Proceedings of ICMMGH 2026 Symposium: Biomedical Imaging and AI Applications in Neurorehabilitation
DOI: 10.54254/2753-8818/2026.AU31066

Applications of AI Agents in the Entire DNA Sequencing Process: What Can Data Science Do?

Mo Liang

School of Statistics, Renmin University of China, Beijing, China
15804837856@163.com

Abstract. This review examines how AI has changed the DNA sequencing pipeline from first-generation Sanger techniques to fourth-generation platforms over 50 years. AI has a role in high-throughput data processing, multi-omics integration, and variant discovery. Deep learning models improve base calling, structural variant identification, and single-cell data fusion. Systems using big language models to translate natural language experimental goals into optimum primer sets and procedures reduce failure rates, turnaround time, and costs in this research. AI and sequencing are enhancing genetic illness diagnosis, cancer risk prediction, therapy selection, and non-invasive disease monitoring by combining genomic, pathological, and clinical data. Ethics and data security, long-sequence modeling, data storage and transmission, and cost management will be examined. Federated learning for privacy-preserving model training, innovative architectures for long DNA sequences, edge computing-based base calling, cost-effective roll-to-roll fluidic sequencing, and error-corrected cfDNA whole-genome sequencing will also be discussed. Targeted technological innovation, including improved single-molecule accuracy and multi-omics AI analytics; expansion of clinical care, public health, and agriculture applications; and interdisciplinary AI–AI-genomics talent development within strong ethical and regulatory frameworks are our recommendations. These advances make AI-powered sequencing essential to precision medicine and life science research.

Keywords: Artificial Intelligence Agent, DNA sequence, Data science, Precision Medicine

1. Introduction: the evolution of DNA sequencing technologies and the inevitable convergence

Over the last five decades, DNA and RNA sequencing technologies have seen significant advancements, marked by several paradigm shifts. (Figure 1) Protein sequencing and the Edman degradation method were among the first sequencing technologies. In the 1970s, Wu Rui innovated DNA sequencing by the primer extension method, then followed by Sanger's dideoxy chain termination and Maxam-Gilbert's chemical degradation procedures [1]. Second-generation sequencing (NGS) emerged in the twenty-first century, allowing enhanced throughput and extensive initiatives such as the 1000 Genomes Project [2]. The third generation of sequencing introduced single-molecule sequencing and longer read lengths, while the fourth generation aims to significantly improve speed and accuracy [3]. AI integration, shown by DeepSeek, has reduced data analysis duration from days to minutes [4].

DOI: 10.54254/2753-8818/2026.AU31066

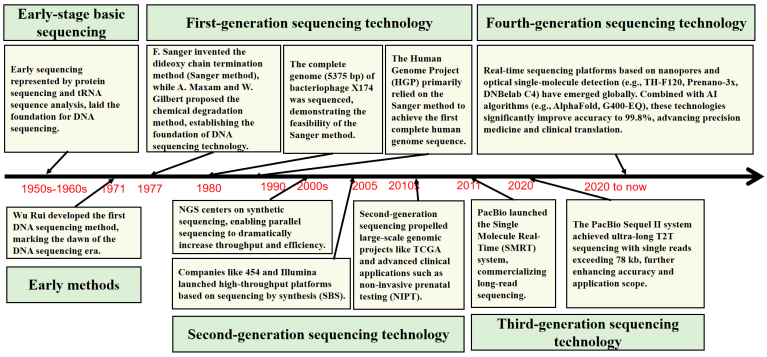

Figure 1. The evolution of DNA sequencing technologies

Concerns over ethics and data security have emerged due to China's Personal Information Protection Law, which mandates data anonymization to safeguard genetic privacy. The integration of CRISPR-Cas9 and sequencing has ignited debates around "designer babies," highlighting the need for global regulation to prevent misuse [5]. Sequencing technologies, varying from expensive, low-throughput methods to contemporary, efficient platforms, are essential in life sciences and precision medicine. Advancements in fourth-generation sequencing and AI-augmented synthetic biology, exemplified as PacBio HiFi technology for artificial genomes, are poised to revolutionize life science research. Integrating innovation with ethics will be essential for sustained advancement and a profound understanding of the fundamental nature of existence. This review explores the transformative role of AI in DNA sequencing, focusing on key applications, challenges, and future directions.

2. AI empowerment in DNA sequencing

2.1. Data processing and analysis

DNA sequencing depends on data processing and analysis. Each sequencing event produces a substantial volume of data, including base signals, quality scores, and sequencing depth [6]. This intricacy often exceeds the capacity of conventional management methods. Conversely, AI-powered systems exhibit superior efficiency in processing these datasets due to their capacity to identify patterns and evaluate data [7]. Multi-omics approaches facilitate the comprehensive analysis of diverse biological data, a task that traditional methods find problematic [8]. Deep learning has significantly enhanced the integration of single-cell data [9]. DNA methylation and chromatin accessibility may be simultaneously evaluated in embryonic cells using techniques such as scCOOL-seq and scNanoCOOL-seq [10,11]. Artificial intelligence systems can identify genomic patterns, regulatory elements, and mutational markers. Utilizing luciferase luminescence data [4], BGI developed an AI model for base calling that incorporates recurrent neural networks to enhance the precision of luminescence signal interpretation [12]. This method effectively reduces errors caused by noise and signal degradation by constantly adjusting categorization criteria throughout sequencing cycles. Moreover, AI has improved variation detection. DeepSV employs a 3D convolutional neural network to detect structural variations over 50 base pairs using PacBio long-read and Illumina short-read data. It surpasses conventional methods by 23%, achieving an F1-score of 0.91 on the GIAB dataset [13].

DOI: 10.54254/2753-8818/2026.AU31066

2.2. Primer design and experimental optimization

On July 30, BGI Genomics announced the launch of PrimeGen, an AI-driven platform designed for primer design in targeted sequencing. PrimeGen is based on a large language model (LLM) that facilitates key functions such as sequence searching, primer design, protocol development, and experiment management [14]. This transition shifts the process from reliance on expertise to a fully automated and efficient workflow. DNA sequencing accuracy is highly reliant on the design of primers that precisely match the target gene sequences. These primers must exhibit exceptional specificity and efficacy, ensuring no interference occurs during multiplex PCR. PrimeGen simplifies this process by transforming complex design phases into an accessible, repeatable solution. Researchers can articulate their objectives in straightforward English, while PrimeGen manages the sequence search, primer arrangement, and protocol development. The Search and Primer Agents work together to identify target sequences and develop optimal primer configurations to ensure comprehensive coverage. The Protocol Agent converts these concepts into operational code compatible with multiple platforms, whereas the Experimental Agent ensures quality assurance during essential stages. This design allows PrimeGen to deliver continuous amplification even with 955 amplicons, reducing primer dimer formation and eradicating the need for repeated experimental cycles. Consequently, it enhances stability, conserves time, and reduces costs.

2.3. Disease diagnosis and prediction

The use of AI in DNA sequencing is advancing healthcare, especially in precision medicine for disease diagnosis and prognosis. BGI's AI model in genetic disease detection integrates contemporary data processing with a stringent quality control system based on extensive validation data, yielding significantly enhanced accuracy in identifying deleterious variants and facilitating early diagnoses. The integration of AI and sequencing technologies in oncology has led to substantial advancements. The MUSK model developed by Stanford University, which combines genetic data with histopathological images and clinical responses, excels in predicting the effectiveness of immunotherapy for lung and gastric cancer. MUSK use deep learning to discern essential factors that affect personalized treatment options, therefore alleviating unnecessary patient strain, both physically and financially. PrimateAI-3D, developed by Illumina's AI Lab in collaboration with prominent universities, forecasts rare illnesses with a deep neural network model that enhances genetic risk evaluations. This AI-based technique aligns with current variant interpretation models and enhances the prediction of genetic risks for rare diseases [15]. Moreover, AI's ability to associate genetic data with environmental and lifestyle factors presents new opportunities for proactive disease risk assessment [16].

3. Challenges and breakthroughs

In this section, we will briefly introduce the challenges and technological breakthroughs encountered in the application of AI combined with whole-genome sequencing of DNA. This discussion is primarily divided into four parts: ethics and security, long sequence processing, data storage and transmission optimization, cost control. (Figure 2)

DOI: 10.54254/2753-8818/2026.AU31066

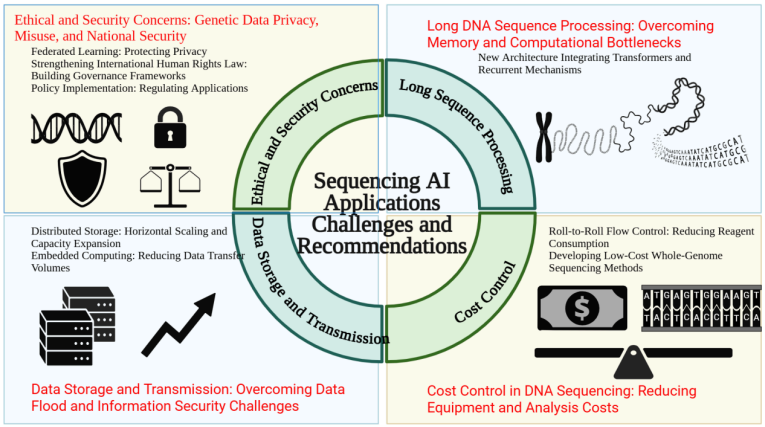

Figure 2. Challenges and breakthroughs of DNA sequencing technologies

3.1. Ethics and security

Life science advances like DNA sequencing raise ethical and safety issues. Data ethics and biosecurity are becoming more important as AI is integrated into sequencing. Human genetic data is sensitive and poses ethical, privacy, and biosafety issues [5]. Genomic data is ethically an individual's genetic heritage. Unauthorized access or disclosure may lead to discrimination in insurance and employment. Gene sequencing and editing have sparked talks about 'designer babies.' Security risks include cross-border transfer or leaking of genomic data, a valuable national resource. Training AI systems on genetic data without authorization may violate privacy. Clinical data is often locked up in institutions, hindering cooperation and data integration.

Recently developed solutions are promising. Federated learning (FL) allows decentralized model training utilizing local data and exchanging just model updates, making it a promising privacy framework [17]. ProCanFDL, a federated deep learning system from the University of Sydney, integrates data from 30 worldwide cohorts, comprising 7,525 human samples and 19,930 mass spectrometry runs. The researchers improved cancer subtype classification accuracy by over 40% by aggregating local characteristics into a global model, akin to centralized training. The model's retraining with additional information identified 16 cancer subgroups [18]. Data isolation, privacy, and model performance are improved by this strategy. FL is growing in proteomics and healthcare, but genomic applications provide unique problems and potential.

International human rights legislation is advancing toward a framework for cross-border data governance that prioritizes human rights while balancing scientific development, individual liberty, and state security [5]. China developed many strategies to achieve this in 2024. The Reference Guidelines for AI Application Scenarios in Health (November 6, 2024) describe AI in healthcare, pharmaceuticals, and public health. The 'Data Element ×' Three-Year Action Plan (2024-2026) aims to improve healthcare data use by building regulated data-sharing frameworks for medical institutions, financial businesses, and elder care sectors.

3.2. Long sequence processing

Certain DNA sequences extend over vast regions, from tens of thousands to millions of base pairs, presenting significant computational challenges for AI systems. Traditional Transformer models exhibit high computational complexity ($O[L^2]$) and memory constraints, rendering them inefficient for the processing of large sequences [19].

DOI: 10.54254/2753-8818/2026.AU31066

New architectures that integrate Transformer and recurrent models have been developed to overcome these limitations. The designs incorporate elements such as linear recurrence, state updates, and gating mechanisms, significantly reducing processing requirements while preserving the models' ability to interpret complex inputs, leading to a phenomenon termed 'recurrent revival' [19]. This technique contests the conventional 'attention-first' paradigm, facilitating more efficient training, accelerated inference, and enhanced hardware compatibility. Shao Bin's team at the Beijing Institute of Technology employed hierarchical DNA encoding to achieve an input length of approximately 100,000 base pairs. The megaDNA model, a multi-scale Transformer incorporating local (16 bp), medium-range (1,024 bp), and long-range (96 kbp) context windows, demonstrates superior performance in predicting essential genes for phage survival and generating synthetic genome segments. The model demonstrates proficiency in identifying genes essential for phage survival and in modeling extensive genomic sequences that are physiologically relevant, thus serving as a crucial resource for research on phage treatment [16]. Sparse attention mechanisms enhance long-sequence processing by focusing on critical segments, leading to improved speed and accuracy [20].

3.3. Data storage and transmission optimization

Data quantities have increased due to DNA sequencing advances, challenging present data storage and transmission methods. Standard whole-genome sequencing (WGS) may generate 150 GB of FASTQ data [21]. Data growth poses three main concerns: growing storage costs, complex data management, and limited scalability. All-flash arrays are expensive yet perform well. While cheaper, tape storage devices have sluggish data retrieval times, making them inappropriate for therapeutic settings. Sequencing data is generally diverse and requires integration with clinical metadata, which standard storage systems sometimes manage inefficiently, slowing data retrieval and complicating system administration. Conventional storage solutions struggle to scale and manage cross-regional data transfers, transmission speed, and data security as sequencing programs grow.

New storage methods and system designs have improved sequencing workflow data management. Dell Technologies' genomic sequencing technology uses a hybrid design to separate computation and storage components for horizontal growth. Dell PowerScale distributed storage and PowerEdge servers create a high-performance infrastructure with simplified operations, superior scalability, easy management, and solid security. PowerScale storage efficiently caches vast sequencing outputs and analytical data for fast I/O and shared analysis. This system scales quickly and improves performance without affecting workloads. Dell offers fast data retrieval, archiving, and catastrophe recovery. The edge-computing module of BGI's DNBSEQ-E25 Flash sequencer uses an 80-MB AI-based base-calling model. This system processes sequencing data in real time and generates variant call files (VCF format), reducing data transfer. Each sequencing cycle takes 75 seconds, allowing full SE50 sequencing in two hours and setting a new speed and throughput standard [22].

3.4. Cost control

In clinical DNA sequencing, managing costs is difficult. Healthcare sequencing is limited by high operating costs and stringent validation. Equipment, reagents, and data analysis contribute to sequencing expenses. Many community and primary healthcare facilities cannot afford them. Whole-genome sequencing costs include equipment, reagents, manpower, and computing resources, which might be difficult for smaller healthcare organizations. The financial burden of expensive

DOI: 10.54254/2753-8818/2026.AU31066

equipment, imported ingredients, and high labor and processing costs limits clinical sequencing technology use.

Recent years have seen notable breakthroughs aimed at reducing costs through the development of low-cost technologies and high-efficiency validation systems. Researchers at the BGI Life Sciences Research Institute, in collaboration with Fuzhou University, reported a pioneering approach in Nature Methods titled “Fast, cost-effective and flexible DNA sequencing by roll-to-roll fluidics.” This method employs roll-to-roll fluidics (r2r-fl) based on planar Couette flow principles to overcome the fluid dynamic limitations inherent to fixed-size flow cells [22]. The system enables continuous reagent flow across flexible biochip substrates, offering high compatibility and cost-efficiency for second-generation sequencing.Experimental results demonstrated that the r2r-fl platform reduces reagent consumption by 85-fold, shortens washing times to less than 2 seconds, and decreases the turnaround time for paired-end 100 bp sequencing from several days to under 12 hours. Moreover, it achieves over 99.9% accuracy and 99.3% sensitivity for single-nucleotide polymorphism (SNP) detection in the human genome, maintains more than 99.9% mapping coverage for Escherichia coli, and exhibits minimal nucleotide substitutions, deletions, or insertions for the SARS-CoV-2 α variant [22]. By simultaneously reducing both cost and time, the r2r-fl technology demonstrates strong scalability and versatility, with promising applications in pathogen detection, cancer diagnostics, and genetic disease analysis.Additionally, cost optimization is being explored in the context of circulating cell-free DNA (cfDNA) sequencing for cancer monitoring [23]. cfDNA analysis is often limited by the scarcity of circulating tumor DNA, low genomic material abundance, and high pre-analytical error rates. Whole-genome sequencing (WGS) can mitigate these issues by integrating signals across the mutation landscape, but its high cost restricts large-scale deployment. To address this, a study led by Dan A. Landau, Bishoy M. Faltas, and Genevieve Boland evaluated a low-cost, error-corrected WGS method for detecting tumor-derived DNA in blood. Their research, which focused on melanoma and urothelial carcinoma patients without matched tumor sequencing, demonstrated that this strategy can effectively assess disease burden and supports the development of non-invasive, affordable cancer tracking through low-cost whole-genome sequencing [23].

4. Recommendations

4.1. Technological innovation

The continued advancement of sequencing technology will transform genomic medicine, demanding targeted innovation driven by AI. Future directions include enhancing AI-assisted error correction for single-molecule sequencing to improve complex genome resolution; reducing sequencing and analytical costs while increasing depth to enable early detection (particularly through liquid biopsy); and developing next-generation technologies such as digital spatial profiling and high-fidelity single-molecule sequencing to address diagnostic blind spots.Equally important is the promotion of data sharing to accelerate biomarker and therapeutic target discovery through the integration of multi-omics datasets with machine learning and AI-based analytics [16]. Furthermore, addressing genomic disparities across different ancestries remains essential to ensure that precision medicine benefits all populations [15].

DOI: 10.54254/2753-8818/2026.AU31066

4.2. Application expansion

Driven by both policy initiatives and rising public awareness, market demand for DNA sequencing technologies continues to expand rapidly [24].On the policy front, newly issued national documents such as the 14th Five-Year Plan for Bioeconomic Development strongly encourage innovation and clinical translation of genetic testing technologies. On the societal front, the growing emphasis on health consciousness has created unprecedented demand for personalized and precision medicine, fueling significant growth in the clinical application of sequencing-based diagnostics. Fields such as early cancer screening, genetic disease detection, and rare disease diagnosis have already shown substantial promise [25].Recent developments have further accelerated this trend. China's first PI3Kα inhibitor (Inaricertib) and AKT inhibitor (Capicertib) were successively approved by the National Medical Products Administration (NMPA), marking a milestone in the evolution of China's breast cancer companion diagnostics market into the era of multi-gene next-generation sequencing (NGS) testing. Updates to the 2025 CSCO Breast Cancer Diagnosis and Treatment Guidelines and the 2025 NCCN Breast Cancer Guidelines have reaffirmed the central role of genetic testing in therapeutic decision-making. Consequently, genetic testing is shifting from an optional to an essential element of cancer care, signaling a new phase in precision oncology.

AI-DNA sequencing integration must go beyond clinical diagnoses. It is necessary and unavoidable to apply to laboratory research, clinical care, public health, and industry [26]. AI-driven metagenomic sequencing technologies improve disease identification and public health epidemic response. The IDbyDNA Explify platform can accurately identify coronaviruses. AI-based systems can speed up pathogen identification during infectious disease outbreaks, enhancing epidemic surveillance and response [25]. AI-driven genomic sequencing spurs industrial and agricultural innovation. Sequencing technologies, genomics, metabolomics, and single-cell omics in agriculture reveal crop genetics and metabolic processes, helping identify trait enhancement genetic markers. An AI-based framework for crop germplasm design was created by Professor Gao Caixia from the Institute of Genetics and Developmental Biology at the Chinese Academy of Sciences and Professor Li Guotian from Huazhong Agricultural University. Biotechnology and AI are used to promote sustainable agricultural growth [24].

4.3. Talent development

President Liu Yi of South-Central University for Nationalities established a 'AI+X' interdisciplinary paradigm that fosters cross-disciplinary credit recognition [27]. Computer science, biology, mathematics, and engineering are needed to train DNA sequencing specialists [28]. Faculty cooperation is encouraged as educational programs adapt to technology and industry demands. The Clinical Medical Education AI Agent and Knowledge Base at Jilin University Third Hospital provides standardized assistance for individualized learning and equal access to high-quality resources. Practical and ethical abilities are equally important. Researchers, engineers, and doctors supervise intensive training programs that teach participants everything from data preparation to AI model deployment, assuring technical and ethical skills [29].

5. Conclusion

To conclude, AI agents have become essential throughout the DNA sequencing process, solving key issues like complex data management, long-sequence analysis, and cost control. Through continuous innovation, broadening cross-disciplinary applications, developing expertise, and strengthening

DOI: 10.54254/2753-8818/2026.AU31066

ethical frameworks, AI-powered genomics will enhance the utility of genetic data and expedite the advancement of precision medicine worldwide.

References

[1] Shendure, J., et al. (2017). DNA sequencing at 40: Past, present and future. Nature, 550(7676), 345–353. https: //doi.org/10.1038/nature24286
[2] Satam, H., et al. (2023). Next-Generation Sequencing Technology: Current Trends. PMC. https: //doi.org/10.3390/
[3] Mardis, E. R., & Wilson, R. K. (2025). Tracing the evolution of sequencing into the era of genomic medicine. Nature Reviews Genetics, 26(10), 719–734. https: //doi.org/10.1038/s41576-025-00712-9
[4] de Ridder, J., et al. (2023). Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing base-calling. Genome Biology, 24, Article 71. https: //doi.org/10.1186/s13059-023-02903-2
[5] Yu, L., et al. (2025). Governance of cross-border genomic data sharing through a human rights approach. Nature Genetics, 57(9), 2090–2098. https: //doi.org/10.1038/s41588-025-02033-3
[6] Napieralski, A., & Nowak, R. (2022). Basecalling Using Joint Raw and Event Nanopore Data Sequence-to-Sequence Processing. Sensors, 22(6), 2275. https: //doi.org/10.3390/s22062275
[7] Alharbi, W. S., & Rashid, M. (2022). A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics, 16(1), 26. https: //doi.org/10.1186/s40246-022-00391-4
[8] Rodriguez-Mier, P., et al. (2025). Unifying multi-sample network inference from prior knowledge and omics data with CORNETO. Nature Machine Intelligence, 7(7), 1168–1186. https: //doi.org/10.1038/s42256-025-01014-2
[9] Wang, X., Yang, S., & Li, H. (2025). scSCC: A swapped contrastive learning-based clustering method for single-cell gene expression data. Quantitative Biology, 13(2), e85. https: //doi.org/10.1007/s40484-025-00348-9
[10] Wang, Y.-N., & Li, J. (2024). Processing pipelines and analytical methods for single-cell DNA methylation sequencing data. Yi Chuan (Hereditas), 46(10), 807-819. https: //doi.org/10.16288/j.yczz.24-154
[11] Frömel, R., et al. (2025). Design principles of cell-state-specific enhancers in hematopoiesis. Cell, 188(12), 3202–3218.e21. https: //doi.org/10.1016/j.cell.2025.05.023
[12] Zhang, T. (2024). Nanopore sequencing: flourishing in its teenage years. Genes & Genomics.
[13] Wang, Y., et al. (2025). Genomic Touchstone: Benchmarking genomic language models in the context of the central dogma. bioRxiv, 2025.06.25.661622. https: //doi.org/10.1101/2025.06.25.661622
[14] Wang, Y., et al. (2025). Accelerating primer design for amplicon sequencing using large language model-powered agents. Nature Biomedical Engineering. https: //doi.org/10.1038/s41551-025-01225-8
[15] Gao, H., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science, 380(6648), eabn8153. https: //doi.org/10.1126/science.abn8153
[16] Shao, B., & Yan, J. (2024). A long-context language model for deciphering and generating bacteriophage genomes. Nature Communications, 15(1), 9392. https: //doi.org/10.1038/s41467-024-47321-0
[17] Calvino, G., et al. (2024). Federated learning: Breaking down barriers in global genomic research. Genes, 15(12), 1650. https: //doi.org/10.3390/genes15121650
[18] Cai, Z., et al. (2025). Federated deep learning enables cancer subtyping by proteomics. Cancer Discovery, 15(9), 1803–1818. https: //doi.org/10.1158/2159-8290.CD-24-0685
[19] Tiezzi, M., et al. (2025). Back to recurrent processing at the crossroad of transformers and state-space models. Nature Machine Intelligence, 7(5), 678–688. https: //doi.org/10.1038/s42256-025-00941-2
[20] Nguyen, E., et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science, 386(6723), eado9336. https: //doi.org/10.1126/science.ado9336
[21] Halldórsson, B. V., et al. (2022). The sequences of 150, 119 genomes in the UK Biobank. Nature. https: //doi.org/10.1038/s41586-022-04965-x
[22] Qin, Y., et al. (2025). Fast, cost-effective and flexible DNA sequencing by roll-to-roll fluidics. Nature Methods, 22(8), 1662–1668. https: //doi.org/10.1038/s41592-025-02503-1
[23] Ochana, B. L., et al. (2024). Time is encoded by methylation changes at clustered CpG sites. bioRxiv. https: //doi.org/10.1101/2024.05.15.593081
[24] Li, G., et al. (2025). Integrated biotechnological and AI innovations for crop improvement. Nature, 643(8073), 925–937. https: //doi.org/10.1038/s41586-025-07573-9
[25] de Almeida, B. P., et al. (2025). A multimodal conversational agent for DNA, RNA and protein tasks. Nature Machine Intelligence, 7(6), 928–941. https: //doi.org/10.1038/s42256-025-00992-5
[26] Yao, W. (2025). An Empirical Study Based on AI+X Program in China. Heliyon / or Journal (mixed-methods)
[27] Wu, Z. (2024). Teaching Practices for the Cultivation of “AI + X” Composite Talents in Higher Education.

DOI: 10.54254/2753-8818/2026.AU31066

[28] Jing, Y. (2024). The Construction of Artificial Intelligence Major (AI+X Direction) in Higher Education. SciOpen
[29] Yang, R., Pang, H., Zang, H., Zhang, R., Zhang, Z., Li, X., & Zhang, L. (2025). Artificial intelligence-enabled DNA computing: Exploring new frontiers in bioinformatics. Daxue Huaxue (University Chemistry), 40(9), 107-117. https: //doi.org/10.12461/PKU.DXHX202412135