Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ35307

# *From Traditional Protein Engineering to AI-Driven Design: Principles, Methods, and Applications*

**Zonghao Cheng**

*School of Biological Sciences and Technology, Beijing Forestry University, Beijing, China*
*Chengzonghao09@sina.com*

***Abstract.*** Protein engineering is a critical technology that modifies protein sequences and structures to optimize specific functions or introduce new ones, and it is widely applied in biomanufacturing, pharmaceutical research and development, synthetic biology and other fields. Traditional protein engineering mainly relies on the modification of natural proteins, and its strategies usually include simulating natural evolution processes or designing based on protein structures, with the aim of enhancing protein stability, activity and other properties. However, the high complexity of protein sequences and the limited understanding of the correspondence between amino acid sequences, protein structures and biological functions have restricted the further development of traditional methods. In recent years, the rapid development of artificial intelligence has greatly driven the paradigm shift in protein engineering research, enabling researchers to efficiently explore proteins with specific structures and functions in a broader sequence space. This paper reviews the basic principles and development history of protein engineering, and introduces traditional strategies such as directed evolution and rational design. It focuses on protein structure prediction tools represented by AlphaFold, generative models such as RFdiffusion and ProteinMPNN, and the de novo protein design methods driven by these tools. Furthermore, combined with classic applications of protein engineering including metabolic engineering, enzyme engineering and antibody engineering, this paper analyzes the progress of artificial intelligence in protein design and optimization. By sorting out relevant studies, this paper aims to provide a reference for understanding the development context of protein engineering and the application of artificial intelligence in this field.

***Keywords:*** protein engineering, directed evolution, artificial intelligence, deep learning, de novo protein design, enzyme engineering

## 1. Introduction

### 1.1. Introduction to protein engineering

Proteins in nature possess extremely extensive and powerful functions, and almost all life activities require the participation of proteins with different functional types. With the development of human society, new challenges have emerged in fields such as healthcare, industry and the environment. The rich structural and functional diversity of proteins, as well as the excellent biological functions

DOI: 10.54254/2753-8818/2026.HZ35307

exhibited by natural proteins, make them widely regarded as an important solution to these challenges. The concept of protein engineering emerged as the times require. Protein engineering refers to the process of endowing proteins with specific functions or strengthening their original properties through modification following the sequence-structure-function logic, and it has extremely broad application prospects in many fields [1, 2]. In the early days after the concept of protein engineering was proposed, traditional methods such as directed evolution and rational screening were usually used to optimize protein functions. Nowadays, with the increasing popularity of artificial intelligence, protein design is undergoing an unprecedented revolution.

### 1.2. Underlying design logic of protein engineering

The core principle of protein engineering lies in the intrinsic correspondence among amino acid sequences, three-dimensional protein structures and biological functions. It is worth noting that the correspondence among sequence, structure and function is highly complex: two polypeptide segments with little difference in primary sequence may have completely different three-dimensional structures, and their corresponding biological functions may also be vastly different [3, 4].

### 1.3. Current status and future of protein engineering

The diverse arrangements of 20 amino acids endow amino acid chains with rich aggregation and assembly behaviors. The permutations and combinations of amino acid sequences, together with the complex folding and self-assembly processes of proteins and polypeptides, jointly endow them with nearly infinite structural possibilities [5, 6]. However, in the vast structural space, only a very small number of conformations can exhibit specific biological functions or activities [6]. Traditional protein design approaches include rational design based on site-directed mutagenesis using known protein structure and function information, and directed evolution based on the principle of random mutation-natural selection in nature. Nevertheless, the inexhaustible possibilities of proteins and polypeptides impose certain limitations on both methods; the huge experimental workload and economic cost of directed evolution restrict its application [7, 8]. The emergence of machine learning has injected new vitality into traditional protein engineering. For example, high-precision structure prediction tools such as AlphaFold have made it possible to obtain a large amount of protein structure information quickly and accurately, and have also enabled the combination of the two strategies of directed mutagenesis and rational design [9]. Machine learning can also complete sequence and structure-function annotation, mutant library construction and other tasks based on existing data, and generate candidate sequences that may have new functions de novo through generative models [10-13]. With the application of more machine learning tools, the prediction and experimental screening of complex combinations that are difficult to achieve with traditional rational design and directed evolution have become achievable goals.

Looking back from the AI era, this paper reviews traditional protein engineering methods dominated by rational design and directed evolution, as well as the new opportunities brought by the AI era to protein engineering. It also illustrates the current status, future development and challenges of this field through practical cases of protein engineering.

DOI: 10.54254/2753-8818/2026.HZ35307

## 2. Traditional protein engineering methods

### 2.1. Directed evolution

Directed evolution is the earliest protein engineering approach and also one of the most classic protein optimization strategies. Its earliest emergence can be traced back to the 1980s, when the concept of an "evolutionary machine" that achieves protein evolution through mutation and screening first appeared in the scientific community, and immediately attracted a group of scientists to conduct research based on this idea. Among them, Szostak and Hageman made particularly prominent contributions, and their experiments laid the foundation for the directed evolution of enzymes [14, 15]. In the early stage of its emergence, directed evolution was mainly used for trait modification of individual proteins [16]. In the subsequent development process, it was gradually extended to modify protein complexes and nucleic acids, and applied in fields such as genetics and biomedicine.

The basic principle of directed evolution originates from the evolutionary process proposed by Darwin, namely mutation-selection [8]. The construction of mutants mainly includes two approaches: random mutagenesis and gene recombination. These two approaches achieve the acquisition of a large number of mutants by creating site-directed mutations and genotypic combination changes respectively.

Error-prone PCR (epPCR), developed by Goeddel's team in 1989, has been the most commonly used *in vitro* random mutagenesis method since its proposal [17], and it can increase the mutation rate up to $10^{-2}$–$10^{-1}$ s.p.b. [18]. In addition, the mixed use of multiple low-fidelity DNA polymerases can obtain a relatively uniform mutation distribution, which is more conducive to the overall mutation-selection effect [19].

Among the methods for constructing mutants through the gene recombination approach, the most commonly used is DNA shuffling first proposed by Stemmer's team in 1994. This method fragments double-stranded DNA with DNase I and further completes fragment recombination using primer-free PCR, thereby obtaining a recombinant sequence library [20]. Compared with non-combinatorial methods represented by epPCR, gene recombination can explore more sequence space, and even achieve simultaneous optimization of multiple properties of the same protein [20]. However, gene recombination also has its limitations. For example, parental genes are prone to reassembly, and the existence of regions with high sequence identity leads to concentrated crossovers [21]. To further improve the efficiency of gene recombination, scientists have also made various improvements. Nevertheless, the complex correspondence between protein structure and sequence also imposes some restrictions on the application of gene recombination in directed evolution. After the construction of the mutant library is completed, variants with target functions need to be screened, and selection is usually achieved through resistance screening, growth dependence and other methods [22].

### 2.2. Rational design

The core idea of rational design is to predict key mutation sites based on the relationship between protein structure and function, and perform precise modification using site-directed mutagenesis technology. Compared with directed evolution, rational design requires a smaller mutation scale, and has the advantages of short implementation cycle, strong pertinence and low resource consumption [23–25]. The traditional goals of rational design are to improve enzyme activity, stability and other properties. The main strategies include multiple sequence alignment, reducing steric hindrance,

DOI: 10.54254/2753-8818/2026.HZ35307

reconstruction based on interaction networks, regulation based on protein dynamic behavior, and overall design based on computational biology.

Although the sequence-structure-function relationship of proteins is highly complex, generally speaking, enzymes with high identity and similar structures tend to have similar functions. The application of multiple sequence alignment is precisely based on this [25]. If the modification scope is narrowed to a single site, scientists often search for CbD (conserved but different) sites through MSA and modify them to be consistent with homologous sequences, so as to improve the properties and functions of the protein to be modified [26]. Modifying specific positions such as the substrate-binding pocket of enzymes for the purpose of reducing steric hindrance can effectively improve the catalytic activity of enzymes [27, 28]. In addition, some modifications based on steric hindrance have also achieved the effect of promoting the binding of target substrates while preventing non-target substrates from entering the active site by using steric hindrance. Through the combined effect of the two directions, the substrate selectivity of enzymes is improved [28].

Optimizing enzyme activity by altering various interactions formed between enzyme active sites and substrates is also an important idea [29]. In a 2021 study, based on the structural analysis results of β-AADH and 3,5-DAHDH, an enzyme with strict substrate specificity, rational design modification was performed on E310G, another mutant that affects the substrate range, to improve its activity towards non-natural substrates. After analyzing the active site of the mutant and the substrate binding mechanism, the research team performed a point mutation on Gly323, mutating it to serine, which enabled it to form additional hydrogen bonds with the substrate, and ultimately increased its specific activity towards the target parent by 17 times [30]. At present, the idea of rationally designing and optimizing enzymes based on enzyme-substrate interactions has been widely used, achieving the goals of improving substrate catalytic efficiency and altering enantioselectivity for target enzymes [28, 31].

The binding of enzymes and substrates is not a direct binding process, and is often accompanied by significant conformational changes of enzymes. Therefore, enzyme catalysis does not only depend on the catalytic pocket; domains related to the entire dynamic binding process are also involved. Thus, rational design of enzymes can be carried out based on the dynamic behavior of enzyme-substrate binding. Commonly used techniques for rational design in this direction include molecular dynamics simulation and conformational dynamics. Mutations can be made on residues in flexible regions far from the active center or on loop regions to alter the dynamic characteristics such as flexibility and length of the regions, ultimately obtaining variants with higher activity, higher stability or selectivity [32-34].

As a highly targeted protein engineering strategy, rational design has often been limited by the incomplete understanding of sequence-structure-function relationships in past applications. However, with the development of machine learning technology, this situation is gradually changing. The addition of machine learning makes it possible to integrate and analyze large-scale previous sequence, structure and function data, and the prediction of functions and mutation sites will be more convenient. Protein engineering will also gradually transform from "trial-and-error driven" to a new stage of "data and prediction driven" rational design.

## 2.3. Semi-rational design

Due to the excessive resource consumption and certain technical difficulties of traditional directed evolution and rational design in obtaining ideal phenotypes, scientists have established a simpler and more efficient semi-rational design strategy. Semi-rational design can construct mutant libraries with smaller scale, higher accuracy and containing conserved regions, which is a relatively concise and

DOI: 10.54254/2753-8818/2026.HZ35307

efficient advanced strategy [9]. In addition, molecular docking simulation is also involved in the semi-rational design process: by analyzing the binding interface between enzymes and substrates, key amino acid residues involved in interactions are directly identified, mutant libraries are constructed, and high-throughput screening is completed [9]. As a more advanced protein engineering strategy, semi-rational design can integrate sequence and structure information for analysis, making it a more effective protein engineering strategy. However, it can only modify existing enzymes, and de novo design requires more advanced protein engineering strategies.

## 3. AI-driven protein engineering

With the breakthrough development in the field of artificial intelligence, tremendous changes have also taken place in the field of protein engineering. New methods incorporating machine learning can break through the limitations of the original Darwinian mutation-selection process, greatly accelerating the development of protein engineering.

The goals of protein design can be divided into two categories: modification of existing proteins based on known proteins, and de novo design. The rational design and directed evolution strategies mentioned above can generally only be used for protein design starting from known proteins, to enhance their existing functions or endow them with new functions. However, natural proteins are limited by their origin from natural evolution, and even after optimization, they are often not truly "optimal designs" [35]. Modified proteins often retain characteristics adapted to the environment of their original organisms, and their structures are also restricted by historical evolutionary paths [36, 37]. In contrast, the de novo design strategy can largely bypass these limitations, and obtain proteins with higher stability and functions more tailored to demands through a modular design approach [38].

With the advent of the AI era, the addition of machine learning (ML) has led to rapid growth in the scale of protein sequence and structure information, making de novo protein design possible. It has also further expanded the application scope of protein engineering.

### 3.1. Protein structure prediction

The function of a protein is closely related to its three-dimensional structure. Protein structure determination has evolved from traditional methods including X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy (cryo-EM) to structure prediction based on computational biology [39, 40]. With the release of AlphaFold2 in 2021, humans finally obtained a high-throughput, high-precision computational method capable of independently completing atomic-level accuracy protein structure prediction [10]. Following AlphaFold2, an important milestone in the field of protein structure prediction, a large number of efficient deep learning-based protein structure prediction models have developed rapidly. Models such as AlphaFold3 released in 2024 not only have higher accuracy in protein structure prediction, but also extend application scenarios to structure prediction of complexes, multimers and more [41]. These models have promoted the development of protein engineering towards de novo design, and further advanced the fields of enzyme engineering, synthetic biology and drug discovery.

Meanwhile, on the basis of the continuous evolution of deep learning frameworks, the field of protein structure prediction is further developing towards larger-scale sequence modeling and unified multimodal modeling. Structure prediction methods based on protein sequence language models (such as ESMFold) show that large-scale pre-trained language models can directly capture protein structural constraints from sequence information, achieving high-precision atomic-level

DOI: 10.54254/2753-8818/2026.HZ35307

structure prediction [42]. On this basis, AlphaFold-Multimer further extends the monomer prediction framework to enable effective modeling of protein complex interfaces, thereby promoting the development of multimer structure determination and protein-protein interaction research [43]. In recent years, structure prediction models have gradually expanded from the "protein structure prediction" paradigm to the "general biomolecular modeling" paradigm. For example, RoseTTAFold All-Atom extends modeling objects from proteins to nucleic acids, small molecules and their composite systems, achieving a more complete atomic-level description of biomacromolecular systems [44].

On this basis, the fields of structure prediction and molecular interaction modeling further show trends of openness and ecological development. Models such as Boltz-1 are committed to lowering the threshold for high-precision biomolecular modeling, and achieve efficient inference of protein interaction networks through a unified framework, thereby promoting the popularization of structural biology tools [45]. Meanwhile, models such as Chai-1 attempt to build a unified learning framework for molecular interactions, and further improve the ability to analyze complex interaction networks in living systems by integrating structure prediction and molecular interaction modeling capabilities [46]. These advances collectively mark that protein structure prediction is gradually evolving from a single structure analysis tool to a general modeling platform for multi-scale biomolecular systems.

### 3.2. De novo protein design — stepwise design based on structure-sequence correspondence

Due to the inherent limitations of traditional protein engineering strategies, scientists have been committed to exploring methods for de novo protein design over the past two decades [47-49]. Successful cases of traditional de novo protein design often lack universal design rules and principles [50]. Subsequently, scientists found that artificial intelligence models can not only achieve efficient and high-precision protein structure prediction, but also directly design protein sequences and structures according to expected structures or functions [13].

AI-based de novo protein design includes two categories: stepwise design based on structure-sequence correspondence and end-to-end generative design.

Among them, stepwise design usually first generates a protein backbone structure according to the target function [38], and then performs inverse folding from the backbone structure to obtain an amino acid sequence that can stably fold into that structure [51, 52]. Two important tools for the stepwise design strategy are diffusion models represented by RFdiffusion and the ProteinMPNN model. RFdiffusion can directly generate structures of novel proteins and protein complexes that can fold stably, based on given structural, functional or even interaction constraints [13]. ProteinMPNN can rapidly design amino acid sequences that are most likely to stably fold into a given three-dimensional protein backbone structure, and is applicable to monomeric proteins, symmetric multimers and more [51].

Diffusion models originally originated from the field of image generation. Their training method mainly involves iteratively adding a small amount of noise to images, so that the neural network can recover the original images from these noise-added samples [53, 54]. Diffusion models for generating protein backbone coordinates are a rapidly developing category, and their application scope includes the generation of diverse monomeric protein and protein complex structures [55]. RFdiffusion is a representative model among them.

After generating the protein backbone structure using diffusion models, inverse folding is performed from the protein structure. The difficulty lies in the complex mapping relationship between protein structure and sequence. The Ingraham team first attempted to introduce graph

DOI: 10.54254/2753-8818/2026.HZ35307

neural networks and autoregressive models to learn such mapping relationships [56]. Based on this idea, the Baker team developed the ProteinMPNN model, which uses a message-passing neural network to encode the protein backbone structure and predicts the amino acid composition of each site in combination with the local geometric environment [51]. Compared with previous RoseTTA models using physical energy functions and earlier models built by the Ingraham team, ProteinMPNN has higher design efficiency. In addition, ProteinMPNN also extends the application scope of neural network models in protein design to multimers and symmetric protein structure design. At present, inverse folding models represented by ProteinMPNN have become an important part of the standard de novo protein design workflow [57, 58].

### 3.3. De novo protein design — end-to-end design

Although the design workflow composed of RFdiffusion, ProteinMPNN and AlphaFold has significantly improved the success rate of de novo protein design, this strategy still essentially belongs to the paradigm of "structure generation — sequence design — structure validation". With the increasing demand for large-scale protein design, researchers have begun to explore "end-to-end" de novo protein design models.

In 2024, ESM3 realized the simultaneous processing of protein sequence, structure and function information under the same framework, and can directly generate proteins from functional requirements [59]. Proteins designed by it can exhibit functions similar to natural proteins with a sequence identity of less than 60%, which is equivalent to completing a span of 500 million years of natural evolution [59]. This success also proves that machine learning-driven de novo protein design can explore a large number of protein possibilities that natural evolution has failed to cover.

In the future, with the emergence of more multimodal foundation models and the continuous accumulation of experimental data, end-to-end protein design is expected to provide more efficient tools for fields such as enzyme engineering, biomanufacturing and synthetic biology.

## 4. Applications of protein engineering

### 4.1. Metabolic engineering

Proteins are the core functional execution units in metabolic networks, and their catalytic efficiency, substrate specificity and stability directly determine the operating efficiency of metabolic pathways. Therefore, protein engineering is an important foundation of metabolic engineering. Traditional metabolic engineering mainly achieves efficient synthesis of target products by mining natural enzyme resources and heterologously expressing key enzymes [60-62]. With the development of artificial intelligence, metabolic engineering in recent years has gradually shifted from optimizing existing pathways to designing pathways. The BioNavi-NP system released in 2022 was designed to address the problem that the biosynthetic pathways of a large number of natural products have not been resolved [63]. It achieved prediction of product biosynthetic pathways through methods such as Transformer neural networks, and achieved a 90.2% pathway search success rate in tests. AI-driven metabolic engineering has led to the discovery of more potential biosynthetic routes, and steps such as protein design and screening in pathways have also entered the era of intelligent design with the development of tools such as AlphaFold, RFdiffusion and ProteinMPNN.

DOI: 10.54254/2753-8818/2026.HZ35307

### 4.2. Enzyme engineering

As one of the most important fields in protein engineering, enzyme engineering often relies on directed evolution strategies, combined with the exploration of the relationship between enzyme structure and function, to improve enzyme performance. Early enzyme engineering focused on improving enzyme stability and catalytic efficiency, thereby enhancing its application effect in industrial production scenarios [64-67]. With the accumulation of enzyme function optimization and remodeling work, enzyme engineering has gradually broken through the limitations of natural enzyme functions. For example, engineered P450 has achieved dual optimization of expanding the substrate spectrum and improving reaction activity of the catalytic system [68]. In 2016, the modification of cytochrome c from *R. marinus* by Frances Arnold's team went further, realizing the synthesis of organosilicon compounds in living cells, which marked the first time that enzyme engineering created a completely new chemical bond formation reaction [69]. Research on the biosynthesis of natural products has also continuously discovered new enzymes with unique catalytic potential [70, 71].

However, as the demand for enzyme engineering in the fields of bioengineering and synthetic biology continues to be refined and increased, enzyme engineering based on traditional protein engineering ideas has limited efficiency in processing large-scale sequence and structure data. The development of AI has become a new driving force for the continued development of enzyme engineering. In 2008, David Baker's team first used RoseTTA to de novo design a series of artificial enzymes [72]. However, the activity of early artificial enzymes was far lower than that of natural enzymes, requiring further directed evolution to achieve activity improvement. In 2012, through a stepwise workflow of first improving the stability of artificial enzymes and then performing directed evolution, Baker's team increased the enzyme activity of multiple Kemp eliminases designed in 2008 by more than 1000 times from the initial computationally designed version, reaching an activity level close to that of natural enzymes [73]. This further proves the feasibility of computationally designed artificial enzymes, but also exposes the shortcomings of current algorithms, namely the difficulty in considering the dynamics of protein structures and the dynamic conformational changes during enzyme-substrate interactions. With the expansion of de novo protein design ideas and the optimization of deep learning models, artificial enzyme design has also developed rapidly. In 2025, Baker's team used models including RFdiffusion and ProteinMPNN, and based on reaction mechanism modeling, designed serine hydrolases and successfully constructed the Ser-His-Asp catalytic triad, with activity close to the level of natural enzymes [74].

After completing the AI design of high-activity artificial enzymes, Baker's team did not stop exploring. In the original workflow, steps such as selection of catalytic residues and protein backbone design needed to be completed first, which still had the disadvantages of insufficient efficiency and heavy workload. RFdiffusion2 launched in 2026 has brought a turning point to this situation. It can complete backbone design based on active site coordinates, substrate types and transition state information, and then obtain the final sequence through inverse folding [75]. This marks that automated design from manually constructed theozymes to enzymes has become possible, laying the foundation for fully automated "given chemical reaction — AI-generated enzyme" processes.

### 4.3. Antibody engineering

Antibody engineering, a branch of protein engineering, is one of the most active directions in modern biomedicine. Traditional antibody engineering ideas are similar to the directed evolution

DOI: 10.54254/2753-8818/2026.HZ35307

strategy in protein engineering, including steps such as antibody library construction, display screening and directed evolution. Its principle is to simulate the antibody production and affinity maturation process in the natural immune system, thereby obtaining candidate antibodies with specific binding capacity [76]. Since the establishment of phage display technology, researchers have successively developed methods such as humanization, fully human antibody screening and *in vitro* affinity maturation, which have greatly promoted the development of antibody drugs [77]. However, traditional antibody engineering methods are highly dependent on large-scale random mutagenesis and experimental screening, with long R&D cycles and high costs, and the process is close to a black box, with limited utilization of antigen structure information [76, 78]. In recent years, breakthroughs in artificial intelligence in the fields of protein structure prediction and de novo design have also brought new ideas to antibody engineering, providing a new technical path for efficiently obtaining high-performance antibodies.

In traditional antibody engineering, researchers often focus on screening antibodies with high affinity. However, with the deepening of antibody drug development, properties such as non-specific binding, self-aggregation, solubility and pharmacokinetics of antibodies have also become important reference factors determining the clinical translation potential of antibodies [79]. Traditional antibody engineering centered on directed evolution strategies is difficult to meet the actual demand for simultaneous optimization of multiple properties. To solve this problem, machine learning and deep learning have been introduced into the field of antibody design. In 2021, Mason's team used the HER2 antibody trastuzumab as a model and realized direct prediction of antigen specificity from antibody sequences using convolutional neural networks [80]. Computational antibody design has also developed into the most cutting-edge research direction in antibody engineering.

## 5. Conclusion

As an important bridge connecting molecular biology, biochemistry and biotechnology, protein engineering has become one of the core technologies driving the development of drug research and development, agricultural improvement and synthetic biology. Protein function essentially originates from the correspondence among amino acid sequences, three-dimensional structures and biological functions. For a long time, researchers have mainly relied on traditional strategies dominated by directed evolution and rational design to modify natural proteins. Through continuous accumulation of experimental data and mechanistic understanding, important breakthroughs have been achieved such as improved enzyme activity, expanded substrate spectrum, optimized antibodies and reconstructed metabolic pathways. However, the huge complexity of protein sequence space and the incomplete understanding of sequence-structure-function relationships have gradually caused bottlenecks in the development and application of traditional protein engineering.

In recent years, the rapid development of artificial intelligence technology has injected new vitality into protein engineering. Deep learning models represented by AlphaFold2 have significantly improved the accuracy and efficiency of protein structure prediction, making it possible to resolve the correspondence between protein sequences and structures [10]. Meanwhile, scientists have also set their sights on the design of entirely new proteins. Generative models such as RFdiffusion and ProteinMPNN have greatly promoted the development of de novo protein design, enabling protein engineering to break out of the framework of natural proteins [12, 13, 51]. On this basis, multimodal foundation models such as ESM3 have realized unified modeling of sequence, structure and function information, making it possible to directly generate candidate proteins from functional requirements [59]. The rapid development of artificial intelligence has driven the

DOI: 10.54254/2753-8818/2026.HZ35307

transformation of protein engineering from experience-driven and trial-and-error driven to data-driven and prediction-driven.

Nevertheless, the addition of artificial intelligence is not intended to completely replace traditional protein engineering, but to deeply integrate with rational design and directed evolution to form a new DTBL paradigm. Machine learning models can identify potential patterns from massive data, assisting researchers in predicting key mutation sites, optimizing screening strategies and generating candidate proteins; while experimental verification and directed evolution continue to undertake the important tasks of function confirmation and performance optimization. The two promote each other and jointly constitute the development framework of modern protein engineering. In the future, with the continuous accumulation of high-quality experimental data and the development of automated experimental platforms, closed-loop iteration between artificial intelligence models and experimental systems is expected to further improve protein design efficiency [35].

## References


[1] Ebrahimi, S. B., & Samanta, D. (2023). Engineering protein-based therapeutics through structural and chemical design. *Nat Commun, 14*(1), 2411.

[2] Koh, H. Y., et al. (2025). AI-driven protein design. *Nature Reviews Bioengineering, 3*(12), 1034–1056.

[3] Webber, M. J., & Pashuck, E. T. (2021). (Macro)molecular self-assembly for hydrogel drug delivery. *Adv Drug Deliv Rev, 172*, 275–295.

[4] Weigmann, K. F. G., Bornscheuer, U. T., & Doerr, M. (2026). Advances and critical evaluation of autonomous protein engineering: Towards transparent, accessible, and reproducible platforms. *Current Opinion in Biotechnology, 97*, 103395.

[5] Dietzen, D. J. (2018). 13 - Amino acids, peptides, and proteins. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Principles and applications of molecular diagnostics* (pp. 345–380). Elsevier.

[6] Schmidt, M. F. (2022). Peptides and proteins. In M. F. Schmidt (Ed.), *Chemical biology: And drug discovery* (pp. 91–123). Springer Berlin Heidelberg.

[7] Zhou, L., et al. (2024). Unlocking the potential of enzyme engineering via rational computational design strategies. *Biotechnology Advances, 73*, 108376.

[8] Wang, Y., et al. (2021). Directed evolution: Methodologies and applications. *Chemical Reviews, 121*(20), 12384–12444.

[9] Xiong, W., et al. (2021). Protein engineering design from directed evolution to de novo synthesis. *Biochemical Engineering Journal, 174*, 108096.

[10] Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

[11] Baek, M., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science, 373*(6557), 871–876.

[12] Watson, J. L., et al. (2022). Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022.12.09.519842.

[13] Watson, J. L., et al. (2023). De novo design of protein structure and function with RFdiffusion. *Nature, 620*(7976), 1089–1100.

[14] Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature, 346*(6287), 818–822.

[15] Liao, H., McKenzie, T., & Hageman, R. (1986). Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proceedings of the National Academy of Sciences, 83*(3), 576–580.

[16] Jäckel, C., Kast, P., & Hilvert, D. (2008). Protein design by directed evolution. *Annual Review of Biophysics, 37*(37), 153–173.

[17] Leung, D. W., Chen, E. Y., & Goeddel, D. V. (1989). A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction.

[18] Zaccolo, M., et al. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *Journal of Molecular Biology, 255*(4), 589–603.

[19] Vanhercke, T., et al. (2005). Reducing mutational bias in random protein libraries. *Analytical Biochemistry, 339*(1), 9–14.

DOI: 10.54254/2753-8818/2026.HZ35307

[20] Stemmer, W. P. (1994). DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America, 91*(22), 10747–10751.
[21] Joern, J. M., Meinhold, P., & Arnold, F. H. (2002). Analysis of shuffled gene libraries. *Journal of Molecular Biology, 316*(3), 643–656.
[22] Xiao, H., Bao, Z., & Zhao, H. (2015). High throughput screening and selection methods for directed enzyme evolution. *Industrial & Engineering Chemistry Research, 54*(16), 4011–4020.
[23] Sellés Vidal, L., et al. (2023). A primer to directed evolution: Current methodologies and future directions. *RSC Chemical Biology, 4*(4), 271–291.
[24] Woolfson, D. N. (2021). A brief history of de novo protein design: Minimal, rational, and computational. *Journal of Molecular Biology, 433*(20), 167160.
[25] Chubb, J. J., Boyle, A. L., & Albanese, K. I. (2026). Rational protein design. *Current Opinion in Structural Biology, 97*, Article 103224.
[26] Yu, H., et al. (2022). Hot spots-making directed evolution easier. *Biotechnology Advances, 56*, Article 107926.
[27] Li, F., et al. (2023). Redesigning an (R)-selective transaminase for the efficient synthesis of pharmaceutical N-heterocyclic amines. *ACS Catalysis, 13*(1), 422–432.
[28] Wang, Z., et al. (2022). Computational redesign of the substrate binding pocket of glutamate dehydrogenase for efficient synthesis of noncanonical l-amino acids. *ACS Catalysis, 12*(21), 13619–13629.
[29] Toscano, M. D., Woycechowsky, K. J., & Hilvert, D. (2007). Minimalist active-site redesign: Teaching old enzymes new tricks. *Angewandte Chemie International Edition, 46*(18), 3212–3236.
[30] Liu, N., et al. (2021). Crystal structures and catalytic mechanism of l-erythro-3,5-diaminohexanoate dehydrogenase and rational engineering for asymmetric synthesis of β-amino acids. *Angewandte Chemie International Edition, 60*(18), 10203–10210.
[31] Calvó-Tusell, C., et al. (2022). Reversing the enantioselectivity of enzymatic carbene N–H insertion through mechanism-guided protein engineering. *ChemRxiv, 2022*(0818).
[32] Heinemann, P. M., Armbruster, D., & Hauer, B. (2021). Active-site loop variations adjust activity and selectivity of the cumene dioxygenase. *Nature Communications, 12*(1), 1095.
[33] Hoque, M. A., et al. (2017). Stepwise loop insertion strategy for active site remodeling to generate novel enzyme functions. *ACS Chemical Biology, 12*(5), 1188–1193.
[34] Li, Z., et al. (2022). Flexibility regulation of loops surrounding the tunnel entrance in cytochrome P450 enhanced substrate access substantially. *ACS Catalysis, 12*(20), 12800–12808.
[35] Yang, W., et al. (2026). The past, present and future of de novo protein design. *Nature, 652*(8112), 1139–1152.
[36] Huang, P. S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature, 537*(7620), 320–327.
[37] Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Sci, 28*(4), 678–683.
[38] Kortemme, T. (2024). De novo protein design—From new structures to programmable functions. *Cell, 187*(3), 526–544.
[39] Shoemaker, S. C., & Ando, N. (2018). X-rays in the cryo-electron microscopy era: Structural biology's dynamic future. *Biochemistry, 57*(3), 277–285.
[40] Gauto, D. F., et al. (2019). Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nature Communications, 10*(1), 2697.
[41] Abramson, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*(8016), 493–500.
[42] Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science, 379*(6637), 1123–1130.
[43] Evans, R., et al. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, p. 2021.10.04.463034.
[44] Krishna, R., et al. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science, 384*(6693), p. eadl2528.
[45] Wohlwend, J., et al. (2025). Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*.
[46] Chai Discovery, t., et al. (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*, p. 2024.10.10.615955.
[47] Kuhlman, B., et al. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science, 302*(5649), pp. 1364–1368.
[48] Siegel, J. B., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science, 329*(5989), pp. 309–313.

DOI: 10.54254/2753-8818/2026.HZ35307

[49] Bale, J. B., et al. (2016). Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science, 353*(6297), 389–94.
[50] Koga, N., et al. (2012). Principles for designing ideal protein structures. *Nature, 491*(7423), 222–7.
[51] Dauparas, J., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. *Science, 378*(6615), 49–56.
[52] Liu, Y., et al. (2025). Enhancing functional proteins through multimodal inverse folding with ABACUS-T. *Nature Communications, 16*, 10177.
[53] Song, Y., et al. (2020). Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456.
[54] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Article 574). Curran Associates Inc.
[55] Anand, N., & Achim, T. (2022). Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *ArXiv*, abs/2205.15019.
[56] Ingraham, J., et al. (2019). Generative models for graph-based protein design. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Article 1417). Curran Associates Inc.
[57] Bennett, N. R., et al. (2023). Improving de novo protein binder design with deep learning. *Nature Communications, 14*(1), 2625.
[58] Fox, D. R., et al. (2025). Code to complex: AI-driven *de novo* binder design. *Structure, 33*(10), 1631–1642.
[59] Hayes, T., et al. (2025). Simulating 500 million years of evolution with a language model. *Science, 387*(6736), 850–858.
[60] Brat, D., Boles, E., & Wiedemann, B. (2009). Functional expression of a bacterial xylose isomerase in Saccharomyces cerevisiae. *Applied and Environmental Microbiology, 75*(8), 2304–2311.
[61] Karimäki, J., et al. (2004). Engineering the substrate specificity of xylose isomerase. *Protein Engineering Design and Selection, 17*(12), 861–869.
[62] Alper, H., & Stephanopoulos, G. (2009). Engineering for biofuels: Exploiting innate microbial capacity or importing biosynthetic potential? *Nature Reviews Microbiology, 7*(10), 715–723.
[63] Zheng, S., et al. (2022). Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nature Communications, 13*(1), 3342.
[64] Chen, K. Q., & Arnold, F. H. (1991). Enzyme engineering for nonaqueous solvents: Random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology (New York), 9*(11), 1073–1077.
[65] Chen, K., & Arnold, F. H. (1993). Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America, 90*(12), 5618–5622.
[66] Zhao, H., & Arnold, F. H. (1999). Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Engineering, 12*(1), 47–53.
[67] Takagi, H., et al. (2000). Engineering subtilisin E for enhanced stability and activity in polar organic solvents. *J Biochem, 127*(4), 617–625.
[68] Fasan, R., et al. (2007). Engineered alkane-hydroxylating cytochrome P450BM3 exhibiting nativelike catalytic properties. *Angewandte Chemie International Edition, 46*(44), 8414–8418.
[69] Kan, S. B., et al. (2016). Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science, 354*(6315), 1048–1051.
[70] Gao, L., et al. (2020). FAD-dependent enzyme-catalysed intermolecular [4+2] cycloaddition in natural product biosynthesis. *Nat Chem, 12*(7), 620–628.
[71] Jiang, B., et al. (2024). Characterization and heterologous reconstitution of Taxus biosynthetic enzymes leading to baccatin III. *Science, 383*(6683), 622–629.
[72] Röthlisberger, D., et al. (2008). Kemp elimination catalysts by computational enzyme design. *Nature, 453*(7192), 190–195.
[73] Khersonsky, O., et al. (2012). Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci U S A, 109*(26), 10358–63.
[74] Lauko, A., et al. (2025). Computational design of serine hydrolases. *Science, 388*(6744), eadu2454.
[75] Ahern, W., et al. (2026). Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods, 23*(1), 96–105.
[76] Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat Biotechnol, 23*(9), 1105–16.
[77] Carter, P. J. (2006). Potent antibody therapeutics by design. *Nat Rev Immunol, 6*(5), 343–57.

DOI: 10.54254/2753-8818/2026.HZ35307

[78] Schier, R., et al. (1996). Isolation of picomolar affinity anti-c-erbB-2 single-chain Fv by molecular evolution of the complementarity determining regions in the center of the antibody binding site. *J Mol Biol, 263*(4), 551–67.
[79] Starr, C. G., & Tessier, P. M. (2019). Selecting and engineering monoclonal antibodies with drug-like specificity. *Current Opinion in Biotechnology, 60*, 119–127.
[80] Mason, D. M., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering, 5*(6), 600–612.