Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

Advancing microbial production through artificial intelligence-aided biology

Xinyu Gong [a], Jianli Zhang [a], Qi Gan [a], Yuxi Teng [a], Jixin Hou [b], Yanjun Lyu [c], Zhengliang Liu [d], Zihao Wu [d], Runpeng Dai [e], Yusong Zou [a], Xianqiao Wang [b], Dajiang Zhu [c], Hongtu Zhu [e], Tianming Liu [d], Yajun Yan [a,*]

[a] School of Chemical, Materials, and Biomedical Engineering, College of Engineering, The University of Georgia, Athens, GA 30602, USA
[b] School of ECAM, College of Engineering, University of Georgia, Athens, GA 30602, USA
[c] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington 76019, USA
[d] School of Computing, The University of Georgia, Athens, GA 30602, USA
[e] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

ARTICLE INFO

Keywords:
Microbial production
Synthetic biology
Genome annotation
Enzyme function prediction
Artificial protein design
Pathway prediction
Artificial intelligence (AI)
Large language models (LLMs)

ABSTRACT

Microbial cell factories (MCFs) have been leveraged to construct sustainable platforms for value-added compound production. To optimize metabolism and reach optimal productivity, synthetic biology has developed various genetic devices to engineer microbial systems by gene editing, high-throughput protein engineering, and dynamic regulation. However, current synthetic biology methodologies still rely heavily on manual design, laborious testing, and exhaustive analysis. The emerging interdisciplinary field of artificial intelligence (AI) and biology has become pivotal in addressing the remaining challenges. AI-aided microbial production harnesses the power of processing, learning, and predicting vast amounts of biological data within seconds, providing outputs with high probability. With well-trained AI models, the conventional Design-Build-Test (DBT) cycle has been transformed into a multidimensional Design-Build-Test-Learn-Predict (DBTLP) workflow, leading to significantly improved operational efficiency and reduced labor consumption. Here, we comprehensively review the main components and recent advances in AI-aided microbial production, focusing on genome annotation, AI-aided protein engineering, artificial functional protein design, and AI-enabled pathway prediction. Finally, we discuss the challenges of integrating novel AI techniques into biology and propose the potential of large language models (LLMs) in advancing microbial production.

1. Introduction

Microbial production provides a sustainable and eco-friendly approach to producing valuable products, like bulk chemicals, pharmaceuticals, and plant natural products (Kim et al., 2023; Son et al., 2023). Compared to traditional chemical synthesis, which uses toxic raw materials and causes severe industrial pollutants, microbial production utilizes simple and inexpensive carbon sources, such as sugar and lignin (Cai et al., 2023; Zhang et al., 2021a). It then assembles appropriate enzymes to achieve green production in microbial cell factories (MCFs). However, conventional microbial production relies on the manual Design-Build-Test (DBT) cycle. This process could be extremely laborious and challenging given the complexity of gene mining, protein discovery/engineering, and biosynthesis pathway construction. To assist microbial production, synthetic biology aims to design, construct, and manipulate biological devices to control biological networks in

Abbreviations: MCFs, Microbial Cell Factories; DBT, Design-Build-Test; DBTLP, Design-Build-Test-Learn-Predict; AI, Artificial Intelligence; LLMs, Large Language Models; CRISPR, Clustered Regularly Interspaced Palindromic Repeats; TFs, Transcriptional Factors; RNAi, RNA interference; ML, Machine Learning; CNN, Convolution Neural Networks; KO, KEGG Ortholog; GO, Gene Ontology; DAG, Directed Acyclic Graph; UniRef, UniProt Reference Clusters; PDB, Protein Data Bank; NLP, Natural Language Processing; MSAs, Multiple Sequence Alignments; EC number, enzyme commission number; GCNs, Graph Convolutional Neural Networks; LSTM, Long-Short-Term-Memory; RNNs, Recurrent Neural Networks; CPD, Computational Protein Design; AIGC, AI-Generated Content; L-HPA, *L*-Homophenylalanine; ER, ene-reductase; NPs, Natural Products.

[*] Corresponding author at: 302 East Campus Road, The University of Georgia, Athens, GA 30602, USA.
E-mail address: yajunyan@uga.edu (Y. Yan).

https://doi.org/10.1016/j.biotechadv.2024.108399
Received 3 January 2024; Received in revised form 20 May 2024; Accepted 23 June 2024
Available online 24 June 2024

multiple levels ranging from genes to proteins to pathways (Choi et al., 2019). The tunable biological devices are usually built based on standardized and modularized genetic elements, such as CRISPR, transcriptional factors (TFs), and nucleic acids (Teng et al., 2022). The CRISPR system is a precise gene editing technique that enables large-scale gene function screening as well as metabolic pathway reprogramming (Teng et al., 2023). Related CRISPR technologies like CRISPRa and CRISPRi act as up- or down-regulatory elements in biosynthesis pathway regulation (Fontana et al., 2020; Wang et al., 2021). Transcriptional factors-based biosensors inherit the regulatory and sensory properties of TFs, being mainly exerted in high-throughput protein screening and dynamic regulation of metabolic pathways (Gong et al., 2022; Jiang et al., 2022; Li et al., 2020). Nucleic acid-based biosensors such as RNA interference (RNAi) and riboswitch provide targeted control of gene expression in either an activated or a repressed manner (Wang and Simmel, 2022; Zhang et al., 2021b). Although microbial production has benefited from synthetic biology strategies, the manual DBT cycle is a time-consuming trial-and-error method that requires laborious laboratory characterizations. Moreover, as genomic, proteomic, and metabonomic data gradually expand, it is impossible to manually extract useful information from vast datasets efficiently and accurately.

The integration of artificial intelligence (AI) and biology has attracted significant attention and demonstrated great success in aiding multiple realms of microbial production, including genome mining, protein discovery/engineering, artificial protein design, and pathway prediction (Fig. 1) (Ferruz et al., 2023; Mullowney et al., 2023; Tan et al., 2023). With the assistance of well-trained AI models, microbial production has been transformed into a multidimensional Design-Build-Test-Learn-Predict (DBTLP) workflow (Fig. 1a). Firstly, for genome mining, abundant genome information can be interpreted by AI models for quick genome annotation (Li et al., 2022), aiding enzyme function prediction (Fig. 1b). Next, for protein discovery/engineering, AI-aided protein engineering can rapidly enrich the existing enzyme pools by computationally analyzing the active sites and substrate-binding pockets of enzymes to facilitate engineering and screening processes (Kouba et al., 2023). Meanwhile, artificial functional protein design is an alternative to enrich enzyme pools and is under the lead of AI techniques, creating novel proteins by computationally designed structures and sequences (Fig. 1c) (Lovelock et al., 2022). With enriched enzyme resources, pathway construction can also utilize AI models for pathway prediction, especially retrobiosynthesis (Yu et al., 2023a), after training on chemical or enzymatic reaction databases (Fig. 1d). Overall, AI-aided microbial production enables systematically analyzing and rapidly solving biological problems computationally from multiple dimensions. Particularly, AI significantly shortens the time and improves the efficiency of microbial production. Instead of having engineers manually sift through vast amounts of data, computational algorithms simplify this by scoring and learning from the given data, offering highly probable solutions automatically (Jang et al., 2022). For example, Lu et al. reported a machine learning (ML)-aided protein engineering method called MutCompute that uses a three-dimensional convolutional neural network (3D CNN) model harboring nine layers. To obtain an optimized PET hydrolase, the model computationally analyzed over 30 residues from PETase crystal structures and generated 159 potentially functional variants. After in-silico metagenesis screening, 29 variants were selected for experimental validation and FAST-PETase presented excellent PET depolymerization activities even for untreated plastics (Lu et al., 2022). This study successfully showcased the efficiency of applying ML-aided protein engineering.

Although AI techniques have already brought significant advancements in microbial production, the emergence of large language models (LLMs) and multimodal learning is bound to lead to a revolution in this field. Every computational model relies on high-quality training

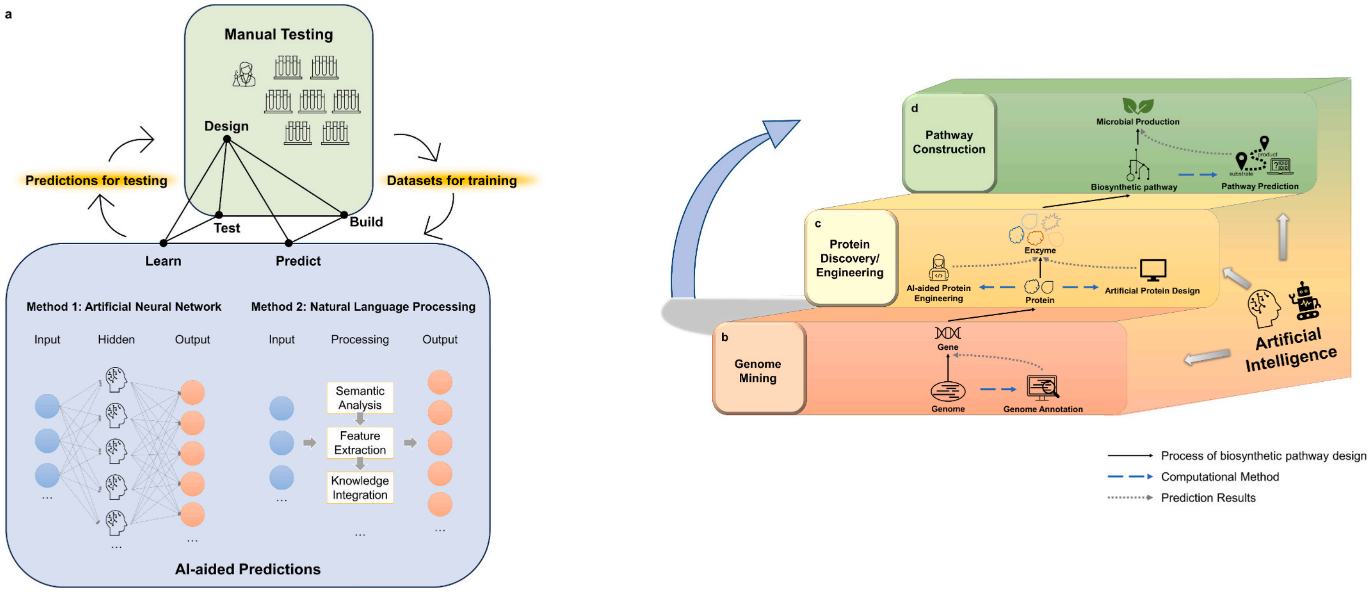

Fig. 1. AI-aided microbial production workflow and main processes. (a) Multidimensional Design-Build-Test-Learn-Predict (DBTLP) workflow. The conventional DBT cycle relies on manual testing, which is labor-intensive. The DBTLP workflow incorporates computational methods, such as artificial neural networks and natural language processing, to assist and accelerate the DBT cycle. (b) Genome mining is the process of searching the genome to identify genes with potential useful functions through genome annotation. AI facilitates genome annotation by developing computational algorithms to extract genetic features, thus efficiently identifying more functional genes in the short term. (c) Protein discovery/engineering is a process of identifying and creating more functional proteins to enrich enzyme resources for microbial production. One common strategy is AI-aided protein engineering which aims to utilize AI techniques to analyze and engineer existing enzymes to predict and screen potential variants for improved enzymatic properties. The other strategy is artificial protein design which seeks to create novel protein sequences or structures with desired functions computationally by learning from curated protein sequence and structure datasets, finally capable of creating novel proteins. (d) Pathway construction is the process of assembling a biosynthesis pathway for microbial production. AI-enabled pathway prediction tools can rapidly propose many potential pathways, reducing the need for tedious trial-and-error efforts. The AI models are trained in curated enzymatic or chemical reactions to learn the rules of chemical synthesis.

datasets. Biological data mainly comprises genomic and proteomic data in multiple modalities. Genome data includes DNA sequences with gene annotations in FASTA format. Protein data includes amino acid sequences in FASTA format and protein structures in PDB format. These datasets serve as valuable resources for training specialized biological LLMs to comprehend diverse biological contents Several biological LLMs have been successfully applied to sequence analysis, genome analysis, proteomics, and other domains, yielding predictions of invaluable bioinformation (Zhang et al., 2023a). The prominent models employed in these bioinformatic analyses are based on the transformer architecture, such as GPT-3.5, GPT-4, and BERT (Devlin et al., 2018; OpenAI, 2023; Talebi et al., 2023). Additionally, the GPT series has demonstrated its utility across broader domains, including education, agriculture, medicine, and biology with impressive performance and satisfactory results (Gong et al., 2023; Latif et al., 2023; Lee et al., 2023; Rezayi et al., 2023). We believe applying LLMs in microbial production scope could be similarly beneficial. Overall, this review highlights the current progress in AI-aided microbial production in the aspects of genome annotation, AI-aided protein engineering, artificial functional protein design, and AI-enabled pathway prediction. Furthermore, it discusses the challenges and potential of integrating LLMs to further propel microbial production.

2. Gene annotation, protein annotation, and enzyme function prediction

Genome resources are an underexplored natural treasure, housing numerous valuable enzymes and diverse functional systems awaiting to be discovered and harnessed. Genome annotation is an important procedure for identifying and labeling functional genomic features known as genes within genetic coding regions (Abril and Castellano, 2019). As the product of genes, proteins are macromolecules comprised of amino acids that collectively constitute functional metabolic networks and proteomes in organisms. Comprehensive annotation of genomic and proteomic features prompts the process of unlocking meaningful genetic characteristics. In genomes, annotatable features cover gene expression levels, regulatory elements loci, TF binding sites, and individual genetic variances. In proteomes, annotatable features encompass key residue identification, post-translational modification recognition, molecular interaction with residues, and 3D molecular structures (Reeves et al., 2009). Recent advances in sequencing and AI techniques enable unprecedentedly rapid access to numerous genomic and proteomic data at both sequence and structure-function levels, deciphering and elucidating intricate mechanisms of metabolic networks.

In recent years, biological datasets have grown tremendously in both size and variety. For instance, NCBI and EBI provide the most comprehensive assembled genome data across species for public use (Mitchell et al., 2020; Sayers et al., 2019). Concurrently, the KEGG Ortholog (KO) and Gene Ontology (GO) Consortium have curated and categorized gene and protein entities using specific rules to form their unique directed acyclic graph (DAG) systems based on entities' functions and interactions (Aleksander et al., 2023; Kanehisa and Goto, 2000). Regarding protein resources, UniProtKB has compiled extensive protein data with sequences and corresponding annotation information (UniProtConsortium, 2022), and SWISS-PROT has provided expertly curated protein annotations within the UniProtKB (Bairoch and Apweiler, 2000). BFD is another clustered protein sequence database created via accurate alignment, clustering 2.5 billion protein sequences from UniProtKB and SWISS-PROT (Steinegger et al., 2019; Steinegger and Söding, 2018). Like UniProt Reference Clusters (UniRef) (Suzek et al., 2015), a database clustered parts of UniProtKB sequences, BFD provides accurate information on protein properties and clustering identity, benefiting automotive data-driven deep model training. Beyond the protein sequence datasets, the Protein Data Bank (PDB) is a repository that provides experimentally validated and computationally predicted 3D protein and nucleic acid structures (Berman et al., 2000). Here, we present some popular biological datasets employed in AI-aided genome and proteome analysis in Table 1.

Furthermore, effective approaches are needed to fully utilize these fast-growing biological data resources. Recently developed AI models inspired by natural language processing (NLP) have enabled automatic, large-scale mining of these datasets for gene annotation, protein annotation, and enzyme function prediction, paving the way to better utilization of multi-modality biological data and providing more available enzymes for efficient microbial production. The multi-modality data often describes functional gene or protein entities from different perspectives, in which the connections are frequently overlooked in traditional DNA or protein annotation studies. By leveraging neural network models such as convolution networks (LeCun and Bengio, 1995), transformer (Vaswani et al., 2017), BERT (Devlin et al., 2018), and graph convolution networks (Kipf and Welling, 2016), recent studies have taken advantage of disparate modalities of genomic and proteomic data to derive deeper insights. Focusing on what microbial production can learn from biological datasets, this section covers the works from two main aspects: DNA elements annotation and protein annotation.

2.1. AI-aided DNA elements annotation

Conventional DNA sequence analysis adopts alignment methods like BLAST (Altschul et al., 1990) or k-mer-based approaches (Koonin et al., 2003). Earlier practices also employed convolution and recurrent neural networks to uncover local patterns and long-range dependencies in DNA sequence data (Amanatidis et al., 2022). As gene expression and regulation involve a variety of nucleotide-binding proteins cooperating with specific DNA or RNA segments, which could potentially be used for biosensor development, many CNN models have been developed to predict these binding sites and motifs. For example, Alipanahi et al. developed DeepBind, which utilized deep CNN to predict binding sites of DNA and RNA-binding proteins by detecting patterns within DNA or RNA sequences. DeepBind can perform downstream tasks like capturing RNA alternative splicing patterns regulated by RNA binding proteins and assessing the effects of deleterious genetic mutations on gene expression and TF binding (Alipanahi et al., 2015). Another study published by Wang et al. focused on predicting the binding affinity of TFs with DNA sequences by introducing the DeFine model, which leveraged deep CNNs to high-throughput classify TF-DNA binding and unbinding sites (Wang et al., 2018). However, these CNN-based methods are limited to extracting local sequence information, lacking the capability to capture distant semantic relationships within the genome.

Recent work has turned to NLP-inspired deep learning techniques for analyzing DNA sequences, considering parallels between natural language grammar and DNA syntax. With the rise of attention-based models in NLP research, transformer-based pre-trained language models like BERT have become prevalent. For instance, DNABERT adapts BERT for DNA sequences by using k-mer as tokens. The model was first pre-trained on a large DNA database using the masking technique and then fine-tuned for downstream tasks. By harnessing contextual information within DNA sequences, DNABERT achieved outstanding performance in predicting promoters and identifying TF binding sites (Ji et al., 2021). DNABERT-2, an updated version of DNABERT, incorporates dynamic byte pair coding instead of fixed-length k-mer tokenization and replaces positional embeddings with attention linear biases to address the input length limitation of DNABERT (Zhou et al., 2023). However, DNABERT-2 showed no substantial improvement over DNABERT. Another example demonstrated the success of combining the BERT and CNN in DNA sequence analysis. Lee et al. transformed pre-trained BERT to specialize for the DNA language domain to extract contextual features from input DNA sequences presented by fixed-length numerical vectors. To better extract features, 2D CNN was integrated to process the numerical vectors. This framework was then applied to identify DNA enhancers, an important type of regulatory element within the genome, exhibiting improved sensitivity and accuracy in capturing

Table 1
The publicly available database utilized by AI for genome annotation.

Name	Abbreviation	Size	Modality	Description	Reference
GenBank at NCBI	GenBank	Over 6.25 trillion	Nucleotide base pairs	The GenBank database, which is created and managed by the National Center for Biotechnology Information (NCBI), contains a large collection of nucleic acid sequences along with bibliographic and biological annotations.	(Sayers et al., 2019)
European Nucleotide Archive	ENA	Over 2.7 billion	Annotated nucleotide sequences	ENA is a comprehensive database that stores and provides open access to nucleotide sequence data and associated metadata. ENA offers a range of services for submitting, searching, and downloading sequence records and enables the global sharing and discovery of genomic data.	(Burgin et al., 2022)
DNA Data Bank of Japan	DDBJ	Over 2.7 billion	Annotated nucleotide sequences	DDBJ mainly operates the DDBJ Sequence Read Archive (DRA), a database that stores sequencing raw data and sequence alignment information by using high-throughput sequencing platforms and analysis pipelines.	(Tanizawa et al., 2022)
Mgnify (formerly EBI Metagenomics)	MGnify	Over 1.9 million	Microbiome datasets	MGnify, a free portal targeted to microbiome data capable of data analysis, investigation, and storage, currently includes three novel analysis pipelines with additional methods for taxonomic classification.	(Mitchell et al., 2020)
KEGG ORTHOLOGY Database	KO	Unknown	Annotated nucleotide and protein sequences	KO database contains functional orthologs, which are manually defined groups of genes or proteins that share the same function across organisms. These functional orthologs are identified in the context of KEGG's molecular networks, including maps of pathways, BRITE functional hierarchies, and modules. Each node in a network is assigned a KO identifier or K number, which defines a functional ortholog group.	(Kanehisa et al., 2022)
Gene Ontology	GO	7.5 million	Annotated nucleotide and protein sequences	The Gene Ontology (GO) knowledgebase contains a comprehensive, structured, and computer-readable representation of gene functions across all cellular organisms and viruses. It standardizes the description of gene roles using a consistent vocabulary to annotate gene products.	(Aleksander et al., 2023)
UniProt Knowledgebase	UniProtKB	Over 227 million	Annotated protein sequences	UniPortKB is the latest version of UniPort, including two protein sets, UniProtKB/Swiss-Port, and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot houses the experimentally verified or computationally predicted annotated protein sequence. UniProtKB/TrEMBL contains the automatic annotations generated by the Association-Rule-Based Annotator (ARBA).	(UniProtConsortium, 2022)
UniProt Reference Clusters (UniRef)					
	UniRef100	Over 44 million	Annotated protein sequences	The UniRef100 database combines identical sequences and sub-sequences of 11 residues or more from proteins across all organisms into consolidated entries.	(Suzek et al., 2015)
	UniRef90	Over 25 million	Annotated protein sequences	The UniRef90 database is constructed by clustering UniRef100 sequences using the MMseqs2 algorithm. The clustering parameters are set such that all sequences within a given cluster share a minimum of 90% sequence identity and 80% overlap in alignment with the longest sequence in that cluster, termed the seed sequence.	(Suzek et al., 2015)
	UniRef50	Over 11 million	Annotated protein sequences	The UniRef50 database is constructed by extracting the seed sequences from UniRef90 and grouping them into clusters if they meet the criteria of having at least 50% sequence identity and 80% alignment overlap with the longest sequence within that cluster.	(Suzek et al., 2015)
Research Collaboratory for Structural Bioinformatics Protein Data Bank	RCSB PDB	Over 200 thousand experimentally determined structures and over 1 million computationally predicted structures	Protein structures	The PDB serves as a comprehensive repository providing free access to experimentally solved and computationally modeled protein structures.	(Burley et al., 2022)
Big Fantastic Database	BFD	2.5 billion	Protein sequences	BFD was generated by clustering 2.5 billion protein sequences sourced from various protein databases, including Uniprot, TrEMBL, Swissprot, etc.	(Steinegger and Söding, 2018)

meaningful features (Le et al., 2021).

The discovery of functional genetic elements like TF binding sequences and DNA enhancers by DNA sequence annotation provides a comprehensive understanding of gene regulations and molecular interactions, inspiring the construction of novel biological systems and devices.

2.2. AI-aided protein annotation

The central dogma states that protein is the product of gene translation. Inheriting genetic information as well, the analysis of protein sequences can be considered an extension of the analysis of DNA sequences (von Heijne, 1991). However, protein sequences are more complex, as the combinations of 20 amino acids confer diverse structural and functional features. Due to remarkable advancements in machine learning and deep learning, protein feature prediction now provides a high-throughput method compared to traditional experimental approaches.

2.2.1. AI-aided protein structural feature annotation

Protein sequences adhere to precise 3D structures that deliver structural features related to specific protein interactions or functions. A key goal of protein structural feature annotation is to discover and understand these structural features, which could demonstrate the mechanism underlying protein functions. AlphaFold models are pioneers in performing protein structure prediction tasks, providing predicted 3D structural information for proteins. The first version of the AlphaFold model, known as AlphaFold1, is a deep-learning model. It interprets amino acid sequences, applies multiple sequence alignments (MSAs), and utilizes deep residual neural networks to forecast the distances between pairs of residues. These distance predictions are then used to generate protein structures via gradient descent optimization (Ruff and Pappu, 2021). AlphaFold1 successfully predicted many high-accuracy structures in the 13th Critical Assessment of Protein Structure Prediction (CASP13). The updated version, AlphaFold2 incorporates new neural network architectures and pairwise features. In CASP14, AlphaFold2 demonstrated accuracy rivaling experimental structures in many cases, substantially surpassing other methods. AlphaFold2 considered protein structure predictions as a graph inference problem within a 3D spatial context, where the graph's edges were determined by the proximity of residues. A key component is the Evoformer building block, which jointly embeds MSAs and pairwise features. The model also integrated physical and geometric constraints inherent to protein structures (Casadevall et al., 2023; Ismi and Pulungan, 2022).

As computational techniques have rapidly advanced, other protein structural feature annotation methods are blooming out. Leveraging the self-supervised learning ideal for language models, ProtTrans models include six different transformer-based protein language models that use protein sequence information alone to capture some protein biophysical features. Two of the models are auto-regressive models called Transformer-XL (Dai, 2019) and XLNet (Yang et al., 2019b). The other four are auto-encoder models named BERT(Devlin et al., 2018), Albert (Lan et al., 2019), Electra (Clark et al., 2020), and T5 (Raffel et al., 2020). The authors trained six models on up to 393 billion amino acids from the UniRef and BFD databases to create context-aware embeddings. The models were validated on downstream tasks including the prediction of secondary structure and protein cellular localization. Notably, when predicting protein secondary structure, the most informative embeddings from ProtT5 surpassed the previous state-of-the-art for the first time, without requiring MSAs or evolutionary information, thus avoiding costly database searches (Elnaggar et al., 2021). The development of protein structural feature annotation methods provides fundamental knowledge and valuable insights into the relationship between protein sequences and structures, facilitating progress in protein design and engineering.

2.2.2. AI-aided protein function annotation and enzyme function prediction

Enzymes are functional proteins capable of catalyzing certain reactions. The sequence-based method BLASTp (Altschul et al., 1990), a very classic sequence alignment tool, could be used for searching protein sequences with high similarities. Based on the sequence homology, we can infer the functions of the query proteins. However, this algorithm is only efficient in predicting the functions of homology (Ejigu and Jung, 2020). As the volume of genome data sets rapidly increases, the unannotated and misannotated proteins are increasing (Schnoes et al., 2009). More advanced and effective methods are needed to meet the demands. To improve enzyme availability for microbial production, AI-aided protein function annotation and enzyme function prediction could directly guide the exploration of functional enzymes in a high-throughput and high-accuracy manner (Ardern et al., 2023; Singh et al., 2016).

In some protein databases, enzymes were cataloged by the enzyme commission number (EC number), which hierarchically reflects the enzyme categories and functions (McDonald and Tipton, 2023). One common type of prediction method uses protein sequence as input and EC number as output. This method analyzes sequence features and extracts motifs from given protein sequences. Combining computational algorithms, the deep learning-based computational framework can be applied to predict protein EC numbers with high quality. Ryu et al. reported a deep learning-based prediction method, DeepEC, which was mainly built with three CNNs for fourth-level EC number prediction. DeepEC was trained on a curated dataset collecting data from both Swiss-Prot and TrEMBL protein databases by examining 4 hyper-parameters and was evaluated with negative testing for enhanced accuracy. However, this model cannot predict enzyme types that have less than 10 curated sequences in the dataset. To remedy this deficiency, homology analysis was used to deal with these exceptions. The proof-of-concept study of DeepEC through functional prediction of an *E. coli* enzyme, YgbJ, proved that the model was good at predicting promiscuous enzymes with multiple functions, especially two functions. Moreover, compared with other prediction tools, DeepEC was more reliable and efficient and required a small disk space (0.045GB) that could flexibly be integrated into third-party software (Ryu et al., 2019). The continuous advancement in the computational field also leads to a better prediction performance of functional enzymes. Yu et al. recently published an ML-based algorithm called CLEAN, using contrastive learning for enzyme function prediction, which was more accurate than the previous ML-based annotation tools. This model was trained on UniProt to form an embedding space of enzymes by calculating Euclidean distance and improved the accuracy by applying the contrastive loss function. The enzymes with the same EC number showed a small Euclidean distance, while enzymes with different EC numbers showed a large Euclidean distance. By using the query protein sequence as input, the sequence was first turned into a vector or matrix and started at the average point of the learned embedding space, and then the Euclidean distance was calculated to find the closest EC cluster. The benchmark study indicated that the prediction performance of CLEAN exceeded BLASTp in three multilabel accuracy metrics. CLEAN was validated not only in silicon but also in vitro. The in vitro validation of the halogenases successfully proved the accuracy in predicting, identifying, and correcting the catalytic functions of MJ1651, TTHA0338, and SsFlA, respectively (Yu et al., 2023b).

Apart from using the EC number for enzyme function classification, the GO Consortium, known as the GO term, is another protein classification scheme and is widely used to describe the enzyme function from three aspects, including molecular function, biological process, and cellular component (Aleksander et al., 2023). The GO term is also used as the output in some computational models for enzyme function prediction. Due to the exquisite 3D protein structure holding tremendous structural biology information, the structural features of enzymes are an ideal input to investigate the structure-function relationship. Gligorijević et al. demonstrated the performance of an enzyme function predictor, DeepFRI, by using three-layer graph convolutional neural networks (GCNs) to analyze the graphic features extracted from protein sequences or structures. DeepFRI had two ways to process the inputs. Protein sequences were processed by a Long-Short-Term-Memory (LSTM) Language Model to get residue-level features and its Cα–Cα contact map. For protein structure (e.g., from PDB), the model extracted the useful sequence information and built its Cα–Cα contact map. The Cα–Cα contact maps were then used as input for the next stage of GCNs, finally giving GO terms as output. This model was trained on the test set with only experimentally determined structures, exhibiting high denoising ability, high tolerance, and high precision. In addition to predicting enzyme function in GO terms, DeepFRI can also show output in EC numbers and can accurately predict the binding site of the input protein, indicating the unprecedented performance of GCNs not only in predicting protein function but also in analyzing protein structure and molecular interactions (Gligorijević et al., 2021). These studies showed the benefits of adapting multi-modality data of protein features in enzyme function prediction.

Drawing an analogy between protein sequence and natural language, protein language models have emerged to aid protein function annotation with protein sequences serving as input (Ofer et al., 2021). The language models and algorithms, which could well comprehend human languages, are now paving the way for protein function annotation. One such self-supervised BERT-based deep language model is ProteinBERT, specialized and tailored explicitly for protein sequences to predict protein functions. Its pre-training strategies include language modeling of protein representations and the prediction of corresponding GO annotations for protein functions. ProteinBERT naturally captures local and global protein representations, enabling seamless end-to-end processing of various input and output types (Brandes et al., 2022). Another notable protein language model is TALE, which jointly embeds protein 1D sequence features and GO terms as labels into a shared latent space. Leveraging the transformer-based architecture for protein function annotation, TALE captures global patterns within protein sequences and exhibits robust generalizability even for novel or rare protein sequences. Furthermore, TALE+, a combination model integrating TALE and DIAMOND, outperforms competing methods when using only the sequence as input. The models with high accuracy and generalizability can annotate protein functions in a high-throughput manner, annotating 1000 sequences in a few minutes (Cao and Shen, 2021). The models' interpretability and generalizability to new sequences make it well-suited for practical protein function annotation.

3. AI-aided protein engineering

Protein engineering is a powerful approach to expanding the pool of available enzymes for microbial production, commonly through directed evolution or rational design, altering enzyme activities and specificities. However, manually constructing and screening a protein mutagenesis library requires extensive analysis of protein structures as well as laborious experimental testing of hundreds of variants. Remarkably, AI-aided protein engineering has emerged as a popular interdisciplinary technique to alleviate the workload associated with manual protein engineering and screening, accelerating the DBT cycle by developing, training, and applying computational models (Fig. 2a). The primary appeal of applying computational models in protein engineering lies in their capacity to creatively generate variants. After the training on protein datasets, computational algorithms can rapidly generate predictions for new variants by assimilating knowledge from known proteins and characterized variants (Yang et al., 2019a). Extensive datasets containing information on protein sequences, structures, functional properties, and biophysical data serve as valuable assets for model training, exemplified by publicly available resources like UniProtKB (UniProtConsortium, 2019), Brenda (Jeske et al., 2019), PDB (Burley et al., 2019), InterPro (Mitchell et al., 2019), and others. Thus, AI-aided protein engineering strategies have been widely applied to enhance protein activities for improved production yields, engineer synthetic biology tools for pathway regulation and reprogramming, or expand enzyme promiscuities for novel enzymatic reactions (Table 2).

Protein sequences play a vital role in AI-aided protein engineering, serving as foundational inputs upon which innovative computational techniques are built. To utilize sequence characteristics of the protein, Alley et al. reported an unsupervised representation learning model called UniRep (Alley et al., 2019) that uses recurrent neural networks (RNNs) to learn statistical representations from a modified UniRef50 dataset comprising approximately 24 million protein sequences. It then

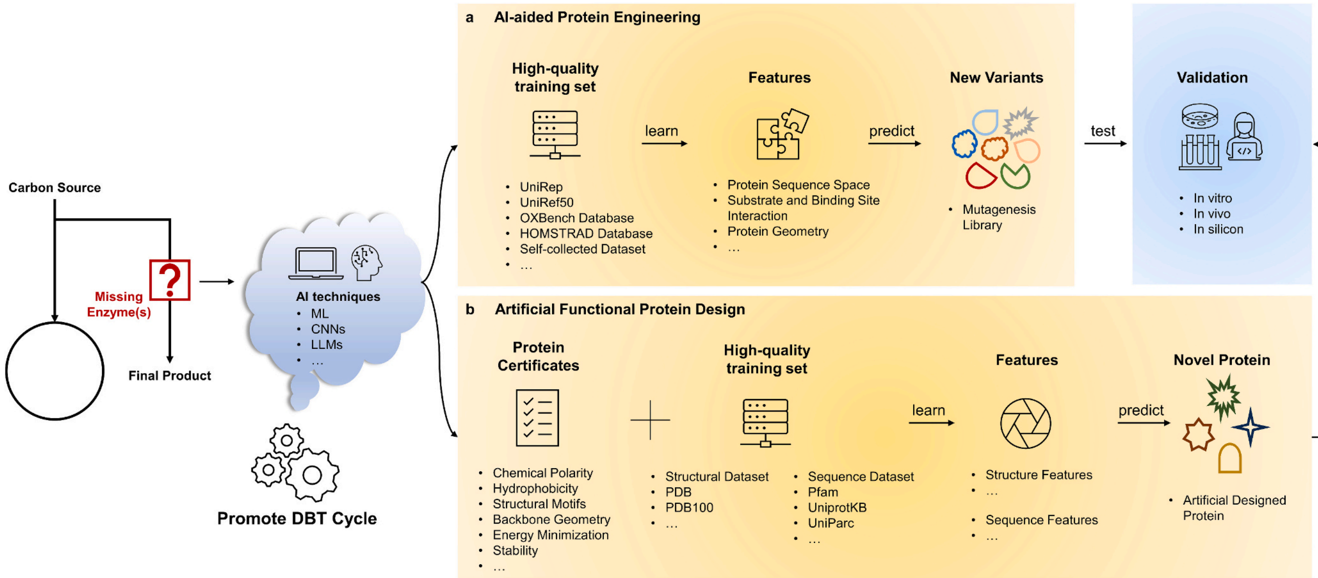

Fig. 2. AI-aided protein engineering and design. (a) AI guides protein engineering by learning protein characteristics from high-quality training data and assists in generating mutagenesis libraries to predict and screen for new functional variants. The selected variants are further validated through in vivo, in vitro, or in silicon experiments. (b) Designing artificial functional proteins requires considering multiple protein properties to create novel proteins. AI models are trained on quality protein sequence or structure datasets to learn patterns and generate proteins with new sequences or structures that may have useful functions. The generated novel proteins are then tested in vivo or in vitro to verify their activities.

Table 2
Selected examples of AI-aided protein engineering applications.

Year	Protein	Target property	Library Size	Learning Approach	Model and method	Screening Results	Performance	Reference
2019	Nitric oxide dioxygenase (NOD)	Improved enantioselectivity	1000 variants in each predicted library	Unsupervised	linear, kernel, neural network, and ensemble methods	90 random variants were tested in vitro	The NOD variant (32 V, 46C, 56H, 97 V, 49Y, 51 V, and 53F) showed selective catalysis of 93% to (*S*)-enantiomers. The NOD variant (32G, 46S, 56S, 97G, 49P, 51R, and 53 L) showed selective catalysis of 79% enantiomeric excess to (*R*)-enantiomers.	(Wu et al., 2019)
2020	Nitrilase	Predict the substrate scope of nitrilases	Over 600 samples for nitrilase screening; 240 nitrilases-substrate activity data.	Supervised	Logistic regression, random forest, gradient-boosted decision trees, and support vector machines	Identified a set of 137 nitrilases and tested 12 putative nitrilases for substrate scopes.	The results of cross-validation indicated that the four machine learning models examined displayed comparable predictive capabilities concerning substrate scope.	(Mou et al., 2021)
2020	Blue fluorescent protein (secBFP2.1), *Candida albicans* phosphomannose isomerase (CaPMI), and TEM-1 β-lactamase	Improved protein folding and function	580 secBFP2.1 variants, 200 TEM-1 variants, and 200 CaPMIvariants	Supervised	3D convolutional neural network	Mutations with improved functional readout (3 for secBFP2.1, 5 for TEM-1, and 9 for CaPMI)	The BFP variant displayed over 6 times higher fluorescence in *E. coli* compared to the wild type BFP. The combined TEM-1 β-lactamase variant (N52K, F60Y, M182T, E197D, and A249V) demonstrated around a 5-fold improvement in its activity. Similarly, the CaPMI variant with the S56A, G119A, Q157I, Q193D, D229T, C295V, L335E, K347R, S368N, K402R, and Q428T mutations showed approximately 5 times better performance for its associated phenotype compared to the wild type.	(Shroff et al., 2020)
2021	Acyl-ACP reductase (ACR)	Improved reduction activity on acyl-ACP substrates	4374 variants	Supervised	Gaussian process (GP) sequence-function models	Top 20 sequences with maximized Gaussian mutual information	ATR-83 produced about 5-fold more fatty alcohols (titer of 54 ± 11 mg/L) than MA-ACR.	(Greenhalgh et al., 2021)
2022	PET hydrolase	Improved hydrolytic activity	159 predicted mutations generated by MutCompute	Self-supervised	3D convolutional neural network (CNN)	29 probable mutations	FAST-PETase presented excellent PET depolymerization activities even for untreated plastics in the temperature range of 30 °C to 50 °C and the pH range of 6.5 to 8.0.	(Lu et al., 2022)
2022	CRISPR-associated protein-9 (KKH-SaCas9)	Improved SaCas9 activity	952 variants	Supervised	Neural Network Models	17 top-performing variants	The screened variant KKH-SaCas9-plus (N888R/A889Q) displayed the highest editing efficiency, with an average activity across 3 guide RNAs targeting GFP that was 127% of the KKH-SaCas9's activity.	(Thean et al., 2022)
2022	Cytochrome P450 (CYP105AS1)	Improved stereoselective hydroxylation of CYP105AS1	9052 virtual variants	Supervised	Rosetta CoupledMoves	8 potential variable mutations	The optimized CYP105AS1 enzyme variant catalyzed the stereospecific hydroxylation of compactin to pravastatin with over 99%	(Ashworth et al., 2022)

(*continued on next page*)

Table 2 (*continued*)

Year	Protein	Target property	Library Size	Learning Approach	Model and method	Screening Results	Performance	Reference
							selectivity, completely eliminating the formation of the byproduct.	
2023	Gaussia luciferase (GLuc)	Improved bioluminescence	164 Gluc variants	Self-supervised	EvoPlay (a self-play reinforcement learning framework)	29 EvoPlay-designed variants	The best luciferase variant (V12A, H62K, P67L, E85S, S86T, A87G, E93P, L107M, V121E) showed a 7.8-fold improved bioluminescence than the wild type.	(Wang et al., 2023b)
2023	Glycolyl-CoA Carboxylase (GCC M5)	Improved enzyme properties	10,019 variants	Supervised	The combination of a GP-based model and a Unirep-random forest model	105 predictions of which 7 of them were tested in vitro.	The GCC M5 G20R variant showed a 2-fold increased carboxylation rate and the GCC M5 L100N showed a 60% reduced ATP demand.	(Marchal et al., 2023)
2023	Renilla luciferase (RLuc) homologs	Improved enzyme activity and stability	219 redesigned active center variants and 394 redesigned scaffold variants	Supervised	The generative Maxent model	8 single variants for redesigned active center and 6 single variants for redesigned scaffold	RLuc (M185L) showed 2.01-fold higher relative activity compared to the wild type, and Rluc (I75A) showed increased stability, with its melting temperature increased by 6.0 °C relative to the wild type.	(Xie et al., 2023)

condenses protein sequences into fixed-length vectors to capture essential characteristics. UniRep has shown superior and broadly applicable performance in protein engineering tasks including stability assessments, function predictions, and protein design (Alley et al., 2019). Most reported machine learning models were built upon enormous training sets. However, the fine-tuning of a well-pre-trained model leads to surprising predictions for specific tasks as well, requiring only a small, high-quality dataset. Biswas et al. integrated the UniRep protein representation model with machine learning to build an ML-guided approach using a small quantity of functionally tested mutant sequences as training sets to facilitate the high-throughput screening of millions of protein sequences via in-silico directed evolution (Biswas et al., 2021). This study involves model fine-tuning, surrogate model building, mutant generation, protein activity optimization, and new protein evaluation. By further incorporating natural protein sequence data, the model learns to discern 'unnaturalness' and avoid nonfunctional sequences. Despite using just mini-sized 24 or 96 characterized variants as training sets, it effectively engineered two distinct proteins: avGFP originating from *Aequorea victoria* and TEM-1 β-lactamase from *E. coli* (Biswas et al., 2021). In these two cases, both engineered variants showed improved activities and low sequence identities compared to their wild types. This study demonstrates the promise of model fine-tuning and machine learning for protein engineering.

Protein structures are important clues for exploring the relationship between sequence, structure, and function. The intricate 3D protein structures contain valuable biological information for structural biology and protein engineering. AI has recently been applied to guide protein evolution through structural considerations. In protein engineering, one considerable feature is the atom interactions between binding pocket residues and ligands. Ao et al. combined structure-based and data-driven machine learning strategies to engineer novel amine transaminase (ATA) variants. They first generated a library of rationally designed ATA mutants and experimentally identified variants exhibiting improved activities. Additionally, protein structure-function relationships were described by one-hot code via analyzing the steric and electronic effects of key binding pocket residues. Collecting assay data as a training set, they developed an ML-based predictor tailored to ATA optimization. After computationally screening thousands of predicted variants, selected variants were validated experimentally, yielding an optimized ATA variant with 3-fold enhanced enzymatic activity in the transaminase-catalyzed reaction (Ao et al., 2023). Another important direction of AI-aided protein engineering is generating variants based on predicted protein structures. Since AlphaFold is able to precisely predict the protein 3D structure, it has been widely applied to the rational design of proteins with unknown crystal structures for improved properties. Wang et al. reported an AlphaFold-aided semi-rational Cas9 engineering strategy to elucidate the mechanism of SeCas9 and then developed the SeCas9 into a CRISPR-based gene regulator, ω-SedCas9-NQ, for microbial production. Guided by the predicted SeCas9 structure, the authors not only expanded the PAM specificity for the development of titratable CRISPRi, but also fused the *E. coli* RNAP ω subunit variant with SedCas9 for the construction of CRISPRa. The bifunctional ω-SedCas9-NQ regulator successfully tuned the 4-hydroxycoumarin production in *E. coli*, achieving 2.6-fold enhancement over the control group without any regulation (Wang et al., 2023a).

To further capitalize on the advantages of multi-modality of protein data, a wide range of sequence- and structure-based machine learning models have expedited the progress in digging out not only the sequence features of protein but also the deep topological properties for protein engineering. Qiu et al. presented the Topology-offered Protein Fitness (TopFit) machine learning model as a complementary approach to protein sequence and structure embeddings for analyzing the intricate geometric complexity of proteins. TopFit combines the persistent spectral theory (PST) to analyze complex protein crystal structures and deep protein language models to interpret the protein sequence features. This combination adeptly captures essential characteristics based on the structure-sequence relationships within the protein fitness landscape. TopFit's efficacy was extensively assessed on more than 30 benchmark datasets totaling over 120,000 variants. The results showed TopFit significantly outperformed purely sequence-based methods and indicated the robustness of TopFit in aiding protein engineering. However, its efficacy relies heavily on high-quality protein structure datasets, which can be limiting in broader contexts (Qiu and Wei, 2023).

Progress in both computational and experimental techniques collectively reshaped the realm of AI-aided protein engineering. It is foreseeable that these advances will accelerate protein engineering and expand enzyme availability, ultimately improving the production yield of current biosynthetic pathways and facilitating the construction of novel biosynthetic pathways.

4. Artificial functional protein design

Artificial functional protein design is a promising method to generate novel proteins to fulfill unsolved enzymatic reactions and build underexplored biosynthetic pathways for valuable compounds. Although genome mining and AI-aided protein engineering offer effective approaches to discovering novel proteins, covering all enzyme types in the short term remains a considerable challenge. In this context, the de novo design of artificial functional proteins emerges as an alternative strategy to address the lack of functional proteins, aiming to effectively fill gaps in biosynthetic pathways for microbial production. As the protein structure fundamentally determines the protein function, de novo protein design must consider a range of interconnected factors (Woolfson, 2021). Firstly, this intricate design involves strategically selecting proper amino acids and arranging them accurately to ensure stability and functionality. Then, to realize the desired fold, structural templates and motifs derived from existing protein structures guide the design and position of secondary structures. Next, proper backbone geometries and torsion angles, in conjunction with energy minimization using molecular mechanics force fields, contribute to the stability of protein conformations. Furthermore, the interplay between the designed protein and its reaction environment must be considered, as solvent effects significantly influence structural stability. Most importantly, protein functionality hinges on the precise positioning of active sites and key residues, as well as the interactions between substrate and active site. Therefore, computational methods aid and drive the design process, simulating the protein's inherent conformational flexibility. Finally, rigorous validation, using techniques like X-ray crystallography and NMR spectroscopy, further physically confirms the accuracy of designed proteins. Successful de novo protein design integrates these interconnected factors and considers available resources and ethical implications, to produce structurally robust and functional proteins.

The field of protein design encompasses a range of methods, including minimal protein design, rational protein design, consensus protein design, and computational protein design, aimed at creating novel protein structures and functions from scratch. Minimal protein design employs simple chemical principles to pattern polar and hydrophobic amino acids, leveraging the hydrophobic effect to guide the folding of secondary structures (Woolfson, 2021). The objective of this strategy is to create proteins with the smallest conceivable number of amino acid residues, while simultaneously maintaining specific structural and functional characteristics. In contrast, rational protein design leverages a profound understanding of structure-function relationships to make informed changes to protein sequences or structures (Korendovych and DeGrado, 2020). By integrating biochemical and evolutionary data, this approach fortifies the robustness of designs (Malbranke et al., 2023). While enhancing design accuracy, this method maintains the speed of design cycles akin to those observed in minimal design. On the other hand, consensus protein design draws inspiration from multiple naturally occurring homologous proteins to create novel sequences that combine favorable features for enhanced stability, functionality, or other desirable characteristics (Sternke et al., 2019). Unlike rational approaches, the consensus design may be confined to mimicking established natural target structures. Recently, computational protein design (CPD) has become a potent paradigm, harnessing protein features, high-quality training data, and computational algorithms and simulations to de novo design proteins with desired properties, functions, and structures (Fig. 2b). This includes energy function-based approaches (Liu and Chen, 2023), exemplified by RosettaDesign (Alford et al., 2017), which optimize sequences to yield thermodynamically favorable protein structures. Despite notable successes, energy function-based methods have limitations in design space exploration and success rates. On the other hand, CPD integrates data-driven approaches, such as structure-based sequence design, that leverage the capability of AI models to infer sequence distributions based on target backbone structures (Baek and Baker, 2022; Bennett et al., 2023; Dauparas et al., 2022; Yeh et al., 2023). These methods draw samples from the learned distributions to optimize sequences to fold into specific structures. Moreover, protein sequence design based on the language models exploits the similarity between protein sequences and natural language data, transferring representation-based models developed in the NLP fields into the protein design domain (Liu and Chen, 2023), witnessing the development of various language-based representation models such as ESMFold (Lin et al., 2023) and ProGen (Madani et al., 2023). Here, we review several CPD studies to demonstrate how computational models advance this field in detail.

With the help of computational methods, the de novo protein design process is simplified. Yeh et al. published de-novo-designed luciferase using a deep-learning-based approach. To generate a protein scaffold that can accommodate appropriate binding pockets for the substrate diphenylterazine (DTZ), the authors chose the structure of nuclear transport factor 2 (NTF2) as the target topology. The deep-learning model was first trained on a dataset with over 7000 NTF2-like protein structures and then retrained by using MSAs to filter the dataset. To idealize the structure generation, a new fold-specific loss function and trRosetta were implemented. For model finalization, they introduced hydrogen bonding networks to specify the backbone conformation and functionalize the binding pocket. In carrying out the de novo luciferase design, hundreds of hallucinated NTF2-like protein scaffolds were generated first, and then the substrate (DTZ) conformers were created for binding pocket design. By using RifGen, a Rotamer Interaction Field (RIF) was produced to stabilize anionic DTZ and enable hydrophobic packing interactions. Finally, RifDock was used to dock the RIF into the modeled NTF2 scaffolds, with further optimization by position-specific score matrix (PSSM)-biased sequence design. Validation experiments showed the computationally designed luciferases had high selectivity, thermostability and sufficient catalytic activity (Yeh et al., 2023). This protein design approach provided a solution to largely expand the de-novo-designed protein scaffold pool, shedding light on the structure-based de novo functional protein design.

Different from structure-based computational de novo protein design, the involvement of LLMs in protein design has revolutionized this field by learning protein language. Especially, the advances in AI-Generated Content (AIGC) can accelerate the artificial protein design process by first learning protein language and then generating protein sequences with natural-like folded structures in seconds. Ferruz et al. developed an unsupervised transformer-based language model for protein design, called ProtGPT2 (Ferruz et al., 2022). Following the grammar logic of human language, i.e., “letter -> words -> sentence -> meaning”, the authors created a protein language akin to the corresponding logic, namely “amino acid -> secondary structure -> domains -> function”. The model was trained on the Uniref50 dataset that contained the entire protein space, including the dark proteome and homology modeling. Additionally, the model learned the relationship between amino acids and generated the next amino acid based on the previous sequence by picking up a high-probability one. To further improve the decoding capability of ProtGPT2, three sampling strategies, Greedy search, Beam search, and random sampling, for sequencing generating were applied, respectively. The results showed that random sampling dramatically improved the amino acid-generated propensities that almost matched the natural protein sequence. As a newly developed protein generator, ProtGPT2 was validated by protein structural properties exerted through protein structure predictor and molecular dynamic simulation. ProtGPT2 can generate natural-like globular protein structures. The structural comparison between the ProtGPT2 proteins and natural proteins indicated that no significant difference was observed overall, but in high-identity regions, they can tell the distance from the natural proteins. Owing to the unsupervised training, ProtGPT2 was able to touch the dark space of protein and surpass the boundaries of the current protein space, providing more potential in designing proteins with novel topologies and functions (Ferruz et al., 2022). ProtGPT2 is available online (https://huggingface.co/nferruz/ProtGPT2) and it can generate proteins in two ways. One way is the zero-shot generation that generates random protein sequences without conditional restriction. The other way is user-defined sequence generation, which needs fine-tuning first and then generates tailored sequences.

ProGene is another transformer-based conditional language model that enables the generation of protein sequences within specific protein families (Madani et al., 2023). Initially trained on a vast dataset spanning over 19,000 families with associated control tags containing protein properties information, the model subsequently underwent fine-tuning tailored to the targeted protein family. This process enhanced its predictive performance for the specific functional protein family. During sequence generation, conditional control tags, as the input, facilitated the generation of protein sequences with related functions within milliseconds. To validate the model, the authors trained it on a lysozyme dataset, yielding artificial lysozymes, notably L056, capable of both in vitro and in vivo expression. The artificial lysozymes folded into structures akin to natural proteins and exhibited normal enzymatic activities and functions, despite low sequence identity to known proteins. Furthermore, the model's application to other protein families, such as chorismate mutase and malate dehydrogenase, demonstrated accurate functional predictions (Madani et al., 2023). Overall, ProGene realized the generation of family-specific protein sequences, leading to the development of LLMs-driven functional protein design. The advancement of CPD methods is poised to propel the design of functional proteins for improved activity, specificity, stability, and other desired properties, to address the limitations imposed by enzyme constraints in microbial production.

5. AI-enabled pathway design

The biosynthetic pathway is the core of microbial production with effective enzymes serving as indispensable building blocks. In metabolic engineering and synthetic biology studies, substantial manpower and resources are typically required for pathway design to synthesize high-

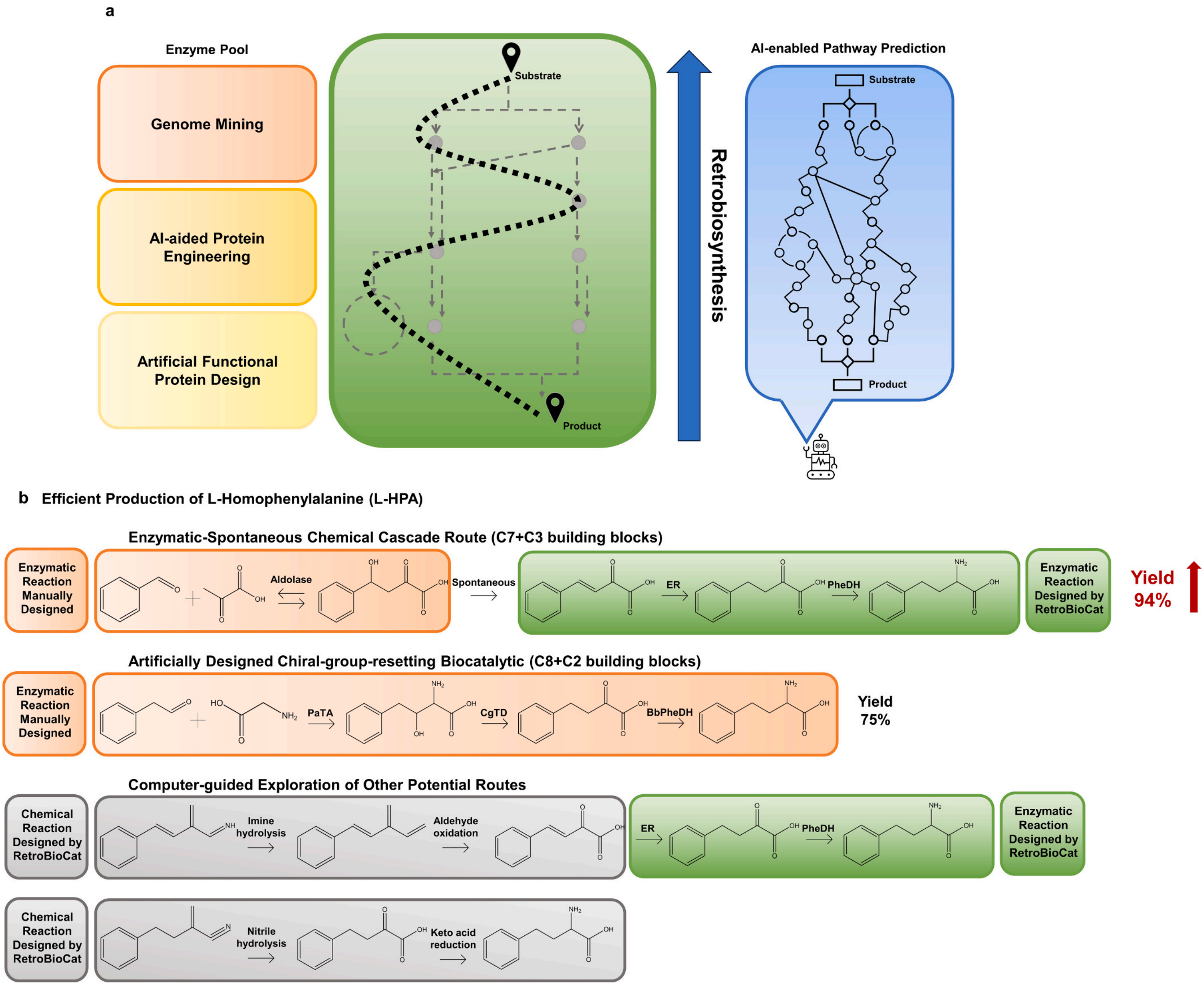
a
Enzyme Pool
Genome Mining
AI-aided Protein Engineering
Artificial Functional Protein Design
Substrate
Product
Retrobiosynthesis
AI-enabled Pathway Prediction
Substrate
Product
b Efficient Production of L-Homophenylalanine (L-HPA)
Enzymatic-Spontaneous Chemical Cascade Route (C7+C3 building blocks)
Enzymatic Reaction Manually Designed
Aldolase
Spontaneous
ER
PheDH
Enzymatic Reaction Designed by RetroBioCat
Yield 94%
Artificially Designed Chiral-group-resetting Biocatalytic (C8+C2 building blocks)
Enzymatic Reaction Manually Designed
PaTA
CgTD
BbPheDH
Yield 75%
Computer-guided Exploration of Other Potential Routes
Chemical Reaction Designed by RetroBioCat
Imine hydrolysis
Aldehyde oxidation
ER
PheDH
Enzymatic Reaction Designed by RetroBioCat
Chemical Reaction Designed by RetroBioCat
Nitrile hydrolysis
Keto acid reduction

Fig. 3. AI-aided pathway design (a) AI-enabled pathway design utilizes the available enzymes characterized through genome mining, engineered by AI-aided protein engineering, or designed by artificial functional protein design. The AI-aided retrobiosynthesis tools predict the pathway by breaking up the given product into the substrate and searching for suitable enzymatic/chemical reactions from reaction databases to propose the pathway. (b) An example of AI-enabled pathway design for efficient production of *L*-Homophenylalanine (L-HPA) by RetroBioCat. The best-predicted pathway is an enzymatic-spontaneous chemical cascade route using benzaldehyde (C7) and pyruvate (C3) as building blocks, reaching 94% yield higher than the previously reported pathway.

value compounds, especially those with complicated structures. Due to the limitations of enzymatic reaction diversity in host strains and special precursor demands for some products, introducing multiple heterologous reactions is often necessary to build complex novel pathways (Lin et al., 2013; Luo et al., 2019). Notably, exploring effective enzymes via genome mining (Refer to Section 2), AI-aided protein engineering (Refer to Section 3), or artificial functional protein design (Refer to Section 4) offers abundant enzyme resources to bridge the substrate to the product (Fig. 3a). To systematically construct optimal biosynthetic pathways, various factors warrant careful consideration, including the availability and promiscuity of heterologous enzymes, the requirement and supply of energy, the circulation of cofactors, and the selection of substrates (Boock et al., 2015). As more and more enzymes and enzymatic reactions are identified and cataloged in publicly available databases like MetaCyc (Caspi et al., 2019), KEGG (Kanehisa and Goto, 2000), and BRENDA (Schomburg et al., 2002), establishing complex biosynthetic pathways becomes increasingly feasible. However, manually screening reactions and enzymes from these voluminous resources to design functional pathways is labor-intensive and challenging. By leveraging computational algorithms and predictive models, biosynthetic pathway prediction facilitates efficient pathway design and construction for microbial production purposes (Fig. 3a), especially with various AI-enabled retrobiosynthesis tools springing up (Campodonico et al., 2014; Delépine et al., 2018; Finnigan et al., 2021). These tools translate the enzyme reactions from the dataset into computational language for training purposes to learn the reaction rules. With a given target compound, possible precursors are then inferred by searching learned reactions and pathways. The predicted pathway will be output when the search result meets the terminating conditions (Yu et al., 2023a).

One common type of retrobiosynthesis tool uses template-based prediction for biosynthetic pathway design by following the reaction rules that are manually curated by experts or automatically extracted by algorithms from reaction databases. For example, GEM-path utilizes an enzyme database generated by processing 3rd-level EC numbers in BRENDA, serving for calculating the reaction promiscuity between the native and non-native substrates. This approach also considers thermodynamics, production potential, and strain design (Campodonico et al., 2014). RetroPath2.0 is another open source retrobiosynthesis tool aiding novel biosynthetic route design. In this model, the enzymatic reaction rules are encoded by SMARTS strings (the text-based reaction representations) with enzymatic promiscuity determined by adjustable reaction centers to show reaction similarity. Numerous rules were extracted from the MetaNetX database, encompassing data from several mainstream metabolic databases like KEGG and MetaCyc (Delépine et al., 2018). Although the rules generated from the existing reactions stored in the database can cover many compounds, it is still a challenge to adjust the rules to adapt to various conditions. Specifically, too-precise rules yield conservative results, while too-general rules give unrealistic and unhelpful results (Yu et al., 2023a). RetroBioCat is a retrobiosynthesis planning tool designed for biocatalytic cascades. To build this tool, the developers created a set of expertly encoded rules describing with reaction SMARTS. Besides, a system that can automatically identify the literature precedent of enzymes was applied, which can provide users with professional suggestions about the applicability of the selected enzymes (Finnigan et al., 2021). Beyond pathway prediction, AI can also remove barriers and troubleshoot the de novo pathway design. Chemical Damage (CD)-MINE is a specialized toolset containing spontaneous reactions that occur under physiological conditions from literature and MetaCyc. By using the CD-MINE prediction tool, researchers can predict the possible spontaneous reactions in the branch of the synthetic pathways of our target metabolites, figuring out the potential problems and reducing the deleterious effects during pathway design (Jeffryes et al., 2022).

With continued optimization and upgrading, the above-mentioned AI-enabled retrobiosynthesis tools are becoming even more powerful, applicable, and user-friendly, making remarkable achievements in biosynthetic pathway design for microbial production. First, retrobiosynthesis pathway prediction can build high-yield biosynthesis pathways. 3-Phenylpropanol is a value-added compound for the food and cosmetics industries. Using RetroPath 2.0, Liu et al. designed a novel pathway for the biosynthesis of 3-phenylpropanol. According to the predicted routes, they screened and introduced the potential enzymes into the engineered *E. coli* strain and produced up to 847.97 mg/L 3-phenylpropanol, which is the highest titer compared with manually generated pathways (Liu et al., 2021). Secondly, retrobiosynthesis pathway prediction provides hints for novel enzyme discovery. *L*-Homophenylalanine (L-HPA) is a typical unnatural amino acid and a common building block for various drugs. However, its previous synthesis used expensive substrates like 2-oxo-4-phenyl-butanoic acid (OPBA). In a recent study, Gao et al. developed an enzymatic-spontaneous cascade for the synthesis of L-HPA with cost-effective substrates, benzaldehyde, and pyruvate (Fig. 3b). The cascade steps were designed by using RetroBioCat. Additionally, through genome mining with the probes suggested by RetroBioCat, an available ene-reductase (ER) that can accept (E)-2-oxo-4-phenylbut-3-enoic acid, called EcQOR, was first reported. It is also noteworthy that the rate-limiting enzyme, TipheDH, was modified according to the model generated by AlphaFold2, improving the specific activity by 82%. The final yield of the predicted pathway reached 94% (Gao et al., 2022). This study provides a paradigm of combining AI technology with microbial production. Moreover, AI-aided retrobiosynthesis tools can contribute to the development of sustainable biosynthesis by relieving multiple problems embedded in traditional chemical synthesis. For example, aliphatic α,ω-diamine (AD) production faces issues like energy waste, tedious steps, and toxic intermediates. Zhang et al. employed RetroBioCat to design biosynthetic routes, and the typical AD, 1,6-hexanediamine (HMD) was selected as the input. According to the prediction, the authors conducted a series of enzyme mining, built a microbial consortia system, and established a complete HMD synthetic pathway, achieving the highest HMD productivity in *E. coli* (Zhang et al., 2023b).

Although the rule-based tools have made impressive achievements in biosynthetic planning, some limitations remain to be addressed. For instance, the process of rules curation and heuristic extraction is time- and labor-consuming, and the manually or automatically generated rules may not be suitable for reactions beyond source databases. Recently, ML has been employed to drive retrobiosynthesis in a template-free way. Probst et al. developed an ML model for both forward and backward template-free prediction of enzyme-catalyzed reactions based on the highly accurate Molecular Transformer (Schwaller et al., 2019) trained in publicly available chemical reaction USPTO datasets. To enable retrobiosynthesis planning and improve accuracy, the authors introduced the enzymatic data set ECREACT, containing the enzymatic reactions with the corresponding EC numbers. Case studies showed successful retrobiosynthesis planning of aminoalcohol, homoaspartate, 4-hydroxy-L-glutamic acid, β-ketoacid and (*S*)-norlaudanosoline under mild enzymatic reaction conditions, although further experimental validation is still needed (Probst et al., 2022). In another study, Zheng et al. developed an ML-based toolkit, BioNavi-NP, for multi-step retrobiosynthesis planning of natural products (NPs) and NP-like molecules. The main component of the model is the single-step prediction composed of transformer neural networks trained by organic and biosynthetic reactions to predict candidate precursors. To improve accuracy, they curated a dataset from BioChem and augmented it with USPTO reactions. Although BioNavi-NP shows superiority in long pathway prediction, it still has a long way to go for the prediction of complex NPs that require many building blocks and reaction steps (Zheng et al., 2022).

6. Challenges and future perspectives

A primary concern of applying AI to microbial production is the availability and quality of publicly accessible biological data. Although recent studies show the potential of LLMs in performing AI-aided microbial production tasks like genome mining, enzyme function prediction, and protein design, to develop mature and robust AI models, expansive, accurate, and diverse datasets are necessary for training purposes. Unfortunately, relevant biological data is often limited in quantity and quality, inherently noisy, or subject to biases. Expanding open-access databases with standardized formats and descriptive metadata is thus essential, enforcing standardized formats and including descriptive metadata. Especially, the advent of LLMs underscores this, as larger models demand correspondingly vast datasets. For example, the recently released open-source Llama 2, scaling from 7B to 70B, was pretrained on 2 trillion tokens and fine-tuned on over 100,000 chat use cases (Touvron et al., 2023). Such data-intensive models risk compromised performance and catastrophic forgetting without extensive, diverse biological datasets. Furthermore, integrating multimodal data types, covering protein and DNA sequences, 3D molecular structures, functional assays, and more, will address data scarcity, paving the way for foundation models with a holistic understanding of biological systems. However, current LLMs primarily address language tasks, lacking capabilities for multimodal functions, especially visual tasks. An emerging avenue involves expanding LLMs to multimodal vision-language models capable of tackling visual challenges in microbial production such as microscopy image analysis, phenotypic characterization, and 3D structural visualization in protein design and engineering.

Another pervasive challenge is the generalizability of AI models beyond the datasets on which they are trained. Many models demonstrate robust performance on narrow or idiosyncratic training data, but they often fail to transfer effectively to new tasks or datasets. Advancing transfer learning approaches, multi-task learning frameworks, and rigorous benchmark testing on diverse data will bolster model generalizability. Additionally, the close integration of computational predictions with wet lab experiments is essential for maximizing practical impact. Leveraging computational methods to build high-throughput screening platforms and laboratory automation will enable tight iterative loops of prediction, validation, and model refinement.

Finally, the ethical and regulatory issues around data privacy, model transparency, and potential LLM biases necessitate community-approved regulatory guidelines and standards. To fully realize the potential of AI-aided microbial production, supporting software infrastructure and interdisciplinary collaboration are crucial. Cohesive pipelines from biological data curation to computational prediction to prospective experimental validation will close the loop on the AI-guided microbial production design cycle. Developing user-friendly, open-source tools for building, sharing, and deploying models will accelerate real-world impact.

7. Conclusion

In conclusion, AI-aided genome mining, protein engineering, protein design, and pathway design demonstrate remarkable and substantial progress and hold tremendous promise in transforming the traditional DBT mode of microbial production into the DBTLP workflow. Exciting opportunities lie ahead, especially as emerging computational approaches can translate biological information into computational readable content. However, further development will require overcoming critical challenges in data availability, model generalizability, experimental validation, and ethical practices. Yet fully unlocking the power of AI in aiding microbial production will depend on sustained cross-disciplinary collaborations between data scientists, biologists, chemists, and engineers across academia and industry.

Author contributions

X.G., T.L., and Y.Y. conceived and designed the review. X.G. drafted the manuscript and drew the figs. X.G. and T.L. wrote section 2 together, and R.D. revised this section. X.G. and Q.G. wrote section 3 together. X. G. and J.H. wrote section 4 together. J.Z. and X.G. wrote section 5 together. Z.L., Z.W., and X.G. wrote section 6 together. Z.L. wrote section 7. X.G. and Y.T. revised the manuscript. All authors reviewed and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support from the College of Engineering, The University of Georgia, Athens.

References

Abril, J.F., Castellano, S., 2019. Genome Annotation. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 195–209. https://doi.org/10.1016/B978-0-12-809633-8.20226-4.

Aleksander, S.A., et al., 2023. The gene ontology knowledgebase in 2023. Genetics 224 (1), iyad031. https://doi.org/10.1093/genetics/iyad031.

Alford, R.F., et al., 2017. The Rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theory Comput. 13 (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

Alipanahi, B., et al., 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. 33 (8), 831–838. https://doi.org/10.1038/nbt.3300.

Alley, E.C., et al., 2019. Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods 16 (12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

Altschul, S.F., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Amanatidis, D., et al., 2022. Deep neural network applications for bioinformatics. In: 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). IEEE, pp. 1–9. https://doi.org/10.1109/SEEDA-CECNSM57760.2022.9932895.

Ao, Y.F., et al., 2023. Structure-and data-driven protein engineering of transaminases for improving activity and stereoselectivity. Angew. Chem. Int. Ed. 62 (23) https://doi.org/10.1002/anie.202301660 e202301660.

Ardern, Z., et al., 2023. Elucidating the functional roles of prokaryotic proteins using big data and artificial intelligence. FEMS Microbiol. Rev. 47 (1), fuad003.

Ashworth, M.A., et al., 2022. Computation-aided engineering of cytochrome P450 for the production of pravastatin. ACS Catal. 12 (24), 15028–15044. https://doi.org/10.1021/acscatal.2c03974.

Baek, M., Baker, D., 2022. Deep learning and protein structure modeling. Nat. Methods 19 (1), 13–14. https://doi.org/10.1038/s41592-021-01360-8.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28 (1), 45–48. https://doi.org/10.1093/nar/28.1.45.

Bennett, N.R., et al., 2023. Improving de novo protein binder design with deep learning. Nat. Commun. 14 (1), 2625. https://doi.org/10.1038/s41467-023-38328-5.

Berman, H.M., et al., 2000. The Protein Data Bank. Nucleic Acids Res. 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

Biswas, S., et al., 2021. Low-N protein engineering with data-efficient deep learning. Nat. Methods 18 (4), 389–396. https://doi.org/10.1038/s41592-021-01100-y.

Boock, J.T., et al., 2015. Screening and modular design for metabolic pathway optimization. Curr. Opin. Biotechnol. 36, 189–198. https://doi.org/10.1016/j.copbio.2015.08.013.

Brandes, N., et al., 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38 (8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020.

Burgin, J., et al., 2022. The European nucleotide archive in 2022. Nucleic Acids Res. 51 (D1), D121–D125. https://doi.org/10.1093/nar/gkac1051.

Burley, S.K., et al., 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47 (D1), D464–D474. https://doi.org/10.1093/nar/gky1004.

Burley, S.K., et al., 2022. Protein data bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. Biomolecules 12 (10), 1425. https://doi.org/10.3390/biom12101425.

Cai, M., et al., 2023. Microbial production of L-methionine and its precursors using systems metabolic engineering. Biotechnol. Adv. 108260 https://doi.org/10.1016/j.biotechadv.2023.108260.

Campodonico, M.A., et al., 2014. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm. GEM-Path. Metab. Eng. 25, 140–158. https://doi.org/10.1016/j.ymben.2014.07.009.

Cao, Y., Shen, Y., 2021. TALE: transformer-based protein function annotation with joint sequence–label embedding. Bioinformatics 37 (18), 2825–2833. https://doi.org/10.1093/bioinformatics/btab198.

Casadevall, G., et al., 2023. AlphaFold2 and deep learning for elucidating enzyme conformational flexibility and its application for design. JACS Au. 3 (6), 1554–1562. https://doi.org/10.1021/jacsau.3c00188.

Caspi, R., et al., 2019. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 48 (D1), D445–D453. https://doi.org/10.1093/nar/gkz862.

Choi, K.R., et al., 2019. Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering. Trends Biotechnol. 37 (8), 817–837. https://doi.org/10.1016/j.tibtech.2019.01.003.

Clark, K., et al., 2020. Electra: pre-training text encoders as discriminators rather than generators. arXiv Prepr. arXiv: 2003.10555. doi: 10.48550/arXiv.2003.10555.

Dai, Z., et al., 2019. Transformer-xl: attentive language models beyond a fixed-length context. arXiv Prepr. arXiv: 1901.02860. doi: 10.48550/arXiv.1901.02860.

Dauparas, J., et al., 2022. Robust deep learning–based protein sequence design using ProteinMPNN. Science 378 (6615), 49–56. https://doi.org/10.1126/science.add2187.

Delépine, B., et al., 2018. RetroPath2.0: a retrosynthesis workflow for metabolic engineers. Metab. Eng. 45, 158–170. https://doi.org/10.1016/j.ymben.2017.12.002.

Devlin, J., et al., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv Prepr. https://doi.org/10.48550/arXiv.1810.04805 arXiv:.04805.

Ejigu, G.F., Jung, J., 2020. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology 9 (9), 295.

Elnaggar, A., et al., 2021. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

Ferruz, N., et al., 2022. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. 13 (1), 4348. https://doi.org/10.1038/s41467-022-32007-7.

Ferruz, N., et al., 2023. From sequence to function through structure: deep learning for protein design. Comput. Struct. Biotechnol. J. 21, 238–250. https://doi.org/10.1016/j.csbj.2022.11.014.

Finnigan, W., et al., 2021. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. Nat. Catal. 4 (2), 98–104. https://doi.org/10.1038/s41929-020-00556-z.

Fontana, J., et al., 2020. Challenges and opportunities with CRISPR activation in bacteria for data-driven metabolic engineering. Curr. Opin. Biotechnol. 64, 190–198. https://doi.org/10.1016/j.copbio.2020.04.005.

Gao, D., et al., 2022. Efficient production of L-homophenylalanine by enzymatic-chemical cascade catalysis. Angew. Chem. Int. Ed. Eng. 61 (36) https://doi.org/10.1002/anie.202207077 e202207077.

Gligorijević, V., et al., 2021. Structure-based protein function prediction using graph convolutional networks. Nat. Commun. 12 (1), 3168. https://doi.org/10.1038/s41467-021-23303-9.

Gong, X., et al., 2022. Engineering of a TrpR-based biosensor for altered dynamic range and ligand preference. ACS Synth. Biol. 11 (6), 2175–2183. https://doi.org/10.1021/acssynbio.2c00134.

Gong, X., et al., 2023. Evaluating the potential of leading large language models in reasoning biology questions. arXiv Prepr. https://doi.org/10.48550/arXiv.2311.07582 arXiv:.07582.

Greenhalgh, J.C., et al., 2021. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. Nat. Commun. 12 (1), 5825. https://doi.org/10.1038/s41467-021-25831-w.

Ismi, D.P., Pulungan, R., 2022. Deep learning for protein secondary structure prediction: pre and post-AlphaFold. Comput. Struct. Biotechnol. J. https://doi.org/10.1016/j.csbj.2022.11.012.

Jang, W.D., et al., 2022. Applications of artificial intelligence to enzyme and pathway design for metabolic engineering. Curr. Opin. Biotechnol. 73, 101–107. https://doi.org/10.1016/j.copbio.2021.07.024.

Jeffryes, J.G., et al., 2022. Chemical-damage MINE: a database of curated and predicted spontaneous metabolic reactions. Metab. Eng. 69, 302–312. https://doi.org/10.1016/j.ymben.2021.11.009.

Jeske, L., et al., 2019. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 47 (D1), D542–D549. https://doi.org/10.1093/nar/gky1048.

Ji, Y., et al., 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 37 (15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.

Jiang, T., et al., 2022. Establishing an Autonomous Cascaded Artificial Dynamic (AutoCAD) regulation system for improved pathway performance. Metab. Eng. 74, 1–10. https://doi.org/10.1016/j.ymben.2022.08.009.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28 (1), 27–30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, M., et al., 2022. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 51 (D1), D587–D592. https://doi.org/10.1093/nar/gkac963.

Kim, G.B., et al., 2023. Metabolic engineering for sustainability and health. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2022.12.014.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv Prepr. https://doi.org/10.48550/arXiv.1609.02907 arXiv:.02907.

Koonin, E.V., et al., 2003. Principles and methods of sequence analysis. In: Sequence—Evolution—Function: Computational Approaches in Comparative Genomics, pp. 111–192. https://doi.org/10.1007/978-1-4757-3783-7_5.

Korendovych, I.V., DeGrado, W.F., 2020. De novo protein design, a retrospective. Q. Rev. Biophys. 53 https://doi.org/10.1017/S0033583519000131 e3.

Kouba, P., et al., 2023. Machine learning-guided protein engineering. ACS Catal. 13 (21), 13863–13895. https://doi.org/10.1021/acscatal.3c02743.

Lan, Z., et al., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv Prepr. arXiv: 1909.11942. doi: 10.48550/arXiv.1909.11942.

Latif, E., et al., 2023. Artificial general intelligence (AGI) for education. arXiv Prepr. https://doi.org/10.48550/arXiv.2304.12479 arXiv:.12479.

Le, N.Q.K., et al., 2021. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. Brief. Bioinform. 22 (5), bbab005. https://doi.org/10.1093/bib/bbab005.

LeCun, Y., Bengio, Y., 1995. Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory Neural Networks, Vol. 3361, p. 1995 (10).

Lee, P., et al., 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New Engl. J. Med. 388 (13), 1233–1239. https://doi.org/10.1056/NEJMc2305286.

Li, C., et al., 2020. Protein engineering for improving and diversifying natural product biosynthesis. Trends Biotechnol. 38 (7), 729–744. https://doi.org/10.1016/j.tibtech.2019.12.008.

Li, R., et al., 2022. Machine learning meets omics: applications and perspectives. Brief. Bioinform. 23 (1), bbab460. https://doi.org/10.1093/bib/bbab460.

Lin, Y., et al., 2013. Microbial biosynthesis of the anticoagulant precursor 4-hydroxycoumarin. Nat. Commun. 4 (1), 2603. https://doi.org/10.1038/ncomms3603.

Lin, Z., et al., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379 (6637), 1123–1130. https://doi.org/10.1126/science.ade2574.

Liu, H., Chen, Q., 2023. Computational protein design with data-driven approaches: recent developments and perspectives. Wiley Interdiscip. Rev. Comput. Mol. Sci. 13 (3) https://doi.org/10.1002/wcms.1646 e1646.

Liu, Z., et al., 2021. Metabolic engineering of *Escherichia coli* for de novo production of 3-phenylpropanol via retrobiosynthesis approach. Microb. Cell Factories 20 (1), 121. https://doi.org/10.1186/s12934-021-01615-1.

Lovelock, S.L., et al., 2022. The road to fully programmable protein catalysis. Nature 606 (7912), 49–58. https://doi.org/10.1038/s41586-022-04456-z.

Lu, H., et al., 2022. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature 604 (7907), 662–667. https://doi.org/10.1038/s41586-022-04599-z.

Luo, X., et al., 2019. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. Nature 567 (7746), 123–126. https://doi.org/10.1038/s41586-019-0978-9.

Madani, A., et al., 2023. Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. 41 (8), 1099–1106. https://doi.org/10.1038/s41587-022-01618-2.

Malbranke, C., et al., 2023. Machine learning for evolutionary-based and physics-inspired protein design: current and future synergies. Curr. Opin. Struct. Biol. 80, 102571 https://doi.org/10.1016/j.sbi.2023.102571.

Marchal, D.G., et al., 2023. Machine learning-supported enzyme engineering toward improved CO2-fixation of Glycolyl-CoA carboxylase. ACS Synth. Biol. https://doi.org/10.1021/acssynbio.3c00403.

McDonald, A.G., Tipton, K.F., 2023. Enzyme nomenclature and classification: the state of the art. FEBS J. 290 (9), 2214–2231. https://doi.org/10.1111/febs.16274.

Mitchell, A.L., et al., 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47 (D1), D351–D360. https://doi.org/10.1093/nar/gky1100.

Mitchell, A.L., et al., 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48 (D1), D570–D578. https://doi.org/10.1093/nar/gkz1035.

Mou, Z., et al., 2021. Machine learning-based prediction of enzyme substrate scope: application to bacterial nitrilases. Proteins. 89 (3), 336–347. https://doi.org/10.1002/prot.26019.

Mullowney, M.W., et al., 2023. Artificial intelligence for natural product drug discovery. Nat. Rev. Drug Discov. 22 (11), 895–916. https://doi.org/10.1038/s41573-023-00774-7.

Ofer, D., et al., 2021. The language of proteins: NLP, machine learning & protein sequences. Comput. Struct. Biotechnol. J. 19, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022.

OpenAI, 2023. GPT-4 Technical Report. ArXiv abs/2303.08774. https://doi.org/10.48550/arXiv.2303.08774.

Probst, D., et al., 2022. Biocatalysed synthesis planning using data-driven learning. Nat. Commun. 13 (1), 964. https://doi.org/10.1038/s41467-022-28536-w.

Qiu, Y., Wei, G.-W., 2023. Persistent spectral theory-guided protein engineering. Nat. Comput. Sci. 3 (2), 149–163. https://doi.org/10.1038/s43588-022-00394-y.

Raffel, C., et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67.

Reeves, G.A., et al., 2009. Genome and proteome annotation: organization, interpretation and integration. J. R. Soc. Interface 6 (31), 129–147. https://doi.org/10.1098/rsif.2008.0341.

Rezayi, S., et al., 2023. Exploring new frontiers in agricultural NLP: investigating the potential of large language models for food applications. arXiv Prepr. https://doi.org/10.48550/arXiv.2306.11892 arXiv:.11892.

Ruff, K.M., Pappu, R.V., 2021. AlphaFold and implications for intrinsically disordered proteins. J. Mol. Biol. 433 (20), 167208 https://doi.org/10.1016/j.jmb.2021.167208.

Ryu, J.Y., et al., 2019. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. U. S. A. 116 (28), 13996–14001. https://doi.org/10.1073/pnas.182190511.

Sayers, E.W., et al., 2019. GenBank. Nucleic Acids Res. 48 (D1), D84–D86. https://doi.org/10.1093/nar/gkz956.

Schnoes, A.M., et al., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comp. Biol. 5 (12) e1000605.

Schomburg, I., et al., 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30 (1), 47–49. https://doi.org/10.1093/nar/30.1.47.

Schwaller, P., et al., 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent. Sci. 5 (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

Shroff, R., et al., 2020. Discovery of novel gain-of-function mutations guided by structure-based deep learning. ACS Synth. Biol. 9 (11), 2927–2935. https://doi.org/10.1021/acssynbio.0c00345.

Singh, A., et al., 2016. Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. 21 (2), 110–124.

Son, J., et al., 2023. Recent advances in microbial production of diamines, aminocarboxylic acids, and diacids as potential platform chemicals and bio-based polyamides monomers. Biotechnol. Adv. 62, 108070 https://doi.org/10.1016/j.biotechadv.2022.108070.

Steinegger, M., Söding, J., 2018. Clustering huge protein sequence sets in linear time. Nat. Commun. 9 (1), 2542. https://doi.org/10.1038/s41467-018-04964-5.

Steinegger, M., et al., 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods 16 (7), 603–606. https://doi.org/10.1038/s41592-019-0437-4.

Sternke, M., et al., 2019. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc. Natl. Acad. Sci. U. S. A. 116 (23), 11275–11284. https://doi.org/10.1073/pnas.1816707116.

Suzek, B.E., et al., 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31 (6), 926–932. https://doi.org/10.1093/bioinformatics/btu739.

Talebi, S., et al., 2023. Beyond the hype: assessing the performance, trustworthiness, and clinical suitability of GPT3. 5. arXiv Prepr. https://doi.org/10.48550/arXiv.2306.15887 arXiv:.15887.

Tan, Z., et al., 2023. Designing artificial pathways for improving chemical production. Biotechnol. Adv. 108119 https://doi.org/10.1016/j.biotechadv.2023.108119.

Tanizawa, Y., et al., 2022. DNA data Bank of Japan (DDBJ) update report 2022. Nucleic Acids Res. 51 (D1), D101–D105. https://doi.org/10.1093/nar/gkac1083.

Teng, Y., et al., 2022. Biosensor-enabled pathway optimization in metabolic engineering. Curr. Opin. Biotechnol. 75, 102696 https://doi.org/10.1016/j.copbio.2022.102696.

Teng, Y., et al., 2023. The expanded CRISPR toolbox for constructing microbial cell factories. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2023.06.012.

Thean, D.G., et al., 2022. Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. Nat. Commun. 13 (1), 2219. https://doi.org/10.1038/s41467-022-29874-5.

Touvron, H., et al., 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv abs/2307.09288. https://doi.org/10.48550/arXiv.2307.09288.

UniProtConsortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47 (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.

UniProtConsortium, 2022. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51 (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

Vaswani, A., et al., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.

von Heijne, G., 1991. Computer analysis of DNA and protein sequences. Eur. J. Biochem. 199 (2), 253–256. https://doi.org/10.1111/j.1432-1033.1991.tb16117.x.

Wang, T., Simmel, F.C., 2022. Riboswitch-inspired toehold riboregulators for gene regulation in *Escherichia coli*. Nucleic Acids Res. 50 (8), 4784–4798. https://doi.org/10.1093/nar/gkac275.

Wang, M., et al., 2018. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. 46 (11) https://doi.org/10.1093/nar/gky215 e69.

Wang, J., et al., 2021. Tunable hybrid carbon metabolism coordination for the carbon-efficient biosynthesis of 1, 3-butanediol in *Escherichia coli*. Green Chem. 23 (21), 8694–8706. https://doi.org/10.1039/D1GC02867G.

Wang, J., et al., 2023a. Exploring and engineering PAM-diverse Streptococci Cas9 for PAM-directed bifunctional and titratable gene control in bacteria. Metab. Eng. 75, 68–77. https://doi.org/10.1016/j.ymben.2022.10.005.

Wang, Y., et al., 2023b. Self-play reinforcement learning guides protein engineering. Nat. Mach. Intell. 5 (8), 845–860. https://doi.org/10.1038/s42256-023-00691-9.

Woolfson, D.N., 2021. A brief history of de novo protein design: minimal, rational, and computational. J. Mol. Biol. 433 (20), 167160 https://doi.org/10.1016/j.jmb.2021.167160.

Wu, Z., et al., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc. Natl. Acad. Sci. U. S. A. 116 (18), 8852–8858. https://doi.org/10.1073/pnas.1901979116.

Xie, W.J., et al., 2023. Enhancing luciferase activity and stability through generative modeling of natural enzyme sequences. Proc. Natl. Acad. Sci. U. S. A. 120 (48) https://doi.org/10.1073/pnas.2312848120 e2312848120.

Yang, K.K., et al., 2019a. Machine-learning-guided directed evolution for protein engineering. Nat. Methods 16 (8), 687–694. https://doi.org/10.1038/s41592-019-0496-6.

Yang, Z., et al., 2019b. Xlnet: generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Proces. Syst. 32.

Yeh, A.H.-W., et al., 2023. De novo design of luciferases using deep learning. Nature 614 (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.

Yu, T., et al., 2023a. Machine learning-enabled retrobiosynthesis of molecules. Nat. Catal. 6 (2), 137–151. https://doi.org/10.1038/s41929-022-00909-w.

Yu, T., et al., 2023b. Enzyme function prediction using contrastive learning. Science 379 (6639), 1358–1363. https://doi.org/10.1126/science.adf2465.

Zhang, R., et al., 2021a. Microbial utilization of lignin-derived aromatics via a synthetic catechol meta-cleavage pathway. Green Chem. 23 (20), 8238–8250. https://doi.org/10.1039/D1GC02347K.

Zhang, R., et al., 2021b. Development of antisense RNA-mediated quantifiable inhibition for metabolic regulation. Metab. Eng. Commun. 12 https://doi.org/10.1016/j.mec.2021.e00168 e00168.

Zhang, S., et al., 2023a. Applications of transformer-based language models in bioinformatics: a survey. Bioinform. Adv. 3 (1) https://doi.org/10.1093/bioadv/vbad001 vbad001.

Zhang, Z., et al., 2023b. Transforming inert cycloalkanes into α,ω-diamines by designed enzymatic cascade catalysis. Angew. Chem. Int. Ed. 62 (16) https://doi.org/10.1002/anie.202215935 e202215935.

Zheng, S., et al., 2022. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nat. Commun. 13 (1), 3342. https://doi.org/10.1038/s41467-022-30970-9.

Zhou, Z., et al., 2023. DNABERT-2: efficient foundation model and benchmark for multi-species genome. ArXiv. https://doi.org/10.48550/arXiv.2306.15006 abs/2306.15006.