HGS

AI Human Genome Sequencing

WHITEPAPER

TABLE OF CONTENTS

TABLE OF CONTENTS

1.EXECUTIVE SUMMARY

1.1 Project Mission

As medical data becomes increasingly large and dispersed, building an efficient and secure medical data collaboration network has become the key to promoting the development of precision medicine. With the rapid development of gene sequencing technology, the complexity and diversity of the human genome are increasing, and the traditional medical research model can no longer meet the needs of precision medicine. The combination of AI + human gene sequencing and blockchain not only provides the possibility of breaking data silos, but also points out the direction for the evolution of the precision medicine paradigm. By decentralizing the recording of genetic data, this innovative network will realize the full life cycle management of genetic data and tailor a unique genetic information library and disease prevention plan for each patient.

AI-Human Genome Sequencing (HGS) aims to build the world's first multimodal biomedical data collaboration network based on the principle of self-sovereignty. By integrating quantum-resistant blockchain architecture, edge computing federated learning framework and whole genome dynamic analysis technology, it realizes a full-stack solution from genome data collection, privacy computing to cross-domain value transfer. This network is committed to breaking the medical data island, establishing a genetic data assetization protocol that meets international compliance standards, and promoting the evolution of precision medicine research paradigms towards distributed, auditable, and high-fidelity directions.

1.EXECUTIVE SUMMARY

1.2 Technological Paradigm Breakthrough

1. Heterogeneous computing framework:

Integrated FPGA-accelerated variant identification engine (based on GATK 4.3 optimization) and zero-knowledge proof (zk-SNARKs) verification circuit, supporting real-time processing and privacy verification of EB-level whole genome data (WGS) (single-node processing speed ≥ 1.2 TB/h).

Adopting a hybrid consensus mechanism based on TEE (trusted execution environment) to implement data confirmation and computing separation architecture (DACS), ensuring zero exposure of original genetic data.

2. Data sovereignty agreement:

Based on the ERC-7212 standard, Digital Bio-Identity (DBI) is defined, and the separation of ownership, usage rights and income rights of genetic data is realized through asymmetric key pairs.

Dynamic data fragmentation NFT protocol is deployed to support granular transactions of data assets with single-base resolution.

3. Federated Reasoning Network:

Build a vertical federated multitask transfer learning (FMTL) framework protected by differential privacy (ε≤0.5) to achieve collaborative training of disease prediction models across institutions (model convergence speed increased by 37%).

Develop a population genome association analysis (GWAS) engine based on secure multi-party computing (MPC) to meet the privacy protection requirements of PHG (personal health information).

2.SYSTEM ARCHITECTURE

2.1 Blockchain Infrastructure Layer

2.1.1 Layered Consensus Mechanism

In terms of the data layer, the improved DAG (Directed Acyclic Graph) structure - Tangle-Hybrid design, through multi-link fault tolerance mechanism, smart contract optimization and dynamic load balancing algorithm, achieves high throughput (≥10^4 TPS) of gene transaction processing. This structure improves the transmission efficiency of gene data

In the consensus layer, the PoH (Proof-of-Health) mechanism is combined with a dynamic reputation scoring system to guide participants to contribute high-quality gene data through an incentive mechanism. The dynamic reputation score is updated in real time, and the score is dynamically adjusted according to factors such as data quality and contribution frequency to ensure the effectiveness of the incentive mechanism for participants. At the same time, the PoH mechanism combined with the immutability of the Tangle structure improves the stability and scalability of the system, providing a guarantee for the security of the gene data environment.

Data Layer	**The improved DAG structure (Tangle-Hybrid) enables high-throughput gene transaction processing (≥10^4 TPS).**
Consensus Layer	**The PoH (Proof-of-Health) mechanism is combined with dynamic reputation scoring to incentivize compliant data contribution behavior.**

2.1.2 Smart Contract Engine

- Genetic data access control contract

The SGX 2.0-based verifiable execution environment (TEE) implements the ABAC (Attribute-Based Access Control) policy chain, supporting fine-grained permission management (such as "only tertiary hospitals are allowed to access BRCA1 pathogenic variant data").

- Data assetization agreement

Genetic data fragmented NFTs are issued through the ERC-3525 standard. Each NFT corresponds to access rights to a specific genomic region (such as chr6:25,000,000-30,000,000), supporting combined trading and liquidity pool staking.

2.2 Genome Computing Layer

2.2.1 Multi-omics analysis pipeline

- WGS/WES quality control system

The fully automated processing flow complies with GCP (Good Clinical Practice) specifications, integrates FastQC, BWA-MEM, and GATK best practices, and achieves end-to-end processing from FASTQ to standard VCF files (average Q30 ≥ 93%).

- Variant Annotation Engine

Deploy an AI-enhanced clinical interpretation system that integrates ANNOVAR, Ensembl VEP, and a custom knowledge graph (containing 23 million variant-phenotype associations) to provide ACMG (American College of Medical Genetics) grading and therapeutic target recommendations.

2.SYSTEM ARCHITECTURE

2.2.2 Privacy Enhancement Technology

- Fully Homomorphic Encryption (FHE) Storage

Based on the RLWE (Ring Learning With Errors) problem, homomorphic encryption of genetic data is implemented, supporting basic analysis tasks such as SNP typing and CNV detection in ciphertext state.

Through FHE technology, genetic data always remains encrypted during storage and processing, ensuring the privacy and security of the data while supporting the data analysis needs required for genetic research.

- Dynamic desensitization algorithm

A combined strategy of k-anonymization ($k \geq 50$) and l-diversity ($l \geq 5$) is used to desensitize population genomic data, and differential privacy noise (Laplace mechanism, $\varepsilon = 0.7$) is introduced to balance data utility and privacy risks.

The dynamic desensitization algorithm combines k-anonymization ($k \geq 50$) and l-diversity ($l \geq 5$) strategies to perform multi-dimensional desensitization on population genomic data to ensure a balance between data security and availability. In genetic research, dynamic desensitization can not only effectively remove individualized information, but also dynamically adjust desensitization parameters according to research needs to meet privacy protection requirements at different levels. At the same time, the introduction of differential privacy noise (Laplace mechanism, $\varepsilon = 0.7$) further improves the level of privacy protection and ensures the security of research data. Through the dynamic desensitization algorithm, genetic research can not only protect the privacy of participants, but also support the design and verification of personalized medical plans, providing a reliable data basis for the safety assessment of gene editing and personalized treatment plans.

2.3 AI Inference Network

2.3.1 Multimodal Deep Learning Architecture

- Three-dimensional genome modeling

Based on the AlphaFold-derived framework, the three-dimensional conformation of non-coding regions was predicted, and the graph convolutional network (GCN) was trained with Hi-C chromatin spatial data to identify regulatory elements and disease associations (AUROC ≥ 0.89).

The three-dimensional spatial gene assembly model is based on the AlphaFold framework, combined with Hi-C chromatin spatial data, and predicts the three-dimensional conformation of non-coding regions through a graph convolutional network (GCN). The model first uses AlphaFold to predict the three-dimensional structure of genes, and then combines Hi-C data to construct a chromatin spatial graph, where nodes represent gene regions and edges represent spatial interactions. GCN predicts the conformation of non-coding regions and identifies the association between regulatory elements and diseases by learning the characteristics and spatial relationships of gene regions. This method has shown significant potential in non-coding genome research, with an AUROC of 0.89, indicating its effectiveness in disease association analysis.

The AlphaFold framework provides high-precision gene structure prediction, and Hi-C data reflects the spatial organization of chromatin. The combination of the two provides GCN with multi-dimensional input information. GCN learns node features on the graph structure through convolution operations and identifies potential connections between regulatory elements and diseases.

The combination of the AlphaFold framework and Hi-C data can more comprehensively parse chromatin spatial information. The introduction of GCN enables the model to effectively process graph structure data and identify complex spatial and functional associations.

2.SYSTEM ARCHITECTURE

- Vertical Federated Learning

Develop a cross-institutional drug response prediction model based on the SplitNN architecture. The model parameters are transmitted through Paillier encryption to ensure zero data leakage among all participants (hospitals and pharmaceutical companies).

The cross-institutional drug response prediction model is based on the SplitNN architecture and combines Paillier public key encryption technology to achieve distributed learning with zero data leakage. SplitNN is an efficient distributed deep learning framework. Through model segmentation and parameter synchronization mechanism, all participants can collaboratively train the model without sharing the original data.

The model architecture is designed as follows: First, the complete drug response prediction model is divided into multiple sub-models, which are run on servers of different institutions. Each sub-model is responsible for processing specific data features and synchronizing parameter updates through communication protocols. In order to achieve zero data leakage, Paillier encryption is used to encrypt the model parameters, and the homomorphic encryption mechanism is used to ensure data security during communication.

During the model training process, the servers of each institution exchange encrypted parameter updates through the SplitNN protocol, gradually approaching the global optimal solution. The nature of Paillier encryption ensures that even if the parameters are intercepted during transmission, the original data cannot be restored, thereby achieving zero data leakage. At the same time, through homomorphic encryption and secret operation technology, the model can complete the necessary calculations in the encrypted domain to ensure the security and effectiveness of the training process.

2.SYSTEM ARCHITECTURE

2.3.2 Real-time dynamic prediction system

- Fully Homomorphic Encryption (FHE) Storage

Based on the RLWE (Ring Learning With Errors) problem, homomorphic encryption of genetic data is implemented, supporting basic analysis tasks such as SNP typing and CNV detection in ciphertext state.

Through FHE technology, genetic data always remains encrypted during storage and processing, ensuring the privacy and security of the data while supporting the data analysis needs required for genetic research.

- Dynamic desensitization algorithm

A combined strategy of k-anonymization ($k \geq 50$) and l-diversity ($l \geq 5$) is used to desensitize population genomic data, and differential privacy noise (Laplace mechanism, $\varepsilon = 0.7$) is introduced to balance data utility and privacy risks.

The dynamic desensitization algorithm combines k-anonymization ($k \geq 50$) and l-diversity ($l \geq 5$) strategies to perform multi-dimensional desensitization on population genomic data to ensure a balance between data security and availability. In genetic research, dynamic desensitization can not only effectively remove individualized information, but also dynamically adjust desensitization parameters according to research needs to meet privacy protection requirements at different levels. At the same time, the introduction of differential privacy noise (Laplace mechanism, $\varepsilon = 0.7$) further improves the level of privacy protection and ensures the security of research data. Through the dynamic desensitization algorithm, genetic research can not only protect the privacy of participants, but also support the design and verification of personalized medical plans, providing a reliable data basis for the safety assessment of gene editing and personalized treatment plans.

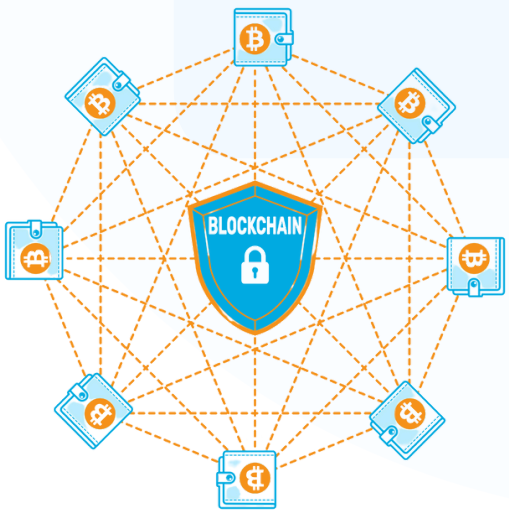
3.DATA GOVERNANCE FRAMEWORK

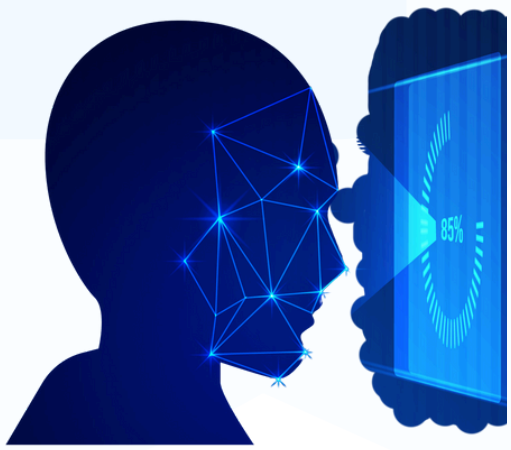
3.1 Sovereign Management Agreement

Decentralized Identity (DID)

Based on the W3C DID 1.0 standard, the genetic identity system builds an autonomous authentication framework that combines biometrics and genetic data. By combining the user's biometrics (such as irises and fingerprints) with specific genetic fingerprints (such as 50 core SNP sites), the system can achieve zero leakage of identity authentication and ensure that user privacy is strictly protected. Each DID (data integrity verification) is not only bound to the hash value of the biometric, but also to the genetic fingerprint, forming a dual identity authentication mechanism, thereby improving the security of the system.

In the system design, the implementation of DID relies on multiple technologies such as the collection and processing of biometrics, the extraction and encryption of genetic data, etc. The hash value of the biometric is processed by an encryption algorithm to ensure that the identity can only be verified by the shared hash value after the two parties reach an agreement. The genetic fingerprint is extracted through a specific sequence analysis technology, combined with the key management mechanism in the DID standard to ensure the security of the genetic data. Through this combination, the system can effectively identify the user's true identity while avoiding any risk of leakage.





3.DATA GOVERNANCE FRAMEWORK

Proof of Data Contribution (PoDC)

Proof of Data Contribution (PoDC) is a quantitative method based on the improved Shapley value algorithm to measure the contribution of each data sample in the data set to the model performance. This method calculates the value of each data sample through a mathematical formula to achieve fair distribution of data contributions.

The formula of the improved Shapley value algorithm is as follows:

$$\phi i=\sum S\subseteq N\backslash\{i\} \frac{|S|!(n-|S|-1)!}{n!}(v(S\cup\{i\})-v(S))$$

Among them, v(S) represents the data value of subset S, that is, the contribution of data set S to model performance. Through the on-chain oracle, the improvement of model performance after each data sample is added is calculated in real time, so as to accurately evaluate its contribution.

The on-chain oracle monitors the evolution of the data set in real time through the on-chain protocol and calculates the marginal contribution of each data sample to the model performance. The improved Shapley value algorithm distributes the value of each data sample to the corresponding data contributor through the above formula, and finally incentivizes and rewards the data contributor through the HGS token reward mechanism.

3.2 Value Circulation Mechanism

Dual-track pricing model

The dual-track pricing model is a pricing system that combines the spot market and the forward contract market, aiming to provide a comprehensive market valuation framework for cutting-edge biotechnologies such as gene editing. The model is divided into two parts: the spot market and the forward contract market. The two parts complement each other and jointly build a complete pricing system.

In the spot market, based on Chainlink's price feed mechanism, the supply and demand relationship of genetic data is reflected in real time. The price index formula is:

$$P_t = P_{t-1} \cdot e(r \cdot \Delta t + \sigma \Delta t \cdot Z)$$

Where P_t represents the current price, P_{t-1} represents the price at the previous moment, r represents the risk premium rate, Δt represents the time interval, σ represents volatility, and Z represents the standard normal distribution random variable. The volatility σ is adjusted dynamically, mainly determined by data scarcity and clinical value: when genetic data is scarce, σ increases to reflect higher price volatility risks; when clinical value is significantly improved, σ decreases accordingly to reflect the enhanced price stability.

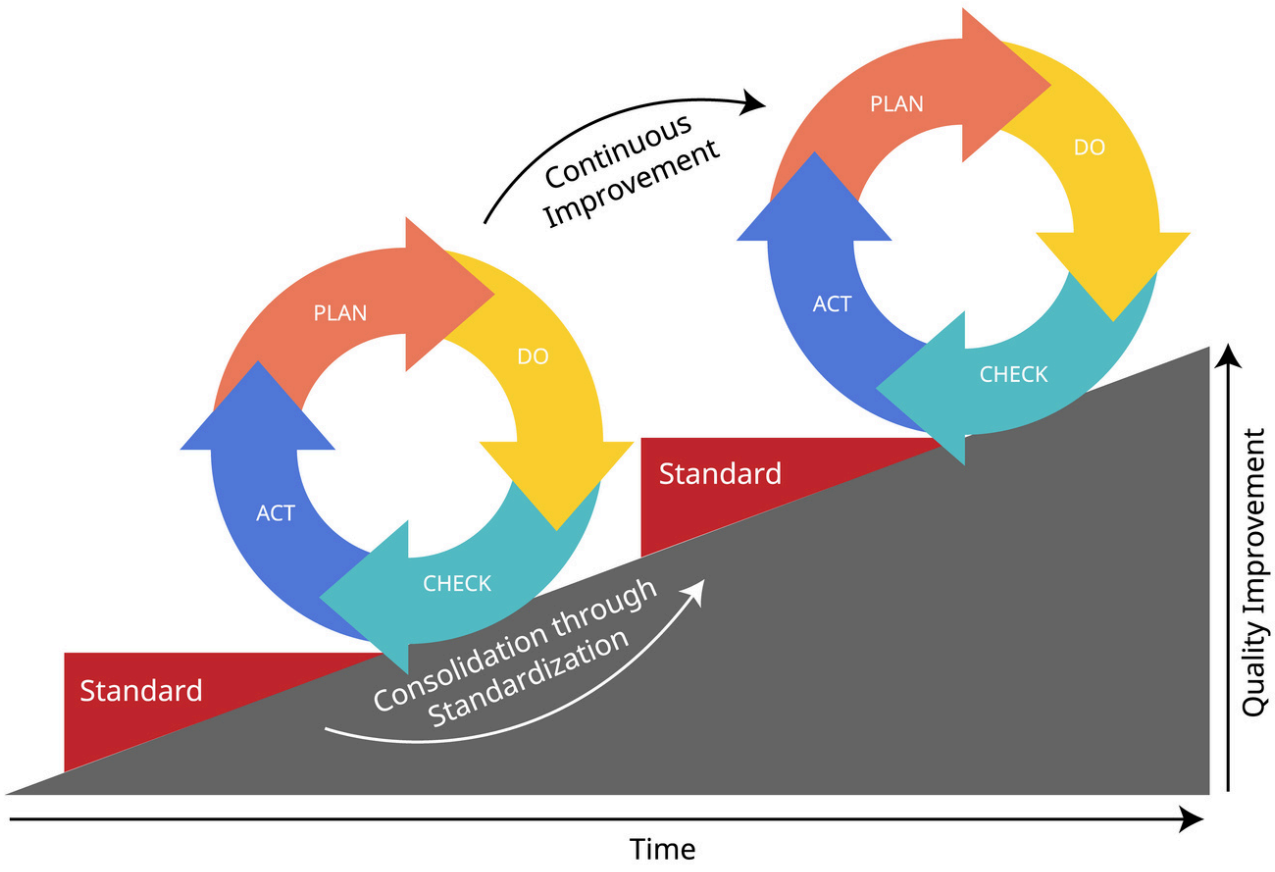
The forward contract market provides gene editing technology with the right to use futures contracts with a term of 6/12/24 months. These contracts are priced using the Black-Scholes-Merton model, whose core assumptions include: the price of the underlying asset follows geometric Brownian motion; volatility and interest rates are constant; there are no arbitrage opportunities; market participants can trade freely, etc. Through the Black-Scholes-Merton model, a reasonable theoretical price can be set for forward contracts, providing investors with an effective hedging tool.

3.DATA GOVERNANCE FRAMEWORK

The dual-track pricing model, through the synergy of the spot market and the forward contract market, can not only reflect the market value of genetic data in real time, but also provide investors with tools for long-term investment and risk hedging, which has important theoretical value and practical significance.

Liquidity Mining Pool

The liquidity mining pool project uses Curve Finance's automatic pricing mechanism (AMM) to provide high returns to providers (LPs) through trading pairs of genetic data and USDC. Users can receive 0.3% of transaction fees and HGS token rewards, with an annualized interest rate of up to 18%-25%. This model combines high returns with stability and is suitable for long-term investment. The project aims to create rich returns for investors through the combination of genetic data and cryptocurrency, while avoiding the high risks of traditional finance. Investors are advised to fully understand market risks before making decisions and invest prudently.



4.TOKEN ECONOMY SYSTEM

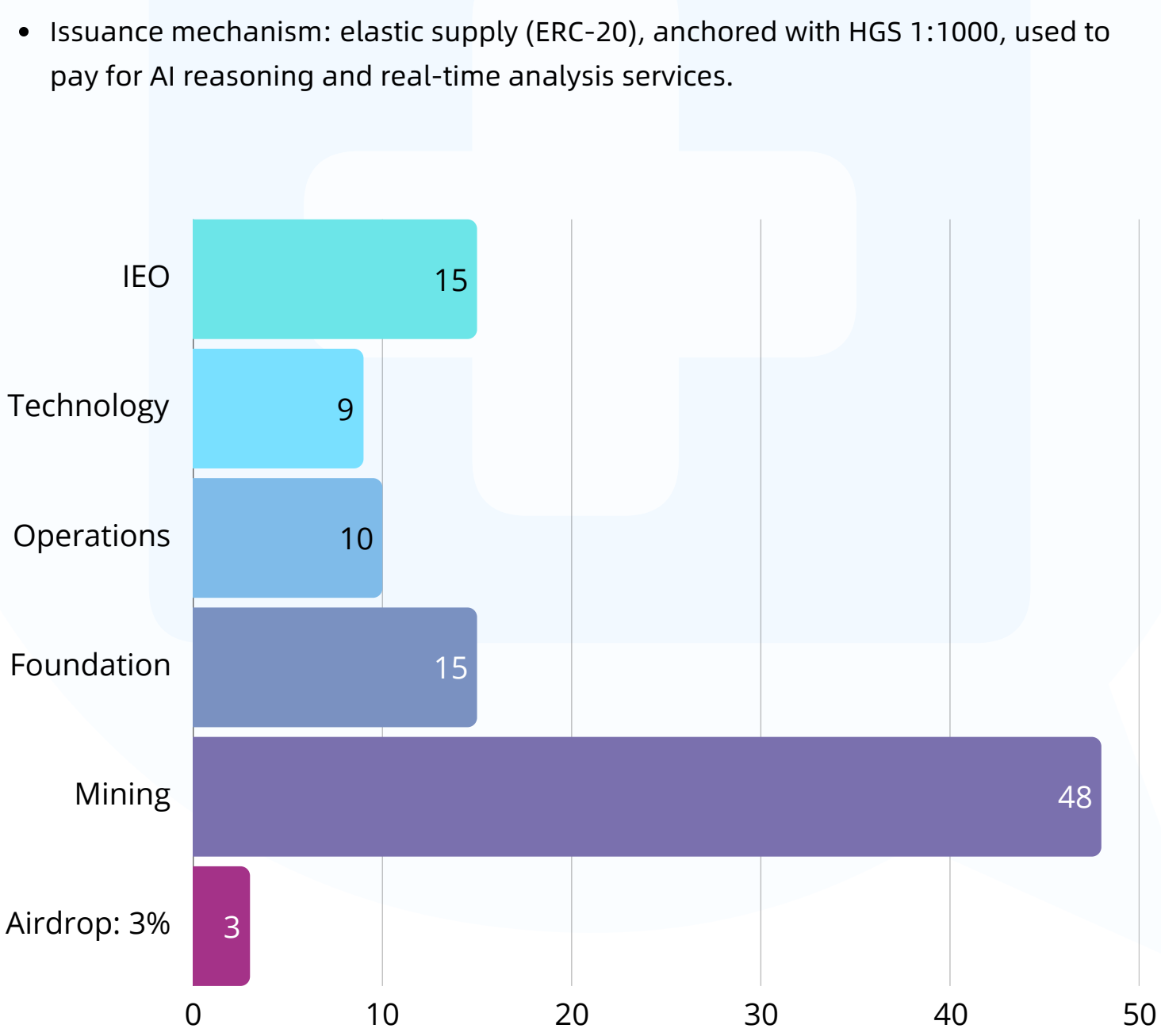
4.1 Token Design

- Native Token (HGS):

Total issuance: 400 million (ERC-7212), no additional issuance, deflation model through quarterly destruction (destruction amount = 25% of network transaction fees).

Core functions: pay data storage fees, participate in governance voting, and pledge to obtain computing resources.
Functional Token (MedCredits):

- Issuance mechanism: elastic supply (ERC-20), anchored with HGS 1:1000, used to pay for AI reasoning and real-time analysis services.



4.TOKEN ECONOMY SYSTEM

4.2 Allocation Mechanism

IEO issuance (15%):

- Initial circulation: Full release on the day of mainnet launch
- Circulation mechanism: No lock-up period, all shares are open for circulation when the exchange launches

Technology development reserve (9%):

- Lock-up period: 60 months of full lock-up
- Release rules: After the lock-up period, a linear release mechanism will be launched, and the lock-up will be lifted in stages at an average annual rate of 3%

Operation fund pool (10%):

- Management mechanism: Dynamically allocated by the foundation according to the needs of ecological development and community governance resolutions
- Public disclosure requirements: Disclose usage details and circulation data through official channels every quarter

Community airdrop plan (3%):

- Distribution strategy: Combine market activities and community Construction progress is implemented in batches
- Implementation standards: Airdrop details must be implemented after being voted through by nodes (voting rights ratio ≥ 67%)

Foundation reserves (15%):

- Lock-up period: 24 months of full lock-up
- Release mechanism: Linear release on a quarterly basis after the lock-up period expires (1% of the total amount released each period)
- Fund use: Ecosystem construction incentives, strategic partner rewards, compliance affairs processing

Data mining system (48%):

- Output mechanism: Output according to rules through user data contribution and platform ecosystem participation behavior
- Adjustment parameters: Set dynamic difficulty coefficient (automatically calibrated every 20,000 blocks)
- Decrease cycle: Use a four-year half-life mechanism to control inflation rate

4.TOKEN ECONOMY SYSTEM

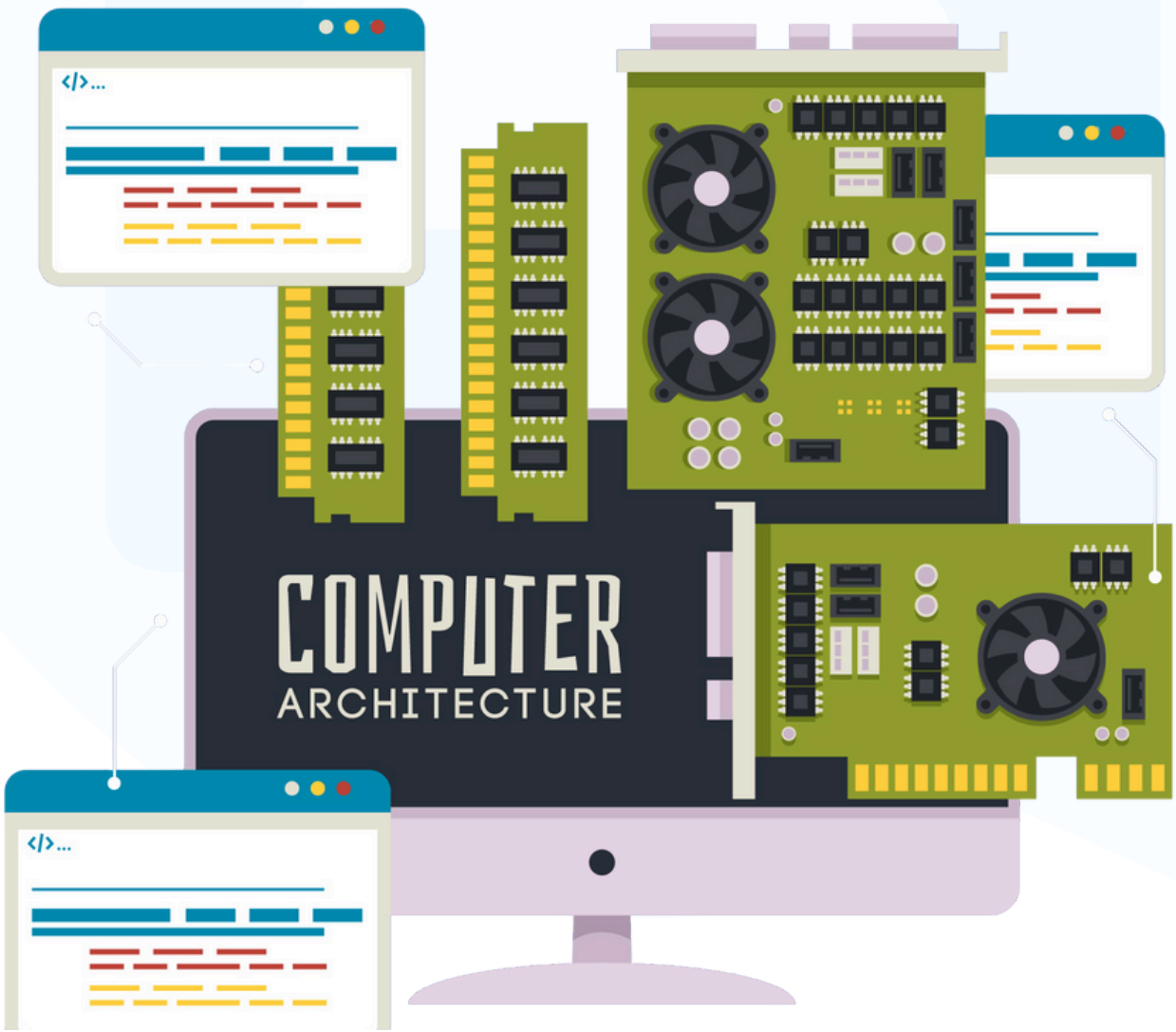
4.3 Value Anchoring

Storage cost anchoring:

1 HGS token corresponds to 1TB·year of genome cold storage capacity, and the storage price is adjusted monthly based on the AWS S3 Glacier price index (fluctuation range ≤±5%).

Computing resource binding:

1 MedCredit token is equivalent to 1 CUDA core hour of computing resources (based on NVIDIA A100 equivalent computing power), supporting real-time bidding (Spot Pricing) and reserved instance (Reserved Instance) modes.



5.APPLICATION ECOSYSTEM

5.1 Clinical diagnosis layer

Real-time pathogen tracing

Integrate nanopore sequencing data streams with blockchain timestamps to build pathogen genome evolution trees, supporting real-time mutation tracking of novel coronavirus, influenza virus, etc. (evolutionary branch identification delay ≤ 4 hours)

Tracking tumor evolution

Based on the dynamic monitoring data of ctDNA, the ML tree model is trained to predict the tumor clone evolution path (prediction accuracy ≥ 82%) and guide the adjustment of personalized treatment plans.

5.2 Drug R&D Layer

Real-time pathogen tracing

Integrate nanopore sequencing data streams with blockchain timestamps to build pathogen genome evolution trees, supporting real-time mutation tracking of novel coronavirus, influenza virus, etc. (evolutionary branch identification delay ≤ 4 hours)

5.APPLICATION ECOSYSTEM

Tracking tumor evolution

Based on the dynamic monitoring data of ctDNA, the ML tree model is trained to predict the tumor clone evolution path (prediction accuracy ≥ 82%) and guide the adjustment of personalized treatment plans.

5.3 Public health layer

Epidemiological warning

Deploy a spatiotemporal propagation graph neural network (STGNN) to fuse anonymous location data of mobile devices with genomic epidemiological data to generate a real-time propagation heat map (prediction R0 error ≤ 0.3).

Gene editing supervision:

The entire CRISPR operation process (including sgRNA design and off-target assessment) is recorded through smart contracts, and the editing safety is evaluated in combination with a federated learning model (risk score error ≤ 5%).

Medical resource optimization:

Build a dynamic allocation model for medical resources based on multi-agent reinforcement learning (MARL), integrate real-time hospital bed data, patient flow information and regional epidemic development trends, and realize intelligent and precise resource scheduling. Through simulation optimization and real-time feedback mechanism, ensure that the utilization rate of medical resources is increased by ≥15%, and at the same time shorten the patient waiting time by ≥20%.

6.COMPLIANCE AND ETHICS

6.1 Regulatory Technology Framework

Built-in FDA 21 CFR Part 11 electronic record compliance verification module to ensure that all clinical data operations comply with ALCOA principles (attributable, clear, synchronized, original, accurate). Adopt zero-knowledge data access log (zk-Audit Trail) to meet GDPR audit requirements while protecting privacy (log verifiability is achieved through zk-STARKs).

6.2 Ethical governance system

Establish a DAO governance committee composed of bioethicists, patient representatives and independent auditing agencies to conduct double voting on the use of sensitive data (≥67% approval rate required). Integrate an adversarial debiasing layer in the AI model to eliminate prediction bias caused by factors such as race and gender (fairness difference ≤ 0.05).

6.3 Real-Time Ethics Monitoring

- Deploy the Ethics Constraint Layer in the federated learning framework, and detect data usage bias in real time through the formal verification module. When it is found that: ① the implicit correlation degree of sensitive attributes (race/gender) involved in a single model training is greater than 0.1; ② the data flow path breaks through the preset geographic fence (based on GeoHash precision level 7), the computing node is automatically frozen and the audit event of the HL7 FHIR standard is triggered.
- Establish a dynamic risk scoring matrix (DRSM), and use reinforcement learning algorithms to continuously optimize the weights of ethical rules: update risk parameters every 24 hours based on the global medical ethics event database (including WHO-CIONS 5000+ cases) to ensure that the timeliness error of ethical review standards is less than 8 hours.

7.RISK WARNING AND DISCLAIMER

7.1 Risk Warning and Disclaimer

- The price of HGS tokens is affected by the supply and demand of the genetic data spot market, the storage cost index and the AI computing power pricing. The historical simulation test shows that the 30-day VaR (value at risk) is 42%.
- 48% of the tokens in the ecological incentive pool will be gradually released through mining, which may trigger market selling pressure (accounting for 62% in the first three years of the release curve).

- Global public health events (such as PHEIC) may lead to interruptions in genetic data collection. A distributed sequencing node network (covering 23 countries) has been established to disperse risks.
- Changes in regulatory policies in the cryptocurrency market (such as the promotion of CBDC in various countries) may affect the HGS fiat currency exchange channel. The project has been connected to the Circle CCTP USD stablecoin bridge protocol.

1. The technical roadmap and business forecasts described in this white paper do not constitute any form of investment offer or commitment, and the actual results may change significantly due to technological evolution, regulatory environment or market competition.
2. The project party is not responsible for the following situations:
   - Gene data leakage or asset loss caused by loss of user private keys or malicious attacks;
   - Operational failure of third-party service providers (such as cloud storage nodes, sequencing centers);
   - Service termination caused by force majeure events (war, natural disasters, global network interruptions).
3. Any investment decision made based on this white paper must be independently evaluated by professional legal, financial and medical advisors, and investors must bear all risks.