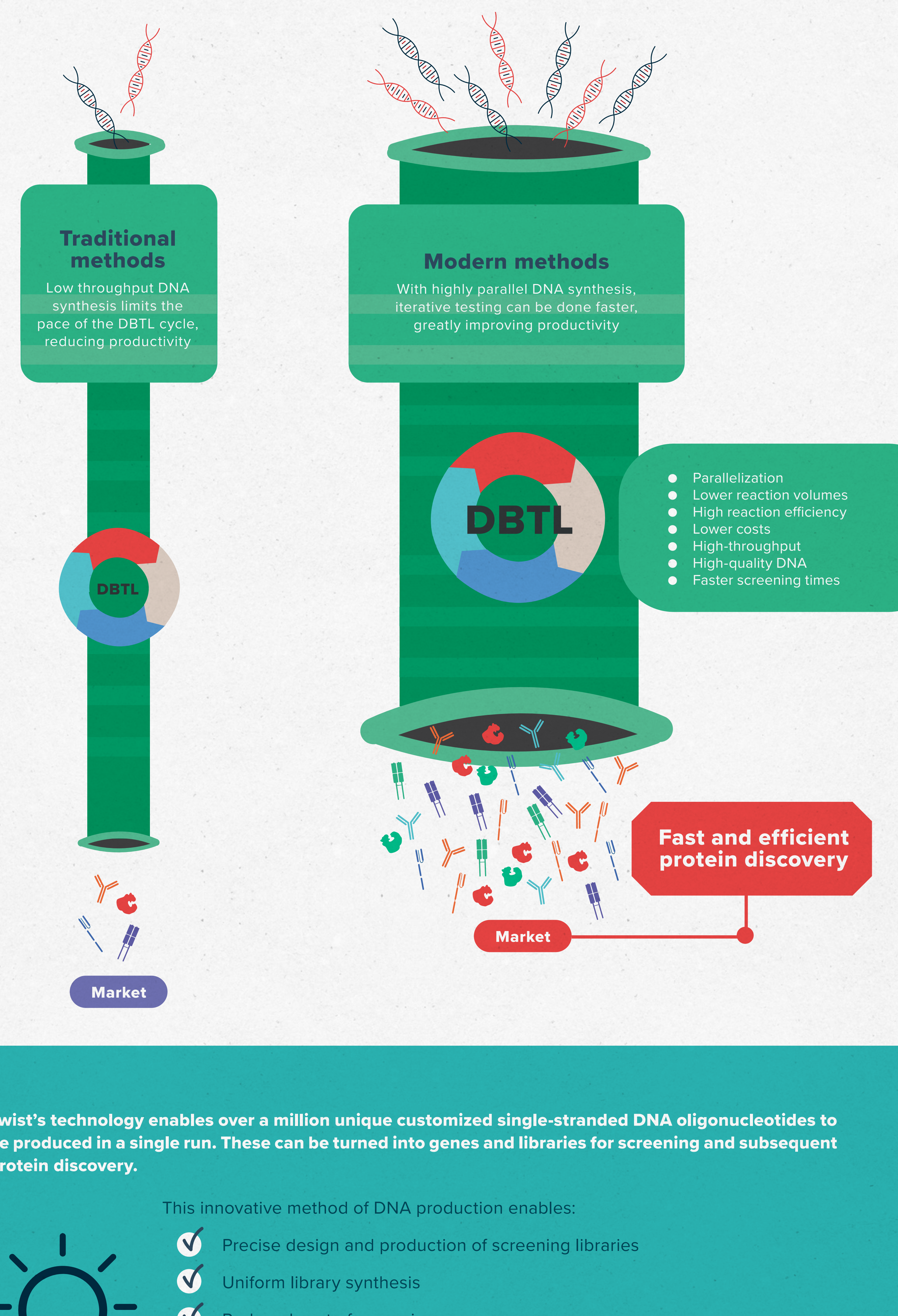
Unblocking the Drug Discovery Pipeline

Protein discovery relies on the design, build, test, learn (DBTL) cycle which needs to be repeated many times, testing different protein-coding DNA sequences, until optimal candidates expressing high-affinity proteins are identified. The cycle time is heavily influenced by how many potential candidates are encompassed within DNA libraries and how quickly these libraries can be created.[5]



Advantages of Modern Silicon-Based DNA Synthesis

Synthesizing DNA libraries with thousands of variants has traditionally been a costly, slow, and error prone process. Twist's silicon-based DNA synthesis platform overcomes these hurdles through miniaturized chemistry, enabling precise and parallel synthesis of thousands of DNA sequences. This enables researchers to produce and test a much wider range of variants, greatly speeding up the DBTL cycle.



Twist's technology enables over a million unique customized single-stranded DNA oligonucleotides to be produced in a single run. These can be turned into genes and libraries for screening and subsequent protein discovery.

This innovative method of DNA production enables:

- Precise design and production of screening libraries
- Uniform library synthesis
- Reduced cost of screening
- Improved productivity when screening
- Shorter DBTL cycles (fewer iterations)
- High-throughput synthesis that is both highly scalable and fully customizable

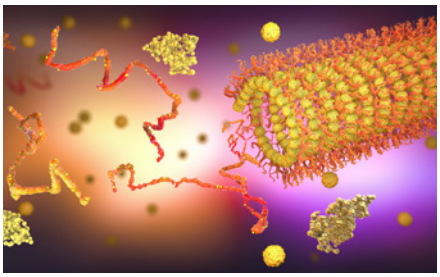
To learn more about Twist Bioscience's silicon-powered DNA synthesis, click here

Design, Build, Test, and Learn with Twist Bioscience

References

1. Paddon C, Keasling J. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* 2014; 12:355–367. doi: 10.1038/nrmicro3240
2. Srinivasan P, Smolke CD. Biosynthesis of medicinal tropane alkaloids in yeast. *Nature.* 2020; 585:614–619. doi: 10.1038/s41586-020-2650-9
3. Gil M. Therapeutic antibody engineering: past, present and future. *Technology Networks*. https://www.technologynetworks.com/tn/lists/therapeutic-antibody-engineering-past-present-and-future-354173. Published: September 28, 2021. Accessed: June 6, 2022.
4. Rohaan MW, Wilgenhof S, Haanen JBAG. Adoptive cellular therapies: the current landscape. *Virchows Arch.* 2019; 474(4):449-461. doi: 10.1007/s00428-018-2484-0
5. Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harb. Perspect Biol.* 2017; 9(1):a023812. doi: 10.1101/cshperspect.a023812

Compendium

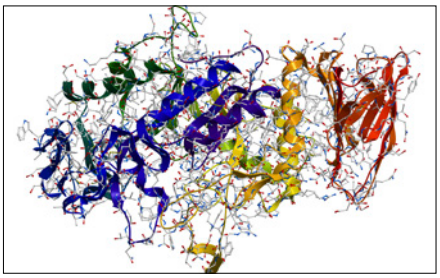


Enzyme identification and engineering for novel industrial enzymes

In this webinar, Derek Quinn, senior biology team leader at Almac Sciences, discusses how to identify new biocatalysts using bioinformatics, gene synthesis, and rapid screening.

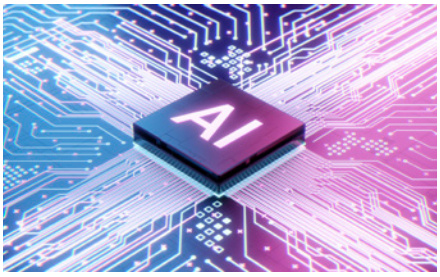
De novo Design of G protein mimetics

In this webinar, Chris Bahl, head of protein design at the Institute of Protein Innovation, explains how to use *de novo* protein design to develop novel mimetic proteins.



Protein predictions: new tools to understand these complex structures of life

Read this blog to learn how AlphaFold, a protein structure solver software, can open new frontiers in protein design.



Finding improved biologics with twist's machine learning and deep learning tools

Download this flyer to learn the value of artificial intelligence tools for protein discovery.

Large language models generate functional protein sequences across diverse families

In this paper, Dr. Ali Madani and colleagues present ProGen, a language model that can generate protein sequences with a predictable function across large protein families.