IECE
Journal of Artificial Intelligence in Bioinformatics
http://dx.doi.org/10.62762/JAIB.2025.194547

REVIEW ARTICLE

Advances in Intelligent Design and Optimization Methods for Nucleic Acid Sequences

Ting Yang[1], **Minxu Han**[1], **Xiaoru Wen**[1] **and Yanfen Zheng**[2,*]

[1] Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian 116622, China
[2] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Abstract

As a carrier of genetic information, the precise design and optimization of nucleic acid (DNA or RNA) sequences are of critical importance for the realization of specific biological functions. From synthesizing genes to designing novel nucleic acid drugs, from constructing efficient expression vectors to modifying microbial metabolic pathways, all are inseparable from the fine regulation of nucleic acid sequences. The primary purpose of nucleic acid sequence design is to generate new sequences tailored to specific requirements for gene expression, function prediction, drug development, and other applications. In this paper, we first review the theoretical basis of nucleic acid sequence design, followed by an overview of current research methods for nucleic acid sequence design. Traditional nucleic acid sequence design methods rely on manual experience and experimentation, and although some progress has been made in the past decades, they still suffer from high cost, long time, and low efficiency in most cases. Therefore, optimizing nucleic acid sequences to improve their performance and stability has become particularly important. In recent years, artificial intelligence technology has provided a new direction for the design and optimization of nucleic acid sequences, opening up new possibilities for more efficient and accurate design methods. These methods include traditional rule-based nucleic acid sequence optimization approaches as well as AI-driven optimization methods for nucleic acid sequence generation. This review systematically examines the latest advancements in both traditional and AI-driven nucleic acid sequence design methods and analyzes the technical details, strengths, and limitations of each application. Finally, the article discusses the current challenges and future development directions of nucleic acid sequence design.

Keywords: nucleic acid sequence design, machine learning, optimization methods, generative models, heuristic algorithms, large language models, nucleic acid structure prediction.

Academic Editor:
Abdur Rasool

Submitted: 03 January 2025
Accepted: 27 March 2025
Published: 30 April 2025

Vol. 1, **No.** 1, 2025.
10.62762/JAIB.2025.194547

***Corresponding author:**
✉ Yanfen Zheng
zhengyanfen95@gmail.com

Citation
Yang, T., Han, M., Wen, X., & Zheng, Y. (2025). Advances in Intelligent Design and Optimization Methods for Nucleic Acid Sequences. *Journal of Artificial Intelligence in Bioinformatics*, 1(1), 12–29.

1 Introduction

In the field of contemporary life sciences, the intelligent design and optimization of nucleic acid sequences have become a hot issue in research, among which the design and optimization of nucleic acid sequence design are particularly critical and at the core, which is not only one of the core issues in the fields of synthetic biology, gene editing, and bio-informatics but also a key link in the development of life science and technology. The fundamental goal is to design and generate sequences that meet the multiple requirements of function, stability, and synthesizability according to the specific application needs, to promote the progress of genome engineering, drug development, and precision medicine [1]. In recent years, nucleic acid sequence design technology has made significant progress, especially in tool development, algorithm optimization, and application expansion [2–4]. Through the introduction of artificial intelligence algorithms and automated design platforms, nucleic acid sequence design has achieved efficient and intelligent closed-loop optimization, significantly improving the efficiency of sequence-function prediction and optimization [5].

Nucleic acid sequence design is widely used in synthetic biology [6], precision medicine [7], agriculture and food [8], industrial production [9] and information storage [10], etc. In synthetic biology, nucleic acid sequence design not only optimizes the efficiency of gene synthesis, but also accelerates the development of synthetic gene circuits and biological components, and promotes the development of fields such as biomanufacturing and environmental protection [11–13]. Meanwhile, the introduction of deep learning and artificial intelligence has brought revolutionary breakthroughs in nucleic acid sequence design, improving the accuracy and efficiency of gene expression efficiency, stability prediction, and sequence optimization [14–16]. With the development of high-throughput screening technology, experimental data-driven models have led to significant improvements in the accuracy and speed of functional verification of nucleic acid sequence design [17]. The rapid development in the field of information storage has also opened up new application prospects for nucleic acid sequence design [18].

With continuous technological innovations, nucleic acid sequence design will play an increasingly important role in the future and drive significant changes in related fields. This review explores intelligent design and optimization methods for nucleic acid sequences, focusing on existing technologies and methods, as well as future challenges and development directions. The paper first reviews the basic theories of nucleic acid sequence design, discussing the structural features, functions, and importance of DNA sequences in gene expression, which lays the foundation for the subsequent chapters. It then introduces traditional rule-driven DNA sequence optimization methods, focusing on strategies based on biological laws and chemical constraints, while also addressing their limitations in practical applications. Next, the paper explores intelligence-driven optimization methods, including generative models such as Variable Auto-Encoders, Generative Adversarial Networks, and Diffusion models, as well as optimization techniques based on large language models. These AI-driven approaches offer greater flexibility and accuracy for handling complex design tasks, overcoming some of the limitations inherent in traditional methods. The paper then discusses the practical significance of nucleic acid sequence design, along with the challenges involved in structure prediction. Finally, it summarizes the current challenges in nucleic acid sequence design and presents potential directions for future development, highlighting how intelligent algorithm-based approaches can drive further innovations in biology and medicine.

2 Basic theories of DNA sequence design

DNA sequence design and optimization aims to accurately construct DNA sequences according to specific functional requirements. This process involves the basic theories of molecular biology, computational biology, optimization algorithms, and artificial intelligence. Therefore, understanding these basic theories is crucial for effective DNA sequence design and optimization [19, 20]. DNA, as a carrier of genetic information, consists of four nucleotides—adenine (A), thymine (T), cytosine (C), and guanine (G)—which are arranged in a specific order to form a double helix structure. The nucleotide order determines the transmission of genetic information and the expression of genes [21, 22]. In addition to encoding proteins, DNA also plays a crucial role in various biological functions, such as regulating gene expression and protein synthesis [23, 24]. Therefore, when designing DNA sequences, several basic principles need to be followed. Firstly, to ensure the realization of gene function, the designed sequence should meet specific biological

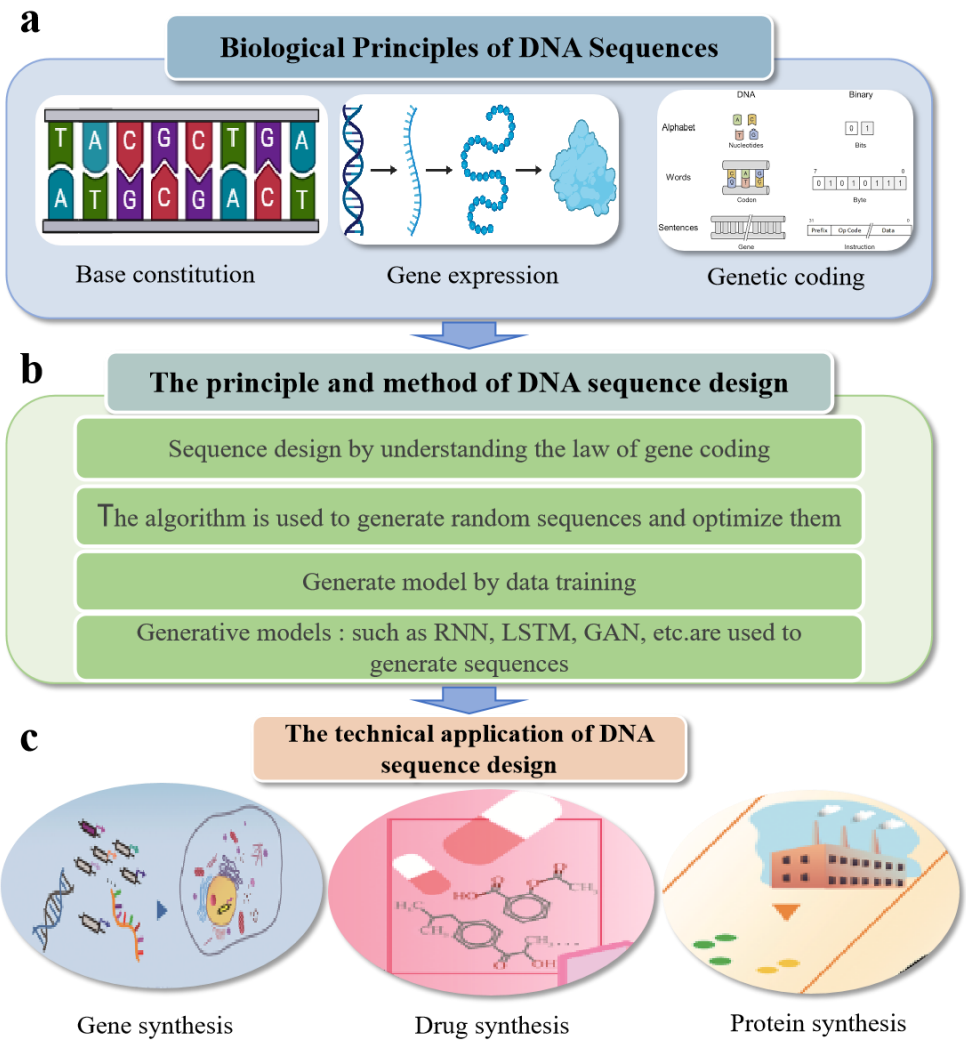

Figure 1. Theoretical foundations to the application of sequences design. (**a**) The biological principles of DNA sequences (DNA structure, gene expression, genetic coding, etc.). (**b**) Principles and methods of DNA sequence design, including rule base, random generation, machine learning, deep learning, and other methods. (**c**) The technical applications of DNA sequence design, including gene synthesis, drug synthesis, and protein synthesis fields.

needs, such as encoding specific proteins or realizing gene regulatory functions [25, 26]. Secondly, high expression efficiency must be ensured, so that the expression level of the target genes meets experimental requirements [27]. The expression level of DNA is influenced by various factors. These factors include promoter strength, codon usage bias, RNA secondary structures, mRNA stability, and compatibility with the host cell's transcriptional machinery. Therefore, the designed sequence must not only meet functional requirements but also optimize these factors as much as possible to improve the efficiency and stability of gene expression [28]. In addition, the feasibility of the synthesis process (e.g., avoiding repetitive sequences or high GC regions) and compatibility with the host cell's genetic background (e.g., codon usage bias, presence of restriction enzyme sites, and promoter or terminator compatibility) must also be considered during sequence design [29].

In the process of DNA sequence design and optimization, multiple optimization theories and methods are usually applied. The optimization goal is often not a single gene expression level but involves the balance and optimization of multiple goals, such as sequence stability, synthesis cost, transcription, translation efficiency, etc.[30]. At this time, multi-objective optimization methods are particularly important, which can find the best balance between multiple conflicting goals. For example, codon optimization can enhance both translation efficiency and mRNA stability, while adjusting sequence complexity (e.g., reducing

homopolymeric runs) may lower synthesis difficulty without compromising functionality [31, 32]. Combinatorial optimization methods, such as genetic algorithms and Monte Carlo simulations, are often used to achieve multi-objective optimization. These methods iteratively evaluate sequence variants based on global parameters (e.g., GC content, thermodynamic stability) and dynamically balance conflicting objectives [33]. Meanwhile, DNA sequence design is often accompanied by constraints such as sequence length, secondary structure, and gene regulatory element limitations [34], which increase the complexity of the design problem, and thus constrained optimization methods have emerged, which help to find the optimal or near-optimal solution under the premise of satisfying biological or engineering constraints. Overall, DNA sequence design is a complex, multi-faceted problem requiring integrated optimization approaches under biological constraints. As illustrated in Figure 1 [35], the theoretical foundation, design process, and application of DNA sequences are presented.

3 Traditional rule-driven DNA sequence optimization methods

Traditional rule-driven DNA sequence optimization methods rely on biological principles and empirical laws to optimize gene expression, stability, and synthesizability by adjusting key elements (e.g. codons, GC content, secondary structure, etc.) in the gene sequences [36–38]. The core idea of these methods is to avoid some negative factors affecting gene expression or synthesis efficiency through the fine design of DNA sequences without relying on complex machine learning or artificial intelligence algorithms.

Codon optimization is a common strategy to enhance gene expression levels [39]. Codon preferences for the same amino acid vary across organisms, so by replacing codons in the target gene that do not match the host cell's tRNA usage preferences, translation efficiency can be significantly improved, thus enhancing the gene expression level [38, 40]. While codon optimization improves expression, excessive replacement of rare codons (over-optimization) may disrupt, over-optimization (e.g., eliminating all rare codons) may inadvertently disrupt co-translational protein folding or mRNA stability [41]. Secondly, rational optimization of GC content is also an important aspect of DNA sequence design. Excessively high or low GC content can lead to difficulties in DNA synthesis or sequence instability, so it is a commonly adopted optimization method by adjusting the GC ratio to avoid regions of extreme GC content [38, 42]. Recent studies suggest combining GC optimization with thermodynamic stability calculations (e.g., using mfold) to balance synthesis feasibility and mRNA function [43]. In addition, repetitive sequences and low-complexity regions in DNA sequences (e.g., long homopolymers of T or A and regions rich in single bases) can affect the effectiveness of PCR amplification and may lead to instability of the expression system [44–46]. For example, the elimination of poly-A/T tracts longer than six nucleotides can reduce polymerase slippage errors during the synthesis process [47]. Therefore, the design of these regions needs to be avoided during optimization to ensure gene stability. The design of promoters and regulatory elements is also crucial [48]. By rationally selecting the strength of promoters and optimizing the location of ribosome binding sites (RBS), transcription and translation efficiency can be effectively enhanced, thereby increasing the level of gene expression [49, 50]. However, promoter optimization must align with host RNA polymerase specificity (e.g., T7 promoters in E. coli) to avoid transcriptional incompatibility [51]. In addition, secondary structures of DNA sequences (e.g., hairpin structures, pseudoknots, stem-and-loop structures, etc.) can affect the translation process of genes and even lead to the stagnation of gene expression [52, 53]. Therefore, avoiding or reducing the formation of these unfavorable secondary structures is another key strategy to optimize gene expression [54]. For instance, stable secondary structures (ΔG <-5 kcal/mol) are minimized using tools like NUPACK, while weak hairpins in 5′ UTRs may be retained to enhance mRNA stability [55, 56].

Traditional rule-driven DNA sequence optimization methods include a variety of strategies, each of which functions according to different optimization goals and application scenarios. The template variation method optimizes gene expression and stability by mutating existing template sequences, but it relies on the template itself, which may lead to the design of sequences that lack innovation in structure or function and cannot avoid the inherent bias of the templates [57, 58]; The random sampling method, on the other hand, generates and screens sequences on a large scale to find an efficient design, and although it is capable of exploring a wider range of possibilities, the search is less efficient due to combinatorial explosion (e.g., the number of possible variants for a 1 kb

gene may exceed 10^{20}) [59]. In contrast, AI-driven methods differ fundamentally in three key aspects: (1) Design Paradigm: AI models (e.g., Transformers) learn hidden rules from large-scale sequence-activity datasets, enabling multi-objective optimization (e.g., balancing translation efficiency and mRNA stability) [60]; (2) Innovation Potential: AI generates non-natural sequences (e.g., synthetic promoters with less than 40% sequence similarity to natural sequences) that bypass template limitations [61]; (3) Efficiency: Compared to traditional random sampling, AI-driven virtual screening can reduce the number of candidates for experimental validation by over 90% [62]. For example, the Linear Design algorithm optimizes mRNA secondary structures 10,000 times faster than manual methods while maintaining codon adaptation [63]. These advancements highlight the complementary roles of traditional and AI-driven approaches in advancing synthetic biology.

4 Intelligence-driven optimization of DNA sequence generation

4.1 Generative Models

Generative models can generate new sequences similar to real sequences by learning the statistical properties of DNA sequences [64, 65]. In addition to data generation, generative models can also enable dimensionality reduction by mapping the data space to the latent space, as well as predictive tasks by utilizing this learned mapping or supervised/semi-supervised generative model design [66, 67].

4.1.1 Variable Auto-Encoders

In the field of machine learning-driven DNA sequence design, generative models have gradually received widespread attention due to their potential to explore complex sequence spaces and discover novel design solutions. Variable Auto-Encoder (VAE) [68], a classical probabilistic generative model consisting of an encoder E and a decoder D, demonstrates unique advantages in sequence design tasks by its effective latent space characterization and flexible data generation capabilities. VAE combines the nonlinear representation capabilities of deep learning with the generative properties of probabilistic models, providing a powerful tool for biological sequence optimization and functional prediction [68, 69]. In DNA sequence design workflows, the VAE converts the input DNA sequences (usually represented by one-hot encoding or tokenized) into the mean and variance of a probability distribution (e.g., Gaussian distribution) in the latent space using an encoder [70]. The latent vector is then obtained by sampling from this distribution and passed to the decoder, which maps the latent vector back into the DNA sequences space to generate a new DNA sequence [71]. The training goal of the VAE is to minimize the reconstruction error and the KL divergence between the latent space distribution and a prior distribution, thereby learning a structured latent space capable of generating sequences similar to the original data [68]. Through this approach, VAE can generate DNA sequences that meet specific functional requirements, such as gene optimization and expression regulation. Figure 2(a) illustrates this application process.

The VAE demonstrates significant potential in biological sequence design, particularly excelling in functional sequence generation and stability optimization [71]. Its core strength lies in enabling controlled generation through probabilistic modeling of the latent space—researchers can efficiently learn latent representations of DNA sequences and sample the latent space to generate optimized sequences based on specific objectives (e.g., high stability) [69]. For instance, Sadeghi et al. [72] utilized VAE to design DNA-stabilized silver nanoclusters, enhancing sequence stability and functionality through automated feature extraction. Hawkins-Hooker et al. [71] generated functional protein variants, validating VAE's efficiency in protein design. Meanwhile, Greener et al. [73] and Moomtaheen et al. [74] applied VAE to metalloprotein design, novel protein fold exploration, and nanomaterial optimization, highlighting its versatility in biomolecular functional innovation. These studies demonstrate that VAE not only generates function-specific sequences but also provides systematic support for sequence stability, synthesizability, and cross-scale optimization.

4.1.2 Generative Adversarial Networks

Generative Adversarial Network (GAN) is a generative model composed of a generator (G) and a discriminator (D), first proposed by Goodfellow et al. [75] in their seminal work. In DNA sequences design, the generator (G) takes a noise vector z as input and generates a new sample G(z) as output [76]. In other words, the generator is responsible for mapping the potential space to the data space. The discriminator (D), on the other hand, takes as input a sample x and outputs a probability value D(x), which is used to evaluate whether x originates from a real data distribution or is synthesized by the generator G. The two networks are trained by

adversarial training. These two networks optimize each other using adversarial training, with the discriminator D aiming to maximize the probability of correct classification, while the generator G attempts to confuse the discriminator by minimizing its probability of misclassifying the generated sample G(z) Figure 2(b) [66]. Essentially, this adversarial mechanism forms a zero-sum game until an equilibrium is reached, where the discriminator D is unable to tell whether G(z) originates from the true distribution or not.

The application of GAN in DNA sequence design has made significant progress in recent years, especially in gene expression optimization, synthetic biology, and new molecule design [77]. By using GAN, researchers can generate DNA sequences that meet specific needs, regulate gene expression, optimize synthetic processes, and even create artificial genomes [78]. Moreover, GANs have an edge in generating sequences with high fidelity, functionality, and preservation of complex structures, especially in tasks that require integration with experimental validation, such as protein design or data augmentation [79, 80]. Yu et al. [81] proposed the MichiGAN model, which samples single-cell data using generative adversarial networks and demonstrates the application of GAN in biological data generation. Yelmen et al. [82] used generative neural networks to successfully create an artificial human genome, further driving innovation in DNA sequence design. Zrimec et al. [14] regulated DNA sequences through deep generative design to optimize gene expression and demonstrated the application of GAN in gene expression regulation. In addition, MedGAN proposed by Macedo et al. [83] optimized GAN in combination with Graph Convolutional Networks(GCN) for generating new molecular structures, further expanding the application of GAN in DNA sequence design and molecule generation. Through these studies, GAN provides a revolutionary technological pathway for DNA sequence design and gene expression optimization, which promotes the development of fields such as precision medicine, gene editing, and synthetic biology.

4.1.3 Diffusion models

Diffusion models (DM) are a class of generative models that have achieved remarkable success in recent years in generative tasks, especially in the fields of image generation, speech generation, and sequence generation. The basic idea is derived from the diffusion process in non-equilibrium thermodynamics [84], where the model simulates a forward stochastic differential equation (SDE) to progressively add noise to data, followed by a reverse SDE to recover the original data through iterative denoising. Unlike VAEs and GANs, DM generates DNA sequences by explicitly modeling the sequential corruption and reconstruction of sequence distributions [85]. Specifically, the diffusion model first performs a 'noise addition' process on the input DNA sequence, which gradually transforms it into a random noise sequence [86], and then gradually denoises it through a learned back-diffusion process to recover a DNA sequence that meets the functional and structural characteristics of the target Figure 2(c) [87]. The strengths of diffusion models stem from their ability to model high-dimensional correlations through the progressive denoising mechanism, as well as their flexible conditional control interfaces, which endow them with broad application prospects in the fields of precision medicine and synthetic biology [85, 88].

In recent years, DM has been gradually applied to DNA sequence design as an emerging method for generating and optimizing gene sequences. The potential diffusion model proposed by DaSilva et al. [85] provides an innovative framework for DNA sequence generation, which generates DNA sequences by mapping them into the potential space and accurately controls the properties of the generated sequences through an optimization process. Sarkar et al. [86] used the discrete diffusion model to design DNA sequences with modifiable activities, providing a new design tool for gene regulation and synthetic biology. Wang et al. [89] proposed the AptaDiff model specifically for the design of aptamers and used the diffusion model for the de novo design and optimization of novel aptamers, which effectively improved the targeting and specificity of the sequences. In addition, the Dirichlet diffusion model introduced by Avdeyev et al. [90] provides a new theoretical basis for biological sequence generation, which is based on the Dirichlet process to optimize the sequence generation process, thus enhancing the diversity and biological functions of the generated sequences. Through these studies, the diffusion model provides a powerful generative capability for the design and optimization of DNA sequences, especially showing great potential in the fields of gene expression regulation, aptamer design, and bioinformatics.

4.1.4 Large Language Models

With the emergence of large-scale pre-trained language models, especially in the field of Natural Language Processing (NLP), many researchers have started to

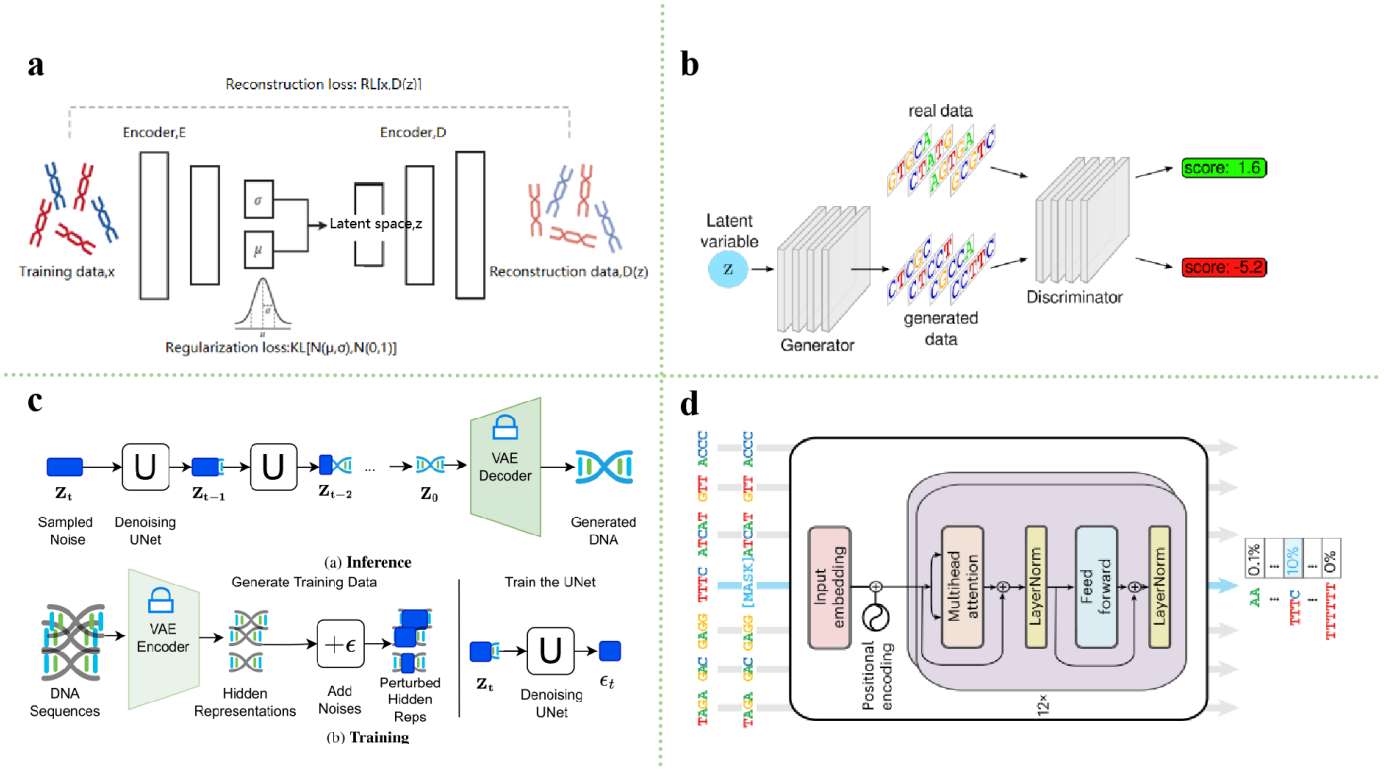

Figure 2. Several different model architectures for generating DNA sequences. (**a**) VAE model: the encoder maps the input data to the latent space, the decoder reconstructs the data from the latent representation, and the model is optimized by the reconstruction loss and the regularization loss. (**b**) GAN model: the generator produces synthetic DNA sequences, the discriminator evaluates the authenticity of the data, and the model is trained by optimizing the loss function of the discriminator. (**c**) Diffusion model: the inference process is shown in detail. Blue squares represent the noise added to the hidden representation of the DNA sequence. (**d**) Language model: after embedding the input data, the model processes the sequence with multi-head attention and layer normalization, then applies feed-forward networks to generate the output embeddings.

draw on the principles of language models to deal with genomic data and develop genomic language models [91]. The basic goal of language models is to predict the next word in a given context by learning the relationship between individual words in a language [92, 93]. In NLP, language models are usually trained based on large amounts of textual data to learn how to predict the next word based on the previous word, thus generating coherent sentences [94]. Analogous to DNA sequence design, DNA sequences can also be viewed as a form of "language," where 'words' can be base pairs (A, T, C, G) or vocabularies obtained through other tokenization techniques (such as K-mer, One-hot encoding) [95, 96]. The model learns the relative positions and order of bases or tokens in various ways, enabling it to generate reasonable and biologically meaningful DNA sequences, as shown in Figure 2(d) [97]. The core advantage of language models stems from the ability of the self-attention mechanism to model long-range dependencies and the multimodal conditional control interface, which has opened up a new paradigm for precision medicine and synthetic biology [98, 99].

Language models have been widely used in the design and analysis of DNA sequences, especially in the fields of genomics and synthetic biology. By using linguistic models, researchers can deeply understand the 'language' of DNA sequences and generate DNA sequences that meet specific needs. For example, Ji et al. [100] proposed the DNABERT model, based on the bidirectional encoder representation (BERT) architecture, which was successfully applied to the pre-training of DNA sequences and demonstrated its powerful language model ability in genome sequence analysis. Shao et al. [101] developed a long context language model for decoding and generating the bacteriophage genome, which provides a new idea for genome sequence generation. Nguyen et al. [13] proposed the Evo model, which combines molecular to genomic-scale sequence modeling and design, demonstrating its potential for large-scale genomic analysis and providing new insights for gene design. Recently, Madani et al. [102] successfully generated

functional protein sequences of multiple families using a large-scale language model. These studies show that language models provide powerful theoretical and technical support for the design of DNA sequences, and can help predict gene function, optimize gene expression, and promote the further development of genomics.

5 Applications of Nucleic Acid Sequence Design and Structure Prediction

5.1 Role of nucleic acid sequence design for structure prediction

Nucleic acid sequence design plays a crucial role in nucleic acid structure prediction. Through in-depth analysis of nucleic acid sequences, researchers can identify potential structural features, folding rules, and functional regions, providing powerful support for structure prediction. In particular, methods such as evolutionary information, multiple sequence comparison, covariance analysis, and secondary structure prediction can significantly improve the accuracy of structure prediction [103]. Meanwhile, sequence analysis models based on machine learning and deep learning can automatically learn the complex relationship between sequence and structure from a large amount of data, further promoting the development of nucleic acid structure prediction technology. For instance, Aslam et al. [104] demonstrated how adaptive machine learning frameworks enhance predictive accuracy in biological systems through domain-specific optimization. With the continuous enrichment of datasets and technological advances, nucleic acid sequence analysis will play a more important role in nucleic acid structure prediction in the future.

The functions of nucleic acids (e.g., catalysis, regulation, binding ligands, etc.) often depend on their 3D structures; therefore, predicting the 3D structures of nucleic acids can help design sequences with specific functions, enhance the accuracy of the design, and advance the understanding of the functions of nucleic acids, as well as optimize the process of sequence analysis, improve the accuracy of comparisons, support the discovery of drug targets, and guide experimental design and genetic engineering. Butt et al. [105] highlighted the integration of intelligent classification models in biomedical applications, which aligns with the need for precision in functional nucleic acid design. Structure prediction promotes a comprehensive understanding of nucleic acid sequences by providing insightful structural context for sequence analysis. DNA or RNA sequences in the genome form specific three-dimensional structures when folded, and these structures determine molecular interactions and biological functions. By accurately predicting these structures, it can help to design more functional molecules, rather than just predicting their basic function based on sequence. For example, in nanobiotechnology, self-assembling DNA or RNA molecules are designed to form specific nanostructures or nanomachines [106]. Rasool et al. [107] further exemplified this by developing a DNA-based file storage system optimized for medical data, demonstrating the synergy between sequence design and structural stability. Using 3D structure prediction, sequences with predetermined structures can be designed and their stability and functionality in practical applications can be ensured. Additionally, Rasool et al. [108] proposed a strategy-based optimization algorithm for DNA data storage encoding, underscoring the importance of sequence-structure co-design in emerging technologies. The two complement each other along with various deep learning approaches paving the way for each other.

5.2 Development and status of nucleic acid structure prediction

At the end of the 20th century, Westhof et al. [109] used molecular mechanics and molecular dynamics simulation software (e.g., AMBER, GROMOS, Xplor, etc.), which are based on classical physical models and algorithms to simulate the changes in nucleic acid structure by calculating the interaction energies between atoms. Multiple sequence comparison tools, such as CLUSTAL, were used to find similarities between sequences based on dynamic programming algorithms or heuristic algorithms. Subsequently, in the 21st century, methods based on comparative sequence analysis such as the RNAfold method [110] are limited by arithmetic power and algorithms and are not accurate enough in predicting complex RNA secondary structures. It is extremely difficult to fully understand the RNA folding mechanism. While data-driven methods are powerful in scenarios with limited mechanistic understanding, models integrating domain knowledge (e.g., thermodynamic rules or evolutionary conservation) often achieve superior interpretability and performance. For instance, hybrid approaches combining deep learning with biophysical principles have advanced RNA structure prediction [111]. These methods can learn the underlying folding patterns from large amounts

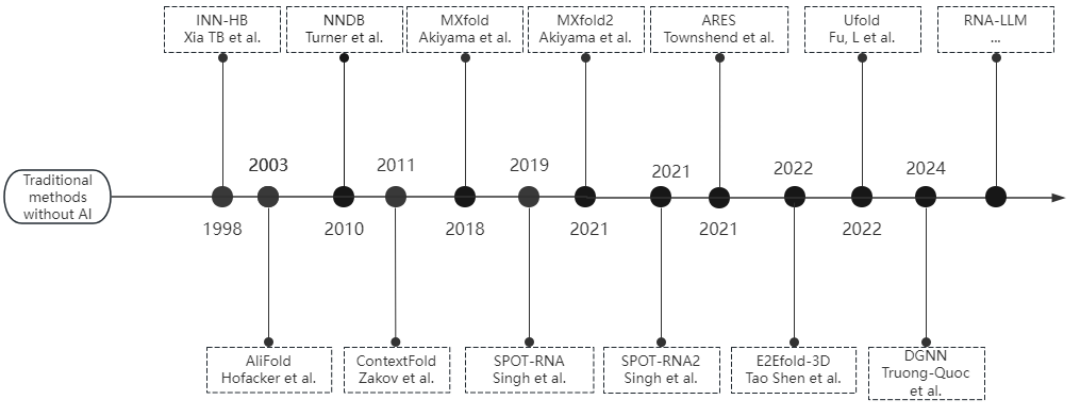

Figure 3. Progress in the field of nucleic acid structure prediction.

of training data. Over the past decades, machine learning and deep learning methods have been used in many aspects of RNA secondary structure prediction methods to improve prediction performance. Figure 3 illustrates the development of nucleic acid structure prediction in the AI era.

RNA secondary structure prediction methods based on machine learning and deep learning typically learn functions that map inputs (features) to outputs by tuning model parameters based on known input and output pairs. Many of them employ free energy parameters, encoded RNA sequences, sequence patterns, or evolutionary information as key features, and their outputs can be categorical labels (e.g., paired or unpaired) or continuous values (e.g., free energy). When new inputs are provided to a trained model, the model can classify the corresponding labels or predict the corresponding values [112]. The nearest neighbor model (NNDB) developed by Turner [113] was an early and fairly commonly used approach, which provided a fairly accurate approximation of RNA free energy. However, several thermodynamic parameters of the NNDB model had to be based on a large number of optimal melting experiments, which were both time-consuming and labor-intensive [55, 114], and due to the associated technical difficulties, not all free energy changes in the structural elements could be measured. Due to the difficulty of obtaining the relevant parameters, several machine-learning techniques have been used to optimize the parameters in energy models. These techniques can use fine-grained models that estimate fractions using known thermodynamic data or RNA secondary structure data to obtain richer and more accurate representations of the features. Xia et al. [115] first trained a linear regression model using known thermodynamic data to infer some of the thermodynamic parameters and extended the neural network model to a more accurate model, the INN-HB model. This model provides a better fit for known experimental data. However, a disadvantage of this approach is that the parameters of some structural elements are fixed before other parameters are calculated, which limits the range of possibilities for the entire parameter set.

Although machine learning-based free energy parameter methods have successfully improved the accuracy of RNA secondary structure prediction, the energy model is still far from ideal. The machine learning-based parameter estimation methods can only replace some wet lab experiments aimed at obtaining energy parameters. As a result, Zakov et al. [116] proposed the ContextFold tool, which not only relies on traditional energy parameters but also takes into account the contextual information in the RNA structure, significantly improving the accuracy and flexibility of RNA structure prediction. Later, Akiyama et al. [117] integrated thermodynamic methods with SSVM and developed MXfold, which overcame the limitations of traditional tools in energy model optimization and large-scale data processing, but had limited prediction accuracy for long-stranded RNAs (e.g., mRNAs or lncRNAs) and did not support the prediction of pseudo-knots. Sato et al. [118] developed MXfold2 to overcome the above-mentioned limitations, marking a new level of performance in RNA structure prediction.

Since 2020, more and more models for predicting the 3D structure of nucleic acids based on deep learning methods have appeared in the public eye with innovations in technologies related to deep learning and nucleic acid sequence design. Linyu Wang's team proposed a new method called DMfold [119] based

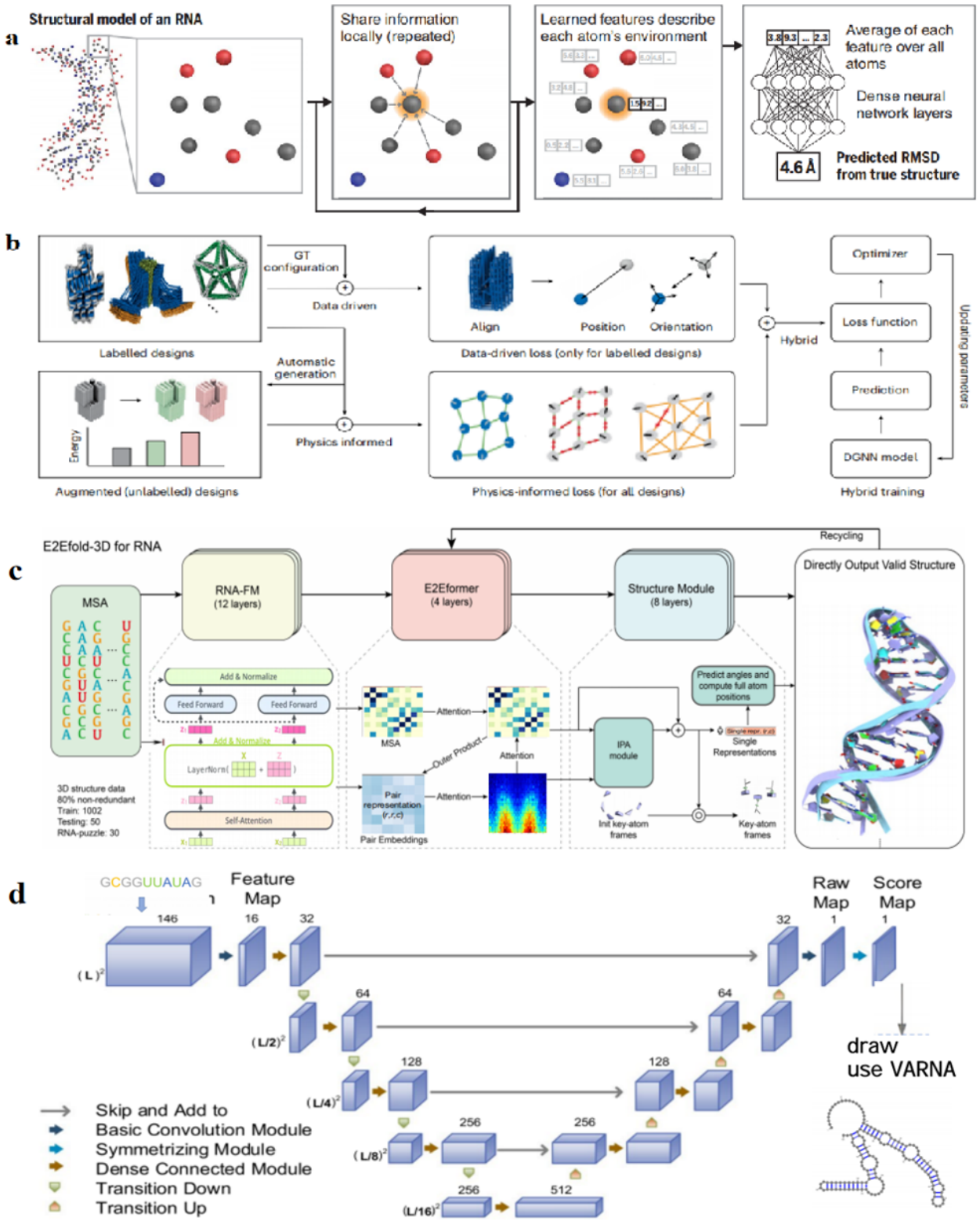

Figure 4. Several different model architectures for predicting nucleic acid 3D structures. (**a**) ARES network updates features based on atomic elements and coordinates and predicts the RMSD of RNA structure by averaging all atomic features. (**b**) DGNN is trained by obtaining 252 labeled designs through SNUPI, generating augmented designs, and using different loss functions with parameter optimization using Adam optimizer. (**c**) E2Efold-3D predicts 3D structures from scratch using RNA sequence information, initializes features by MSA and pre-trained models, generates structures using E2Eformer, and optimizes the loss function by structural constraints. (**d**) REDfold network converts RNA sequences to conformations and predicts secondary structure score maps through feature extraction and encoder-decoder network, and finally draws structure images.

on Bi-LSTM to predict the secondary structure of RNA containing pseudoknots, which combines deep learning and the improved base pair maximization principle, fully absorbing the advantages and avoiding some of the disadvantages of the multi-sequence method (MPM) and the single-sequence method (SPM), but DMfold does not achieve the optimal accuracy in the case of known However, the accuracy of DMfold is not optimal when the sequence is known to be insufficient, and with the continuous optimization of sequence design technology, DMfold will gain more space for development. Kunitski et al. [120] proposed the first end-to-end deep learning model SPOT-RNA to predict the secondary structure of RNA.SPOT-RNA treats RNA secondary structure as a CT table and uses an ultradeep hybrid network that combines ResNet and 2D-BBN to predict the secondary structure of RNA. SPOT-RNA treats RNA secondary structures as CT tables and uses an ultra-deep hybrid network that combines the ResNet and 2D-BLSTM networks to make predictions, with the ResNet network capturing contextual information across the entire sequence and the 2D-BLSTM efficiently propagating long-range dependencies in RNA structures. Through migration learning, SPOT-RNA can achieve good results even with limited samples. Experimental results showed that SPOT-RNA performed well on multiple RNA benchmark datasets, and in a follow-up development, the SPOT-RNA2 model was proposed by the same research group [121]. This model employs evolutionarily derived sequence profiles and mutation coupling as inputs, uses the same migration learning approach, and outperforms SPOT-RNA in all types of base pair prediction. Shen et al. [122] developed the E2Efold-3D model, as shown in Figure 4(c), which is another deep learning for predicting RNA secondary structure, E2Efold-3D is another method that can directly predict the 3D structure of RNA without external templates, which greatly improves the accuracy of RNA structure prediction through secondary structure-assisted self-distillation, multidimensional information fusion, and joint training, especially in addressing data scarcity and structural complexity, showing its unique advantages. It is worth mentioning that E2Efold uses a deep learning approach to achieve end-to-end RNA structure prediction by obtaining binary classification scoring matrices of base pairs from input RNA sequences, whereas the ARES model, as shown in Figure 4(a), which is different from the classification model E2Efold, is a regression model based on geometric deep learning developed by Townshend et al. [123] only trained a new RNA tertiary structure scoring model from 18 known RNA tertiary structures published between 1994 and 2006. The input to ARES is the 3D coordinates and chemical element type of each atom, and the output is the root-mean-square deviation (RMSD) between the predicted structural model and the true structure, which means that ARES learning is completely featureless without any predefined features, and the model directly learns and extracts features from the raw data, effectively reducing human bias and significantly excelling in the discovery of new features. ARES significantly outperforms other scoring functions and models despite using a limited number of known RNA structures. The REDfold model, as shown in Figure 4(d) [124], which uses a CNN-based encoder-decoder network to learn the dependencies in RNA sequences and effectively propagates information across layers through symmetric skip connections, achieves better performance in both efficiency and accuracy for RNA secondary structure prediction. Another RNA prediction model, Ufold, instead of directly inputting the 3D coordinates and chemical element type of each atom, inputs a matrix of all possible base pairs (canonical and non-canonical base pairs) and pairing features of the RNA sequence. Ufold converts the input matrix and pairing features into base pairing probabilities for predicting the RNA secondary structure through the use of a fully convolutional network (FCN) [125].

The outstanding contribution of nucleic acid sequence design in structure prediction is not only in the direction of RNA prediction but also in the direction of DNA, Chien Truong-Quoc's team has developed a DNA-origami-based graph neural network (DGNN), as shown in Figure 4(b) [126], which is more focused on the instantaneous and accurate prediction of DNA origami structures. The authors' innovative hybrid data-driven and physically-guided training approach greatly alleviates the difficulty of training purely data-driven models for DNA origami design, as the dataset is very scarce, and the authors also integrate pre-trained models corresponding to various shapes of DNA origami, which enhances the adaptability and performance of the DGNN for various types of data. The DGNN also makes outstanding contributions in the areas of structure prediction of supramolecular assemblies and inverse design of DNA origami. DGNN has also made outstanding contributions in the areas of supramolecular assembly structure prediction and DNA origami inverse design. Recently, many

large language models (RNA-LLM) for RNA structure prediction have appeared [127], and perhaps shortly, there may be new advances in the combination of nucleic acid structure prediction and large language models.

6 Challenges and future directions

Despite significant progress made by intelligent design and optimization methods in the field of nucleic acid sequences, multiple technical bottlenecks and systemic challenges remain. Existing computational models can perform preliminary predictions using rule engines, machine learning, and generative approaches; however, considerable uncertainty persists when dealing with complex nonlinear systems such as gene regulatory networks [128, 129]. The experimental validation phase relies on high-throughput screening technologies, which are plagued by the high cost of equipment and consumables, thereby limiting the overall research and development efficiency of the "computation-experiment" dual-track validation model. Moreover, practical applications of artificial intelligence face three key constraints: computational resources, data quality, and ethical standards. Training deep generative models consumes enormous computational power—for instance, handling databases with millions of sequences requires terabyte-level storage and GPU clusters [130]; noise data in public nucleic acid sequence databases can account for as much as 30%, and the scarcity of data for certain disease-related targets severely undermines model reliability [131, 132]; additionally, the "black box" nature of AI may lead to issues such as untraceable decision-making and algorithmic bias. For example, in gene editing design, the absence of an interpretable mechanism might conceal potential off-target risks [133].

Nevertheless, AI technologies have demonstrated revolutionary empowerment potential across the entire nucleic acid sequence design chain. From RNA three-dimensional structure prediction algorithms expanded from AlphaFold3 to target-sequence matching systems incorporating Transformer architectures, AI has reduced traditional design cycles from months to weeks [134]. Furthermore, deep generative models can automatically generate candidate sequences that satisfy specific binding energy, stability, and functionality criteria; when combined with automated experimental platforms, they enable high-throughput synthesis and validation [16]. This closed-loop approach not only accelerates the development of antiviral nucleic acid drugs but also provides a new paradigm for the design of cancer vaccines and gene editing tools (such as CRISPR). However, the ultimate challenge of technological implementation lies in balancing the pace of innovation with risk management. Only by establishing a multidimensional governance framework that encompasses technological interpretability, data ethics, and public engagement can we ensure that AI truly advances nucleic acid sequence design toward precision and responsible innovation.

Future breakthroughs should focus on multimodal data integration and interdisciplinary collaborative innovation. By integrating multidimensional data from genomics, epigenetics, and proteomics, a more comprehensive framework for sequence function prediction can be constructed. At the same time, developing lightweight model architectures (such as those based on knowledge distillation) and distributed computing solutions is expected to lower the computational threshold, while the adoption of privacy-preserving techniques like federated learning can help alleviate data silo issues. On the ethical front, it is necessary to establish a transparent mechanism that spans the entire lifecycle of model design, training, and deployment—for example, by introducing linear artificial chromatography imaging methods to elucidate neural network decision paths and by supervising technological compliance through a dynamic ethics review committee.

The application of artificial intelligence in nucleic acid sequence design has already demonstrated enormous potential, yet its practical implementation still requires balancing the speed of innovation with risk management. Only by building a multidimensional governance system that includes technological interpretability, data ethics, and public participation can AI truly drive nucleic acid sequence design toward precision and responsible innovation. Through technological breakthroughs, expanded applications, and improved governance, AI is poised to usher in a new revolution in the life sciences, providing a powerful engine for human health and sustainable development.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported in part by the 111 Project under Grant D23006; in part by the National Natural Science Foundation of China under Grant 62272079; in part by the National Foreign Expert Project of China under Grant D20240244; in part by the Natural Science Foundation of Liaoning Province under Grant 2024-MS-212; in part by the Scientific Research Project of Liaoning Provincial Department of Education under Grant LJ222411258005; in part by the LiaoNing Revitalization Talent Program under Grant XLYC2403039; in part by the Artificial Intelligence Innovation Development Plan Project of Liaoning Province under Grant 2023JH26/10300025; in part by the Dalian Outstanding Young Science and Technology Talent Support Program under Grant 2022RJ08; in part by the Dalian Major Projects of Basic Research under Grant 2023JJ11CG002; in part by the Interdisciplinary Project of Dalian University under Grant DLUXK-2024-YB-001; in part by the Joint plan of Liaoning Province science and technology plan under Grant 2024JH2/102600064 and Grant 2024-MSLH-009.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Shin, S., Lee, I., Kim, D., & Zhang, B. (2005). Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing. *IEEE Transactions on Evolutionary Computation, 9*(2), 143-158. [CrossRef]

[2] Feldkamp, U., Rauhe, H., & Banzhaf, W. (2003). Software tools for DNA sequence design. *Genetic Programming and Evolvable Machines, 4*, 153-171. [CrossRef]

[3] Raab, D., Graf, M., Notka, F., Schödl, T., & Wagner, R. (2010). The GeneOptimizer algorithm: Using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. *Systems and Synthetic Biology, 4*(3), 215-225. [CrossRef]

[4] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet., 9*(1), 387-402. [CrossRef]

[5] Bates, M., Lachoff, J., Meech, D., Zulkower, V., Moisy, A., Luo, Y., Tekotte, H., Franziska Scheitz, C. J., Khilari, R., Mazzoldi, F., Chandran, D., & Groban, E. (2017). Genetic constructor: An online DNA design platform. ACS Synthetic Biology, 6(12), 2362-2365. [CrossRef]

[6] Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J., & Govindarajan, S. (2006). Gene designer: A synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics, 7*(1). [CrossRef]

[7] Angelbello, A. J., Chen, J. L., Childs-Disney, J. L., Zhang, P., Wang, Z., & Disney, M. D. (2018). Using genome sequence to enable the design of medicines and chemical probes. *Chemical Reviews, 118*(4), 1599-1663. [CrossRef]

[8] Alarcon, C. M., Shan, G., Layton, D. T., Bell, T. A., Whipkey, S., & Shillito, R. D. (2018). Application of DNA- and protein-based detection methods in agricultural biotechnology. *Journal of Agricultural and Food Chemistry, 67*(4), 1019-1028. [CrossRef]

[9] Tyo, K. E., Kocharin, K., & Nielsen, J. (2010). Toward design-based engineering of industrial microbes. *Current Opinion in Microbiology, 13*(3), 255-262. [CrossRef]

[10] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature, 494*(7435), 77-80. [CrossRef]

[11] Zhang, P., Wang, H., Xu, H., Wei, L., Liu, L., Hu, Z., & Wang, X. (2023). Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nature Communications, 14*(1). [CrossRef]

[12] Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nature Reviews Chemistry, 7*(3), 144-161. [CrossRef]

[13] Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., Ng, M. Y., Lewis, A., Patel, A., Lou, A., Ermon, S., Baccus, S. A., Hernandez-Boussard, T., Re, C., Hsu, P. D., & Hie, B. L. (2024). Sequence modeling and design from molecular to genome scale with Evo. *Science, 386*(6723), eado9336. [CrossRef]

[14] Zrimec, J., Fu, X., Muhammad, A. S., Skrekas, C., Jauniskis, V., Speicher, N. K., Börlin, C. S., Verendel, V., Chehreghani, M. H., Dubhashi, D., Siewers, V., David, F., Nielsen, J., & Zelezniak, A. (2022). Controlling gene expression with deep generative design of regulatory DNA. *Nature Communications, 13*(1), 5099. [CrossRef]

[15] Keskin Karakoyun, H., Yüksel, Ş. K., Amanoglu, I., Naserikhojasteh, L., Yeşilyurt, A., Yakıcıer, C., Timuçin, E., & Akyerli, C. B. (2023). Evaluation of AlphaFold structure-based protein stability prediction on missense variations in cancer. *Frontiers in Genetics, 14*, 1052383. [CrossRef]

[16] Gosai, S. J., Castro, R. I., Fuentes, N., Butts, J. C., Mouri, K., Alasoadura, M., Kales, S., Nguyen, T. T., Noche, R. R., Rao, A. S., Joy, M. T., Sabeti, P. C., Reilly, S. K., & Tewhey, R. (2024). Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature, 634*(8036), 1211-1220. [CrossRef]

[17] Cox, B., Denyer, J. C., Binnie, A., Donnelly, M. C.,

Evans, B., Green, D. V., ... & Watson, S. P. (2000). Application of high-throughput screening techniques to drug discovery. *Progress in Medicinal Chemistry, 37*, 83-133. [CrossRef]

[18] Gervasio, J. H. D. B., da Costa Oliveira, H., da Costa Martins, A. G., Pesquero, J. B., Verona, B. M., & Cerize, N. N. P. (2024). How close are we to storing data in DNA?. *Trends in Biotechnology, 42*(2), 156-167. [CrossRef]

[19] Clark, D. P., & Pazdernik, N. J. (2012). *Molecular biology*. Elsevier.

[20] Waterman, M. S. (2018). *Introduction to computational biology: maps, sequences and genomes*. Chapman and Hall/CRC.

[21] Watson, J. D., & Crick, F. H. (1953, January). The structure of DNA. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 18, pp. 123-131). Cold Spring Harbor Laboratory Press. [CrossRef]

[22] Travers, A., & Muskhelishvili, G. (2015). DNA structure and function. *The FEBS journal, 282*(12), 2279-2295. [CrossRef]

[23] Komili, S., Farny, N. G., Roth, F. P., & Silver, P. A. (2007). Functional specificity among Ribosomal proteins regulates gene expression. *Cell, 131*(3), 557-571. [CrossRef]

[24] Neylon, C. (2004). Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic acids research, 32*(4), 1448-1459. [CrossRef]

[25] Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in Protein-DNA recognition. *Annual Review of Biochemistry, 79*(1), 233-269. [CrossRef]

[26] Beerli, R. R., Segal, D. J., Dreier, B., & Barbas III, C. F. (1998). Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proceedings of the National Academy of Sciences, 95*(25), 14628-14633. [CrossRef]

[27] Goldenzweig, A., Goldsmith, M., Hill, S., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R., Aharoni, A., Silman, I., Sussman, J., Tawfik, D., & Fleishman, S. (2016). Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Molecular Cell, 63*(2), 337-346. [CrossRef]

[28] Matange, K., Tuck, J. M., & Keung, A. J. (2021). DNA stability: A central design consideration for DNA data storage systems. *Nature Communications, 12*(1), 1-9. [CrossRef]

[29] Szostak, J. W., Bartel, D. P., & Luisi, P. L. (2001). Synthesizing life. *Nature, 409*(6818), 387-390. [CrossRef]

[30] Bulyk, M. L. (2003). Computational prediction of transcription-factor binding site locations. *Genome biology, 5*, 1-11. [CrossRef]

[31] Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science, 324*(5924), 255-258. [CrossRef]

[32] Francis, D. M., & Page, R. (2010). Strategies to optimize protein expression inE. coli. Current Protocols in Protein Science, 61(1), 5.24. 1-5.24. 29. [CrossRef]

[33] Murakami, S., & Jaffrey, S. R. (2022). Hidden codes in mRNA: Control of gene expression by m6A. *Molecular Cell, 82*(12), 2236-2251. [CrossRef]

[34] Dowell, R. D., & Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics, 5*(1), 1-14. [CrossRef]

[35] WANG, Y., WANG, H., YAN, M., HU, G., & WANG, X. (2021). Design of biomolecular sequences by artificial intelligence. *Synthetic Biology Journal, 2*(1), 1-14.

[36] Condon, A. (2006). Designed DNA molecules: Principles and applications of molecular nanotechnology. *Nature Reviews Genetics, 7*(7), 565-575. [CrossRef]

[37] Lathe, R. (1985). Synthetic oligonucleotide probes deduced from amino acid sequence data. *Journal of Molecular Biology, 183*(1), 1-12. [CrossRef]

[38] Newman, Z. R., Young, J. M., Ingolia, N. T., & Barton, G. M. (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proceedings of the National Academy of Sciences, 113*(10). [CrossRef]

[39] Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., & Gileadi, O. (2008). Codon optimization can improve expression of human genes in escherichia coli: A multi-gene study. *Protein Expression and Purification, 59*(1), 94-102. [CrossRef]

[40] Gingold, H., & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. Molecular Systems Biology, 7(1), 481. [CrossRef]

[41] Sharp, P. M., & Li, W. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research, 15*(3), 1281-1295. [CrossRef]

[42] Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T., Puetz, L., Rasmussen, M., Zervas, A., & Hansen, L. H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience, 9*(2), giaa008. [CrossRef]

[43] Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology, 288*(5), 911-940. [CrossRef]

[44] Brahmachari, S. K., Meera, G., Sarkar, P. S.,

Balagurumoorthy, P., Tripathi, J., Raghavan, S., Shaligram, U., & Pataskar, S. (1995). Simple repetitive sequences in the genome: Structure and functional significance. *ELECTROPHORESIS, 16*(1), 1705-1714. [CrossRef]

[45] Van Belkum, A., Scherer, S., Van Alphen, L., & Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews, 62*(2), 275-293. [CrossRef]

[46] Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36-46. [CrossRef]

[47] Kosuri, S., & Church, G. M. (2014). Large-scale de Novo DNA synthesis: Technologies and applications. *Nature Methods, 11*(5), 499-507. [CrossRef]

[48] Chen, Z., Pan, N., & Beachy, R. N. (1988). A DNA sequence element that confers seed-specific enhancement to a constitutive promoter. *The EMBO Journal, 7*(2), 297-302. [CrossRef]

[49] Slobodin, B., Han, R., Calderone, V., Vrielink, J. A., Loayza-Puch, F., Elkon, R., & Agami, R. (2017). Transcription impacts the efficiency of mRNA translation via Co-transcriptional N6-adenosine methylation. *Cell, 169*(2), 326-337.e12. [CrossRef]

[50] Shaul, O. (2017). How introns enhance gene expression. *The International Journal of Biochemistry & Cell Biology, 91*, 145-155. [CrossRef]

[51] Studier, F., & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology, 189*(1), 113-130. [CrossRef]

[52] Kozak, M. (1986). Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences, 83*(9), 2850-2854. [CrossRef]

[53] Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene, 361*, 13-37. [CrossRef]

[54] Milenkovic, O., & Kashyap, N. (2005). DNA codes that avoid secondary structures. *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005*, 288-292. [CrossRef]

[55] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for molecular biology, 6*, 1-14. [CrossRef]

[56] Schlake, T., Thess, A., Fotin-Mleczek, M., & Kallen, K. (2012). Developing mrna-vaccine technologies. *RNA Biology, 9*(11), 1319-1330. [CrossRef]

[57] Arita, M., & Kobayashi, S. (2002). DNA sequence design using templates. *New Generation Computing, 20*(3), 263-277. [CrossRef]

[58] Ling, M. M., & Robinson, B. H. (1997). Approaches to DNA mutagenesis: An overview. *Analytical Biochemistry, 254*(2), 157-178. [CrossRef]

[59] Wachsmuth, M., Findeiss, S., Weissheimer, N., Stadler, P. F., & Morl, M. (2012). De Novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Research, 41*(4), 2541-2551. [CrossRef]

[60] Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology, 12*(7), 878. [CrossRef]

[61] Brophy, J. A., & Voigt, C. A. (2014). Principles of genetic circuit design. *Nature methods, 11*(5), 508-520. [CrossRef]

[62] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature, 596*(7873), 583-589.

[63] Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., ... & Huang, L. (2021). Lineardesign: Efficient algorithms for optimized mrna sequence design.

[64] Liu, J., Li, J., Wang, H., & Yan, J. (2020). Application of deep learning in genomics. *Science China Life Sciences, 63*, 1860-1878. [CrossRef]

[65] Li, X., Cao, B., Wang, J., Meng, X., Wang, S., Huang, Y., ... & Song, T. (2025). Predicting mutation-disease associations through protein interactions via deep learning. *IEEE Journal of Biomedical and Health Informatics*. [CrossRef]

[66] Killoran, N., Lee, L. J., DeLong, A., Duvenaud, D., & Frey, B. J. (2017). Generating and designing DNA with deep generative models. *arXiv preprint arXiv:1712.06148*.

[67] Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature Methods, 15*(10), 816-822. [CrossRef]

[68] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv e-prints*, arXiv-1312. [CrossRef]

[69] Sumi, S., Hamada, M., & Saito, H. (2024). Deep generative design of RNA family sequences. *Nature Methods, 21*(3), 435-443. [CrossRef]

[70] Seo, E., Choi, Y., Shin, Y., Kim, D., & Lee, J. (2023). Design of synthetic promoters for cyanobacteria with generative deep-learning model. *Nucleic Acids Research, 51*(13), 7071-7082. [CrossRef]

[71] Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., & Bikard, D. (2021). Generating functional protein variants with variational autoencoders. *PLOS Computational Biology, 17*(2), e1008736. [CrossRef]

[72] Sadeghi, E., Mastracco, P., Gonzàlez-Rosell, A., Copp, S. M., & Bogdanov, P. (2024). Multi-objective design of DNA-stabilized Nanoclusters using variational Autoencoders with automatic feature extraction. *ACS Nano, 18*(39), 26997-27008. [CrossRef]

[73] Greener, J. G., Moffat, L., & Jones, D. T. (2018).

Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports, 8*(1), 16189. [CrossRef]

[74] Moomtaheen, F., Killeen, M., Oswald, J., Gonzàlez-Rosell, A., Mastracco, P., Gorovits, A., Copp, S. M., & Bogdanov, P. (2022). DNA-stabilized silver Nanocluster design via regularized variational Autoencoders. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3593-3602. [CrossRef]

[75] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. [CrossRef]

[76] Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems, 28*.

[77] Barazandeh, S., Ozden, F., Hincer, A., Seker, U. O. S., & Cicek, A. E. (2023). Utrgan: Learning to generate 5'utr sequences for optimized translation efficiency and gene expression. *bioRxiv*, 2023-01. [CrossRef]

[78] Dai, J., Zhang, Y., Shi, C., Liu, Y., Xiu, P., & Wang, Y. (2024). BEGAN: Boltzmann-Reweighted Data Augmentation for Enhanced GAN-Based Molecule Design in Insect Pheromone Receptors. *The Journal of Physical Chemistry B, 128*(47), 11666-11675. [CrossRef]

[79] Chiquitto, A. G., Oliveira, L. S., Bugatti, P. H., Saito, P. T. M., Basham, M., Raittz, R. T., & Paschoal, A. R. (2024). Generative Approaches for Nucleotide Sequences to Enhance Non-coding RNA Classification. *bioRxiv*, 2024-11. [CrossRef]

[80] Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., & Jay, F. (2023). Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLOS Computational Biology, 19*(10), e1011584. [CrossRef]

[81] Yu, H., & Welch, J. D. (2021). MichiGAN: sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome biology, 22*(1), 158. [CrossRef]

[82] Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., ... & Jay, F. (2021). Creating artificial human genomes using generative neural networks. *PLoS genetics, 17*(2), e1009303. [CrossRef]

[83] Macedo, B., Ribeiro Vaz, I., & Taveira Gomes, T. (2024). MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Scientific reports, 14*(1), 1212. [CrossRef]

[84] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256-2265). pmlr.

[85] DaSilva, L. F., Senan, S., Patel, Z. M., Reddy, A. J., Gabbita, S., Nussbaum, Z., ... & Pinello, L. (2024). DNA-diffusion: leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements. *bioRxiv*. [CrossRef]

[86] Sarkar, A., Tang, Z., Zhao, C., & Koo, P. K. (2024). Designing DNA with tunable regulatory activity using discrete diffusion. *bioRxiv*, 2024-05. [CrossRef]

[87] Li, Z., Ni, Y., Huygelen, T. A. B., Das, A., Xia, G., Stan, G. B., & Zhao, Y. (2023). Latent diffusion model for dna sequence generation. *arXiv preprint arXiv:2310.06150*. [CrossRef]

[88] Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. (2022). Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems, 35*, 9754-9767.

[89] Wang, Z., Liu, Z., Zhang, W., Li, Y., Feng, Y., Lv, S., ... & Li, X. (2024). AptaDiff: de novo design and optimization of aptamers based on diffusion models. *Briefings in Bioinformatics, 25*(6), bbae517. [CrossRef]

[90] Avdeyev, P., Shi, C., Tan, Y., Dudnyk, K., & Zhou, J. (2023, July). Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning* (pp. 1276-1301). PMLR.

[91] Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., ... & Wang, B. (2023). To transformers and beyond: large language models for the genome. *arXiv preprint arXiv:2311.07621*. [CrossRef]

[92] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137-1155.

[93] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[94] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[95] Cao, B., Wang, B., & Zhang, Q. (2023). GCNSA: DNA storage encoding with a graph convolutional network and self-attention. *Iscience, 26*(3). [CrossRef]

[96] Zheng, Y., Cao, B., Zhang, X., Cui, S., Wang, B., & Zhang, Q. (2024). DNA-QLC: an efficient and reliable image encoding scheme for DNA storage. *BMC genomics, 25*(1), 266. [CrossRef]

[97] Sanabria, M., Hirsch, J., Joubert, P. M., & Poetsch, A. R. (2024). DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence, 6*(8), 911-923. [CrossRef]

[98] Zhang, D., Zhang, W., Zhao, Y., Zhang, J., He, B., Qin, C., & Yao, J. (2023). DNAGPT: a generalized pre-trained tool for versatile DNA sequence analysis tasks. *arXiv preprint arXiv:2307.05628*. [CrossRef]

[99] Chen, Y., & Zou, J. (2024). GenePT: a simple but

effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2023-10. [CrossRef]

[100] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics, 37*(15), 2112-2120. [CrossRef]

[101] Shao, B., & Yan, J. (2024). A long-context language model for deciphering and generating bacteriophage genomes. *Nature Communications, 15*(1), 9392. [CrossRef]

[102] Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., ... & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. *Nature biotechnology, 41*(8), 1099-1106. [CrossRef]

[103] Tinoco Jr, I., & Bustamante, C. (1999). How RNA folds. *Journal of molecular biology, 293*(2), 271-281. [CrossRef]

[104] Aslam, S., Rasool, A., Li, X., & Wu, H. (2025). Cel: A continual learning model for disease outbreak prediction by leveraging domain adaptation via elastic weight consolidation. *Interdisciplinary Sciences: Computational Life Sciences*, 1-19. [CrossRef]

[105] Butt, M. H. F., Li, J. P., Ji, J., Riaz, W., Anwar, N., Butt, F. F., ... & Uddin, M. Y. (2024). Intelligent tumor tissue classification for Hybrid Health Care Units. *Frontiers in Medicine, 11*, 1385524. [CrossRef]

[106] Dutt, Y., Pandey, R. P., Dutt, M., Gupta, A., Vibhuti, A., Vidic, J., ... & Priyadarshini, A. (2023). Therapeutic applications of nanobiotechnology. *Journal of nanobiotechnology, 21*(1), 148. [CrossRef]

[107] Rasool, A., Hong, J., Hong, Z., Li, Y., Zou, C., Chen, H., ... & Dai, J. (2024). An Effective DNA-Based File Storage System for Practical Archiving and Retrieval of Medical MRI Data. *Small Methods, 8*(10), 2301585. [CrossRef]

[108] Rasool, A., Qu, Q., Jiang, Q., & Wang, Y. (2021, December). A strategy-based optimization algorithm to design codes for DNA data storage system. In *International Conference on Algorithms and Architectures for Parallel Processing* (pp. 284-299). Cham: Springer International Publishing. [CrossRef]

[109] Westhof, E. R. I. C., Auffinger, E., & Gaspin, C. (1996). DNA and RNA structure prediction. *DNA–Protein Sequence Analysis, Oxford*, 255-278.

[110] Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic acids research, 31*(13), 3429-3431. [CrossRef]

[111] Wen, X., Sun, L., Xie, L., Zheng, Y., Cao, B., & Wang, B. (2024, December). MFN: Explainable DNA triple helixes Stabilized Design based on mCGR and flow network. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 248-253). IEEE. [CrossRef]

[112] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. [CrossRef]

[113] Turner, D. H., & Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic acids research, 38(suppl_1), D280-D282. [CrossRef]

[114] Bellaousov, S., Reuter, J. S., Seetin, M. G., & Mathews, D. H. (2013). RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic acids research, 41*(W1), W471-W474. [CrossRef]

[115] Xia, T., SantaLucia Jr, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., ... & Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry, 37*(42), 14719-14735. [CrossRef]

[116] Zakov, S., Goldberg, Y., Elhadad, M., & Ziv-Ukelson, M. (2011). Rich parameterization improves RNA structure prediction. *Journal of Computational Biology, 18*(11), 1525-1542. [CrossRef]

[117] Akiyama, M., Sato, K., & Sakakibara, Y. (2018). A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *Journal of bioinformatics and computational biology, 16*(06), 1840025. [CrossRef]

[118] Sato, K., Akiyama, M., & Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. Nature communications, 12(1), 941. [CrossRef]

[119] Wang, L., Liu, Y., Zhong, X., Liu, H., Lu, C., Li, C., & Zhang, H. (2019). DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Frontiers in genetics, 10*, 143. [CrossRef]

[120] Kunitski, M., Eicke, N., Huber, P., Köhler, J., Zeller, S., Voigtsberger, J., ... & Dörner, R. (2019). Double-slit photoelectron interference in strong-field ionization of the neon dimer. *Nature communications, 10*(1), 1. [CrossRef]

[121] Singh, J., Paliwal, K., Zhang, T., Singh, J., Litfin, T., & Zhou, Y. (2021). Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics, 37*(17), 2589-2600. [CrossRef]

[122] Shen, T., Hu, Z., Peng, Z., Chen, J., Xiong, P., Hong, L., ... & Li, Y. (2022). E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv preprint arXiv:2207.01586*. [CrossRef]

[123] Townshend, R. J., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., & Dror, R. O. (2021). Geometric deep learning of RNA structure. *Science, 373*(6558), 1047-1051. [CrossRef]

[124] Chen, C. C., & Chan, Y. M. (2023). REDfold: accurate

RNA secondary structure prediction using residual encoder-decoder network. *BMC bioinformatics, 24*(1), 122. [CrossRef]

[125] Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2022). UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic acids research, 50*(3), e14-e14. [CrossRef]

[126] Truong-Quoc, C., Lee, J. Y., Kim, K. S., & Kim, D. N. (2024). Prediction of DNA origami shape using graph neural network. *Nature Materials, 23*(7), 984-992. [CrossRef]

[127] Zablocki, L. I., Bugnon, L. A., Gerard, M., Di Persia, L., Stegmayer, G., & Milone, D. H. (2025). Comprehensive benchmarking of large language models for RNA secondary structure prediction. *Briefings in Bioinformatics, 26*(2), bbaf137. [CrossRef]

[128] Tušek, A., & Kurtanjek, Z. (2012). Mathematical modelling of gene regulatory networks. Applied Biological Engineering—Principles and Practice.

[129] Lu, T., Liang, H., Li, H., & Wu, H. (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association, 106*(496), 1242-1258. [CrossRef]

[130] Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023, March). Sparks of artificial general intelligence: Early experiments with gpt-4.

[131] Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics, 14*(10), 681-691. [CrossRef]

[132] Jucker, M., & Walker, L. C. (2018). Propagation and spread of pathogenic protein assemblies in neurodegenerative diseases. *Nature neuroscience, 21*(10), 1341-1349. [CrossRef]

[133] Thalpage, N. (2023). Unlocking the black box: Explainable artificial intelligence (XAI) for trust and transparency in ai systems. *J. Digit. Art Humanit, 4*(1), 31-36.

[134] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*(8016), 493-500. [CrossRef]

Ting Yang graduated from Nanchang Institute of Technology with a Bachelor's degree in Software Engineering. She is currently pursuing a Master's degree in Software Engineering at Dalian University. (Email: yting0784@gmail.com)

Minxu Han received the B.E. degree in Computer Science and Technology from Weifang University in 2020. Now, he is working on his Master's degree in software engineering at Dalian University. (Email: hanminxu248@gmail.com)

Xiaoru Wen graduated from Hunan Institute of Technology with a Bachelor's degree in Computer science and Technology. She is currently pursuing a Master's degree in Software Engineering at Dalian University. (Email: xiaoruwen39@gmail.com)

Yanfen Zheng graduated from Dezhou University and Dalian University with a Bachelor and Master degree in Computer Science and Technology. Now, she is working on her PhD in Computer Science at Dalian University of Technology. Her current main research interests are DNA storage and neural network. (Email: zhengyanfen95@gmail.com)