



OPEN ACCESS

EDITED BY

Ingrid Vasilii Feltes,
University of Miami, United States

REVIEWED BY

Mustaqeem Khan,
Mohamed bin Zayed University of Artificial
Intelligence (MBZUAI), United Arab Emirates
Divyashree D.,
CMR University, India

*CORRESPONDENCE

Yu Wang,
✉ yosean.wang@cityu.edu.hk
George M. Church,
✉ gchurch@genetics.med.harvard.edu

*Lead Contact

RECEIVED 19 May 2025
REVISED 25 December 2025
ACCEPTED 29 December 2025
PUBLISHED 02 February 2026

CITATION

Wang Y, Fan J, Bao Z, Zhang S, Liu L, Yang M and
Church GM (2026) Cryptographic open science:
enabling secure and incentivized biomedical
data sharing with web 3.0 technologies to
overcome the open science dilemma.
Front. Blockchain 8:1631217.
doi: 10.3389/fbloc.2025.1631217

COPYRIGHT

© 2026 Wang, Fan, Bao, Zhang, Liu, Yang and
Church. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is permitted
which does not comply with these terms.

Cryptographic open science: enabling secure and incentivized biomedical data sharing with web 3.0 technologies to overcome the open science dilemma

Yu Wang^{1,2,3,4*†}, Junfeng Fan^{2,5}, Zhiwei Bao^{3,6}, Sheng Zhang³,
Liang Liu², Mengsu Yang^{1,4} and George M. Church^{7*}

¹Institute of Digital Medicine, City University of Hong Kong, Hong Kong SAR, China, ²DeSci Sino, Hong Kong, Hong Kong SAR, China, ³Zhiyu DAO, KanDaoShanWei Technology Inc., Shanghai, China, ⁴Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China, ⁵Open Security Research, Shenzhen, Guangdong, China, ⁶College of Biomedical Engineering and Instrument Science, Ministry of Education Key Laboratory of Biomedical Engineering, Zhejiang University, Hangzhou, China, ⁷Department of Genetics, Harvard Medical School, Boston, MA, United States

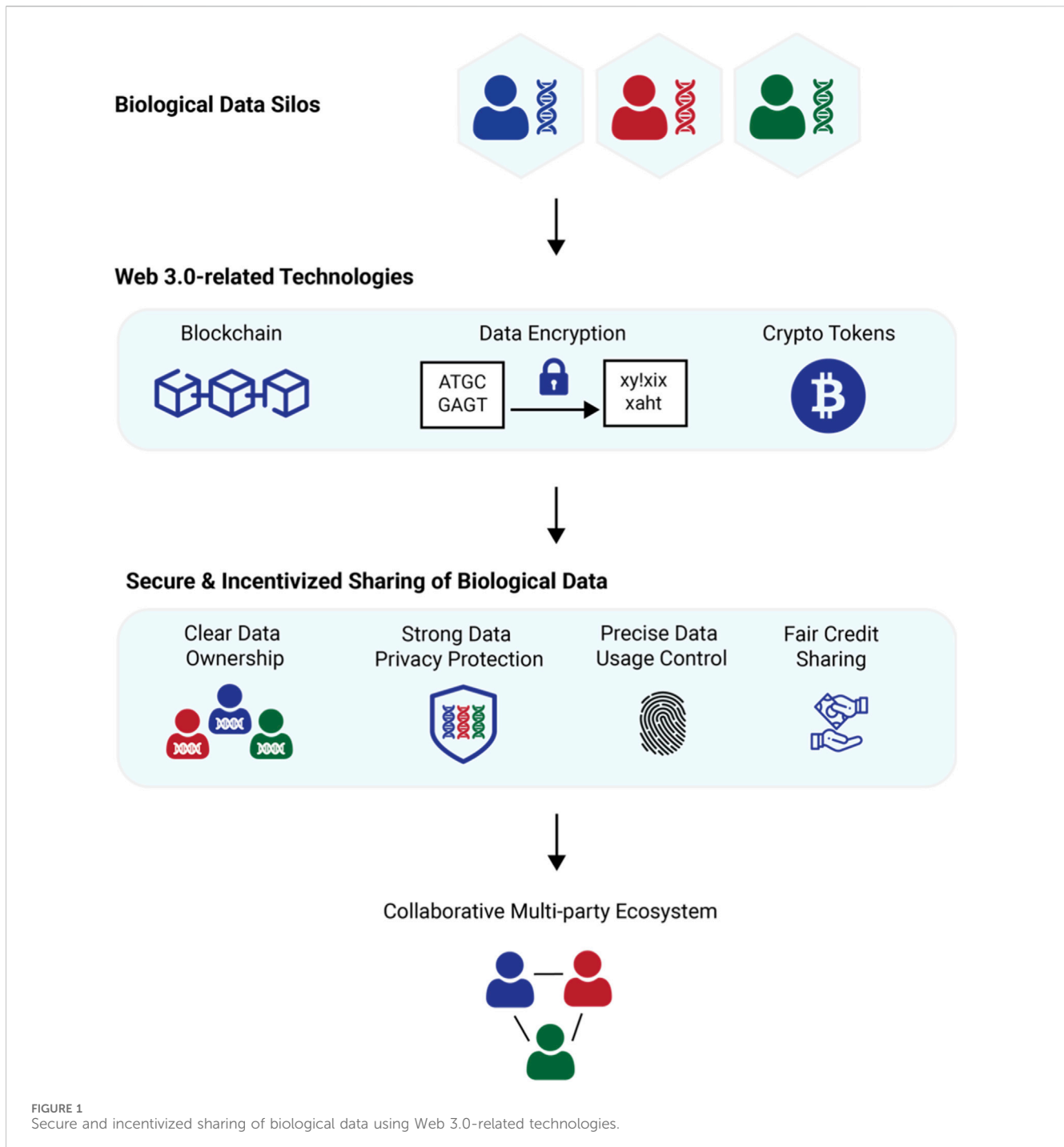
Biomedical data underpins scientific discovery, and open science offers the potential to accelerate innovation through the unrestricted sharing of knowledge, methodologies, and datasets. However, the open science dilemma persists, as researchers hesitate to share data due to privacy concerns, intellectual property risks, and lack of recognition. Stringent data privacy regulations further compound these challenges, limiting data sharing. To address these barriers, we propose the Cryptographic Open Science (COS) framework, integrating advanced technologies for secure, privacy-preserving, and incentivized data sharing. Blockchain technology provides immutable records of data ownership and usage, enhancing transparency and trust, while smart contracts automate access controls and enforce compliance. However, blockchain alone does not prevent loss of control over plaintext data once released. COS incorporates Fully Homomorphic Encryption (FHE) to allow computations on encrypted data, ensuring end-to-end confidentiality and maintaining full ownership control. Recognizing that privacy alone does not incentivize sharing, COS introduces a crypto token-based system to create a market-driven flywheel. This system rewards contributors and aligns stakeholder interests, promoting active data sharing. By integrating blockchain, FHE, and token-based incentives, COS bridges the gap between the ideals of open science and the practical concerns of data providers, accelerating progress in fields like precision medicine and genomics.

KEYWORDS

biomedical data, blockchain, cryptocurrency, fully homomorphic encryption, open science, web 3.0 technologies

1 Introduction

In today's digital age, especially with the rapid growth of artificial intelligence, data has become a vital resource driving advancements across numerous scientific fields. For instance, the second phase of the Human Genome Project (HGP2) has set an ambitious goal to collect multi-omics data from 1% of the global population—equating to 80 million



people out of 8 billion (Liu, 2024). Open science initiatives, which advocate for the open sharing of scientific knowledge, data, and methodologies, have the potential to revolutionize research by facilitating data exchange (Commission, 2024; NASA, 2025; UNESCO, 2025). Significant scientific breakthroughs, such as the rapid development of COVID-19 vaccines, have been greatly accelerated by open data sharing and collaborative efforts within the global scientific community (Zastrow, 2020).

Despite the promising benefits of open science, a fully free and transparent system is not always practical (Scheliga and Friesike, 2014). This issue, often referred to as the “Open Science Dilemma,”

is analogous to the classic “Prisoner’s Dilemma” in game theory, where individuals must choose between acting for the collective good or prioritizing personal interests. In the realm of open science, researchers face similar trade-offs. Although open data sharing can greatly benefit the scientific community by fostering faster discovery and wider collaboration, individual researchers often hesitate to participate due to concerns over losing intellectual property rights, recognition, or competitive advantage (Scheliga and Friesike, 2014). Additionally, the absence of sufficient incentives discourages participation, especially when maintaining high data quality requires substantial extra effort. This problem is particularly

pronounced in the private sector, where significant volumes of valuable data are held but cannot be shared openly due to IP constraints (Micheli, 2022).

In addition, data safety and privacy concerns, particularly for sensitive human-related data, further compound these challenges. Strict data protection laws in regions such as the U.S., Europe, and China—like HIPAA, GDPR and PIPL—create significant legal and ethical barriers to open sharing (Ness and Committee, 2007; Voigt, 2017; Personal Information Protection Law, 2021). According to IBM's 2024 Cost of a Data Breach Report, the global average total cost of a data breach is \$4.88 million (IBM, 2024). In 2023, the genetic testing company 23andMe experienced a major hacker attack, compromising the personal information of nearly 7 million customers, including sensitive genetic data. This incident not only led to widespread public concern but also sparked class-action lawsuits against the company (Tidy, 2023). Another notable breach is the 2018 data breach at MyHeritage, affecting 92 million users (MyHeritage, 2018). These incidents underscore the vulnerability of biological data security. Studies have shown that even anonymized genetic data can be re-identified through cross-referencing with other publicly available data sources (Bonomi et al., 2020; He and Zhou, 2020).

To address these challenges, we propose a 'Cryptographic Open Science' framework that leverages advanced Web 3.0-related cryptography technologies such as blockchain, data encryption (particularly Fully Homomorphic Encryption, FHE) and crypto tokens to create robust data protection mechanisms that safeguard security, intellectual property, and privacy while incentivizing participation (Figure 1). Although it may seem counterintuitive, sharing data with cryptographic protections allows for secure, controlled data exchange without compromising privacy, data ownership, or competitive advantage. This balanced approach between openness and security provides a viable path forward for advancing open science, ensuring that data can be shared in a way that maintains trust and encourages collaboration.

2 Related research

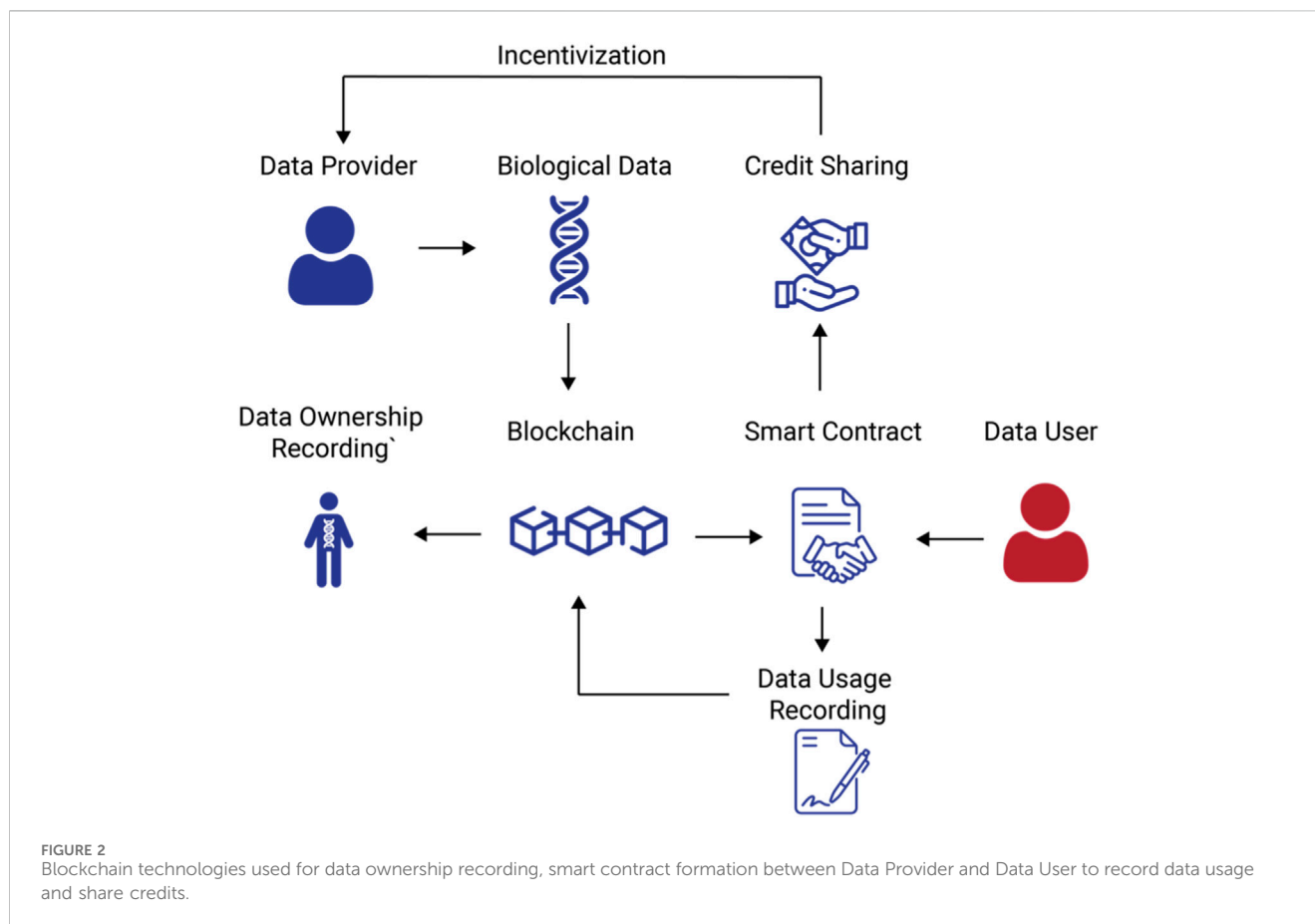
Blockchain is a decentralized, distributed ledger technology that allows data to be recorded, shared, and maintained across a network of computers without the need for a central authority (Crosby et al., 2016; Sarmah, 2018). Each block in the blockchain contains a list of transactions or data entries, a timestamp, and a cryptographic link to the previous block, forming a secure and immutable chain. The use of blockchain techniques ensures that once data is recorded, it cannot be altered without the consensus of the network participants (Figure 2). MedRec is an early exemplar that uses a blockchain layer to manage medical data permissions and access events, while FHIRChain targets clinical interoperability by binding access management to HL7 FHIR-based workflows (Azaria et al., 2016; Zhang et al., 2018).

However, blockchain alone is insufficient to protect data due to the inherent nature of digital assets, where duplication costs are virtually zero. Once data are released publicly, it becomes impossible to maintain control over them, even if data ownership is recorded on

blockchains. Privacy concerns and restrictive government policies further compound the challenges of sharing data in plain formats. Before modern cryptographic approaches were widely usable, many biomedical consortia adopted federated query architectures that keep data behind institutional firewalls and only return restricted outputs (typically aggregate counts or vetted summaries). Early systems such as SHRINE operationalized cross-site cohort discovery by sending standardized queries to local clinical repositories and returning aggregate results under institutional review board (IRB) constraints (Weber et al., 2009). These systems provide an important baseline: they preserve institutional control and reduce raw-data transfer, but privacy protection largely relies on governance rules, output restrictions, and trust in the hosting institution rather than cryptographic confidentiality guarantees.

Federated Learning is a machine learning paradigm where a global model is trained collaboratively across multiple decentralized devices or servers holding local data samples, without exchanging the data itself (Kairouz and McMahan, 2021; Mammen, 2021). Each participant trains the model on their local data and shares only the model updates (e.g., gradients) with a central server that aggregates them to improve the global model. Federated learning has achieved significant traction in biomedical applications (Crowson et al., 2022; Rieke et al., 2020). Notable large-scale initiatives include the MELLODDY project, where 10 pharmaceutical companies (including Amgen, Bayer, Janssen, Novartis, etc.) collaboratively trained drug discovery models without sharing proprietary data (Heyndrickx et al., 2023). The experiments involved a data set of 2.6+ billion confidential experimental activity data points cross-pharma companies, documenting 21+ million physical small molecules and 40+ thousand assays in on-target and secondary pharmacodynamics and pharmacokinetics. The results demonstrated markedly higher improvements in multi-partner federated models compared to single-partner models. The Federated Tumor Segmentation (FeTS) initiative spanning 30 international healthcare institutions focusing on training deep learning models for brain tumor segmentation (Pati et al., 2022). While FL enables collaborative model training without data pooling, recent work has exposed critical vulnerabilities—gradient leakage attacks can reconstruct training images (Zhu et al., 2019). Another significant challenge is Statistical Heterogeneity. In healthcare, data distributions vary significantly across institutions due to differences in scanner protocols, patient demographics, and labeling practices. This heterogeneity can lead to slow convergence or poor performance of the global model. Additionally, System Heterogeneity—the varying computational power and network bandwidth of client nodes—poses logistical challenges, as the speed of the training round is often bottlenecked by the slowest participant (Barona López and Borja Saltos, 2025; Li et al., 2025; Shah et al., 2025).

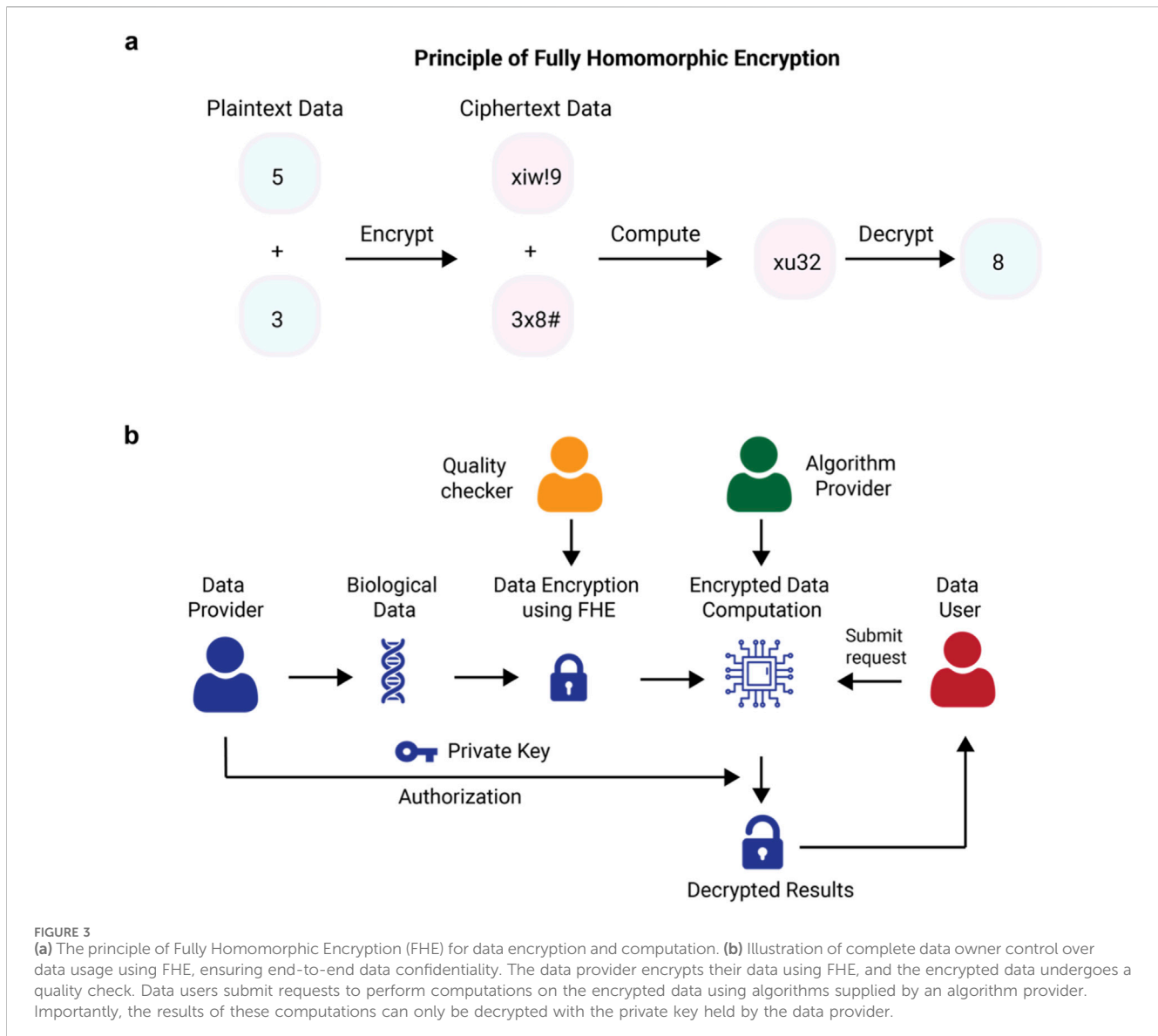
Secure Multi-Party Computation (SMPC) is a set of cryptographic protocols that enable multiple parties to jointly compute a function over their inputs while keeping those inputs private (Zhao et al., 2019; Tran et al., 2021). Each party learns only the output of the computation and gains no additional information about the other parties' inputs. Cho, et al. introduce a framework that enables large-scale genome-wide association studies (GWAS) to be conducted directly on privately held genomic and phenotypic data using secret sharing and secure multiparty computation. By



addressing key computational bottlenecks—particularly quality control and population stratification correction—and achieving linear scalability with respect to cohort size, this work demonstrates that rigorous privacy guarantees can be reconciled with the practical demands of modern GWAS (Cho et al., 2018). Differential privacy (DP) offers a mathematically grounded approach to limiting information leakage from released statistics or trained models by injecting calibrated noise. In biomedical settings, DP is most naturally aligned with releasing aggregates, summary statistics, or privacy-preserving model updates. Yet practical deployments must confront domain-specific constraints: biomedical variables are often highly correlated (e.g., linkage disequilibrium in genetics, repeated measures in EHR trajectories), and naively applying DP mechanisms can either weaken privacy guarantees under dependency assumptions or degrade utility significantly. Recent work has explicitly studied DP under dependent records (including genomic-style dependencies), emphasizing that privacy/utility tradeoffs depend strongly on the statistical structure of the data and on how privacy budgets are tracked across repeated queries. This motivates hybrid approaches where DP is applied selectively (e.g., only to certain outputs or query types) and coupled with cryptographic controls that reduce exposure in the first place (Almadhoun et al., 2020). Zero Knowledge Proofs (ZKP)s are cryptographic protocols that allow one party (the prover) to prove to another party (the verifier) that a certain statement is true without revealing any additional information beyond the validity of the statement itself (Sun et al.,

2021; Berentsen et al., 2023). ZKPs enable verification of knowledge or attributes without disclosing the underlying data. Trusted Execution Environments (TEE)s are hardware solutions that provide secure areas within a processor that isolate code execution and data processing from the rest of the system (Sabt et al., 2015; Jauernig et al., 2020). They protect sensitive computations from being accessed or tampered with by unauthorized parties, even if the operating system or other software is compromised. While TEEs are generally faster, they require trust in the hardware manufacturer and have been shown to be vulnerable to side-channel attacks (e.g., Spectre, Meltdown), which limits their utility for extremely sensitive genomic data compared to pure cryptographic solutions.

FHE is a form of encryption that allows computations to be performed directly on encrypted data (ciphertexts) without needing to decrypt it first (Figure 3). The result of these computations remains encrypted, and when decrypted by the data owner, matches the result of operations performed on the plaintext (Marcolla et al., 2022; Kun, 2024). This enables secure data processing in untrusted environments such as cloud platforms. Recent advancements in Fully Homomorphic Encryption (FHE) have significantly enhanced the potential for secure and privacy-preserving analysis of genomic and clinical data at scale (Blatt et al., 2020; Kim et al., 2020; Ya et al., 2022; Grishin et al., 2019; Grishin et al., 2021). For example, Cheon et al. who developed an FHE-based method for genome-wide association studies (GWAS), and their approach allows chi-square statistics to be computed on encrypted



genomic data, achieving accuracy comparable to traditional methods while preserving data privacy (Kim et al., 2020).

Each of these technologies offers a unique approach to addressing data privacy challenges, and their combined use can ensure the highest level of data protection tailored to specific application scenarios. For example, private keys used to decrypt FHE results can be securely stored within a Trusted Execution Environment (TEE), preventing unauthorized access (Behnke, 2023).

3 Cryptographic Open Science framework

3.1 Blockchain for data ownership and transparent data usage

In the context of open science, blockchain can be utilized to create immutable records of data, including data provenance, ownership, and usage (Figure 2) (Pi et al., 2022; Khan et al., 2023). These immutable

records provide proof of data origin and ownership. By ensuring that original data contributors can be properly traced, blockchain enhances accountability and recognition within the research community. Additionally, the transparent nature of blockchain allows for easy auditing of data collection, access, and usage, enabling data users to verify whether the data resources and collection procedures meet their standards—a factor particularly important when handling sensitive human-related data (Kaaniche and Laurent, 2017). A key concept embedded in the Web 3.0 community is “Don’t trust, verify,” derived from the immutability and transparency of blockchain. Recording data usage on the blockchain enables data providers and owners to track and examine each instance of their data being used. This transforms the traditional collaboration paradigm from building trusted networks to establishing verifiable networks, which is critical for accelerating open science among a large group of scientists.

Researchers can also leverage blockchain-enabled smart contracts, which are self-executing agreements that automatically enforce predefined terms and conditions, to facilitate secure and efficient data sharing (Khan et al., 2021). As a demonstration, we

TABLE 1 Comparison of traditional and encrypted data sharing approaches. This table compares traditional and encrypted data sharing approaches across five key aspects. Encrypted data sharing shows clear advantages in privacy protection, data utility, collaborative potential, and regulatory compliance, while traditional methods have an edge in lower computational overhead.

Aspect	Traditional data sharing	Encrypted data sharing
Privacy protection	Low	High
Data utility	Limited	Full
Collaborative potential	Restricted	Extensive
Regulatory compliance	Challenging	Simplified
Computational overhead	Low	High

designed a soul-bound token (SBT)-anchored biological data access-control system that establishes verifiable, non-transferable ownership of datasets. Each Data Owner receives a SBT that binds data control to an immutable digital identity. The SBT encodes (i) a persistent subject identifier (ownerDid), (ii) the owner's registered public-key fingerprints for policy signatures, and (iii) optional compliance attestations such as IRB or ethics approvals. Authorization is governed by two core smart contracts:

AccessACL, which defines and stores access policies—including identity bindings, dataset scope, operation masks, purpose hashes, time windows, geographic constraints, and optional differential-privacy budgets—and UsageLog, an append-only ledger that records immutable state transitions (Requested, Granted, Used, Revoked, JobResult).

When a user submits an access request, the Gateway verifies the corresponding on-chain policy and issues a short-lived signed ticket containing the dataset ID, permitted operations, expiration timestamp, and a cryptographic digest of the encrypted object. This ticket functions as a capability token for encrypted-storage systems, which validate its signature before serving ciphertext blocks. Every read or compute operation triggers an on-chain log event to ensure full accountability. For computational tasks, the same workflow issues a JobPermit binding a dataset, algorithm identifier, and purpose hash. Off-chain orchestrators subscribe to these events, execute approved encrypted workflows, and publish result commitments (hashes of ciphertext outputs and associated metadata) back to the ledger. Post-processing smart-contract hooks handle payment distribution, enforce rate limits, and finalize the audit trail.

3.2 Data encryption for enhanced data privacy and complete owner control

Among the cryptographic technologies, Fully Homomorphic Encryption (FHE) is often regarded as the “holy grail” of encryption, offering a critical advantage: end-to-end data confidentiality, which entirely eliminates exposure of plaintext data (Figure 3). Unlike other methods, FHE does not rely on external parties, hardware manufacturers, or other participants. It allows computations to be securely outsourced to untrusted third-party servers or cloud providers without risking data exposure. By keeping data

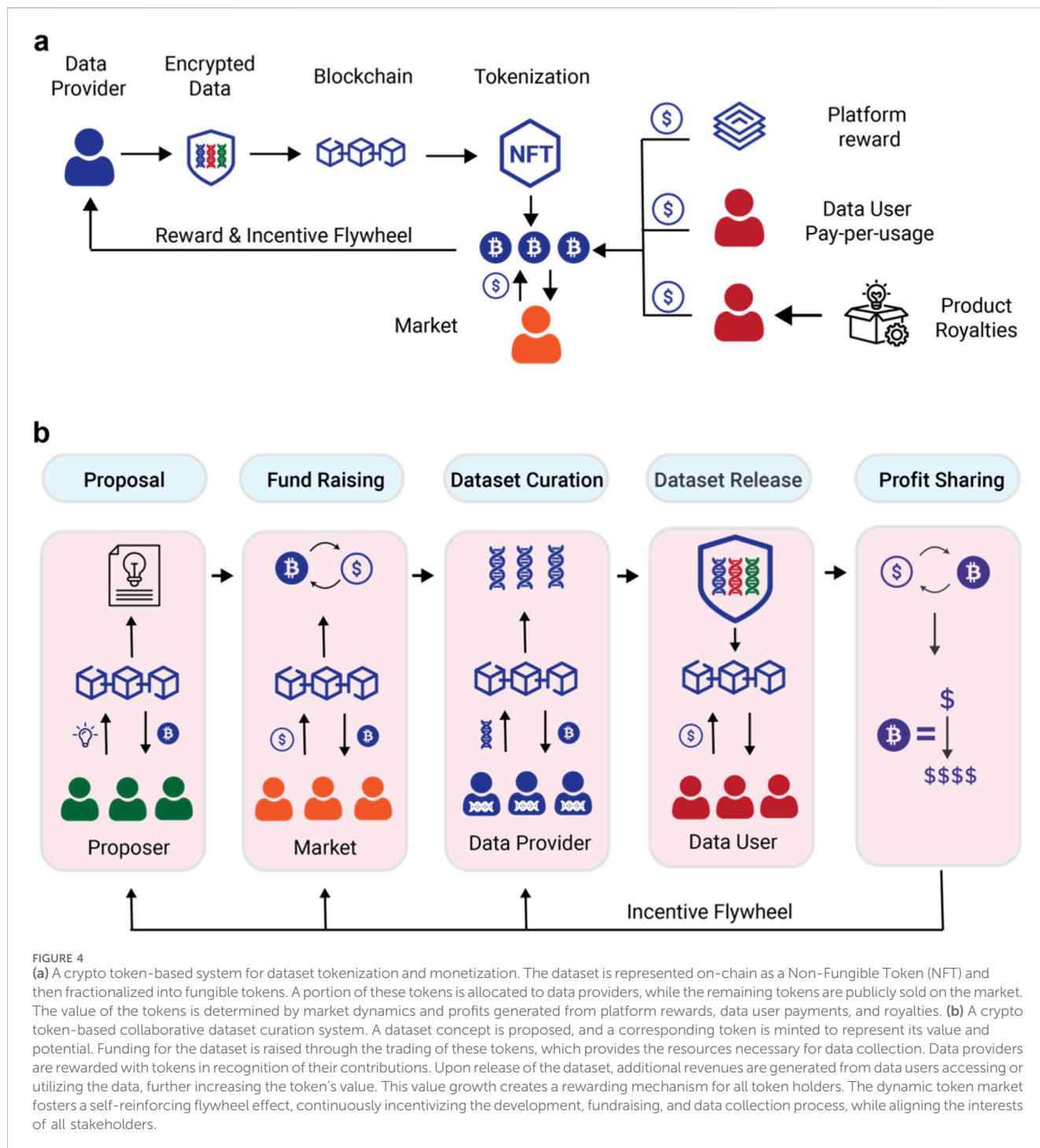
encrypted during processing, FHE also simplifies compliance with data protection regulations like GDPR, HIPAA and PIPL, offering significant advantages over traditional data sharing methods (Table 1).

To further enhance data security, we propose developing blockchain- and FHE-integrated DePIN (Decentralized Physical Infrastructure) devices such as DNA sequencers that perform FHE encryption and blockchain ownership recording at the point of data generation. This ensures that genetic sequencing data is protected from the moment it is created, greatly reducing the risk of data leakage. This approach not only safeguards user privacy but also builds public confidence in genetic sequencing technologies, encouraging broader participation in genomic research. Collaborations with leading sequencing device manufacturers such as Illumina, Oxford Nanopore, and BGI could facilitate the integration of such technologies into existing devices. To ensure the usability of encrypted data, standardized analysis workflows and APIs tailored to common bioinformatics tasks are essential. We propose developing a suite of encrypted data analysis tools, including those for GWAS, epigenomics, transcriptomics, proteomics, and multi-omics integration. These tools will abstract the complexity of encrypted computations, providing researchers with familiar interfaces and output formats. This approach enables biologists who may lack expertise in encryption technologies to conduct sophisticated data analyses effortlessly. We also envision creating an encrypted bioinformatics platform similar to Galaxy (Jalili et al., 2020), equipped with a graphical user interface, workflow management, visualization tools, integrated development environment, data management system, computational resource management, and collaboration tools. This platform will significantly lower the technical barrier to using encrypted data, fostering wider participation in secure data sharing and analysis. Standardized workflows will further reduce these barriers, promoting broader scientific collaboration.

3.3 Crypto token-based incentive system for high-quality data sharing

Addressing data privacy concerns alone does not automatically result in active participation in data sharing. To further incentivize the sharing of high-quality datasets, we propose a crypto token-based incentive system, demonstrated here under the name OMICS (an umbrella term encompassing biological and medical data) as a conceptual framework (Figure 4).

In the proposed OMICS system, data providers submit information about their encrypted datasets onto a blockchain platform and mint a Non-Fungible Token (NFT) representing each unique dataset. This NFT serves as a digital certificate of ownership and provenance, ensuring the dataset's integrity and authenticity within the decentralized network. The NFT can then be fractionalized into fungible tokens, enabling divisible ownership or profit rights associated with the dataset. A portion of these fungible tokens is allocated back to the data provider, granting them continued stake and potential revenue from their data asset. The remaining tokens are listed on decentralized markets or exchanges for trading, providing liquidity and allowing investors or other researchers to acquire stakes in the dataset.



Beyond capital raised through token trading, there are three primary mechanisms to generate additional funding and enhance the value of the tokens:

1. *Upfront Submission Rewards from the Platform*: the upfront submission reward is provided to data providers for submitting datasets and passing quality checks. These rewards are issued in OMICS tokens by the platform and tiered based on data quality and completeness. The data quality can be reviewed through a

- crowd-source peer review system by peer scientists who have expertise in the corresponding domain and the peer scientists who complete the review can also obtain OMICS token awards.
2. *Pay-Per-Use Payments*: the pay-per-use payment system allows data consumers to access and perform computations on the encrypted datasets by paying usage fees facilitated through smart contracts. These payments are automatically and transparently distributed to token holders, providing an ongoing revenue stream linked directly to the dataset's utilization.

3. *Long-Term Credit-Sharing Royalties*: the credit-sharing royalty mechanism tracks the profits gained by data users from utilizing shared data, such as revenue from publications, patents, or product commercialization. Through smart contracts, a predetermined percentage of these profits is distributed to the original data providers and token holders using OMICS tokens in proportion to their data's contribution. Data users are required to declare revenues generated from shared datasets through the platform. This information is cross-verified using blockchain-based data usage records and publicly available records, such as patent databases or publication citations. Non-compliance or detected underreporting can result in penalties, suspension, or exclusion from the platform. This ensures that data providers continue to benefit from the long-term value generated by their data.

The abovementioned system is often challenged by limited dataset size, variable dataset quality, lack of funding for data curation, ambiguous data pricing strategy and unclear identification of data buyers. We here propose a collaborative dataset curation mechanism that allows individuals, organizations, or investors to crowdfund or invest in the creation, curation, quality-control and commercialization of datasets in a community-driven, decentralized, and transparent manner.

The workflow is structured into five key phases: *proposal, funding, dataset curation, dataset release, and profit-sharing*. This organized approach promotes transparency, accountability, and fair rewards for all stakeholders. Notably, the incentive system is not a one-directional, fixed process; it is dynamic. As token prices increase, they create a positive feedback loop that drives greater participation and engagement from stakeholders across all stages, reinforcing the ecosystem's growth and sustainability.

1. *Proposal Phase*: A proposer submits a comprehensive dataset proposal, detailing the scope (description, use cases, and scientific objectives, such as rare disease genomics), funding goals (budget requirements), milestones (phased deliverables and timelines), and contributor incentives. Contributors may include data providers, funding supporters, quality checkers, annotators, brokers, community operators, and the platform itself. These elements form the foundation for a community-driven dataset creation process.
2. *Funding Phase*: The project is launched on the platform, enabling investors to pledge OMICS tokens during the offering period. In return, investors receive dataset-specific tokens, which grant equity or utility rights such as access, profit-sharing, and voting power in the dataset's governance. These tokens can be traded on the market, with a portion of the proceeds allocated to the project's funding pool.
3. *Dataset Curation Phase*: Funds raised are used to recruit data providers, collect data, ensure quality control and proper dataset annotation. Data providers receive tokens as compensation, though the tokens are locked until the data passes quality checks. Funds are released in tranches upon milestone completion, with each milestone verified and recorded on the blockchain. The dataset may establish a Decentralized Autonomous Organization (DAO) to govern

the curation process and manage funding allocation, with token holders actively participating in decision-making, such as setting data quality standards.

4. *Dataset Release Phase*: The curated dataset is encrypted and securely stored on the platform. Additional stakeholders, such as data brokers, may be introduced to promote and negotiate usage deals with data consumers, expanding the dataset's reach and facilitating its commercialization.
5. *Profit-Sharing Phase*: Revenues generated from dataset usage (e.g., pay-per-use fees, licensing royalties, or downstream commercialization) are distributed among token holders. A transparent revenue-sharing mechanism allocates earnings proportionally to token ownership. Additionally, a token burning mechanism reduces token supply, potentially increasing token value over time. All transactions and revenue flows are immutably recorded on the blockchain, ensuring transparency, fairness, and trust for all contributors and investors.

The incentive system can also be designed to reduce malicious behaviors. A reputation system can be implemented, where users earn reputation points based on their contributions, behavior, and the quality of their data. Higher reputation levels would unlock additional benefits, such as increased OMICS rewards or reduced transaction fees, incentivizing users to maintain high standards of participation and collaboration. Additionally, a staking mechanism can be introduced, allowing users to stake OMICS tokens on the platform in exchange for rewards that are proportional to the amount staked. This mechanism encourages users to actively participate and align their interests with the platform's success. Any malicious behavior or breach of platform rules would result in the forfeiture of staked OMICS and a corresponding reputation score drop, creating a deterrent against unethical actions and fostering a culture of accountability. Staking OMICS tokens could also grant users voting rights in platform governance, particularly if the platform operates as a Decentralized Autonomous Organization (DAO). This empowers users to participate in decision-making processes, ensuring that the platform evolves in a decentralized and community-driven manner. Moreover, the staking mechanism not only incentivizes long-term engagement but also contributes to token stability by reducing the circulating supply and encouraging committed, responsible participation.

The crypto token-based incentive mechanism provides a decentralized and transparent approach to funding and managing high-value datasets, offering significant benefits for all stakeholders. For data providers, it ensures fair compensation for their contributions through milestone-based funding, token rewards, and profit-sharing from dataset usage. For investors, it creates opportunities for financial returns and strategic access to curated datasets. Data users benefit from access to high-quality, encrypted datasets with clear pricing and flexible payment options. The platform fosters collaboration through decentralized governance, allowing stakeholders to influence decisions and monitor progress transparently via blockchain.

It is important to emphasize that the goal of this incentive system is not to create paywalls that restrict access to datasets. Instead, it aims to acknowledge the time, effort, and resources invested by data providers and other contributors, ensuring they receive appropriate recognition and fair compensation for their contributions. Data providers retain the option to donate their datasets freely or waive fees for non-profit researchers. However,

TABLE 2 Cryptographic open science platform features.

Feature	Description
Zero trust	Data providers' information is fully protected through fully homomorphic encryption (FHE), allowing them to share data securely without needing to trust other entities
Per-use authorization	Data providers maintain complete control over data usage by authorizing each access with private keys, addressing concerns over loss of control after data is shared
Provable privacy	The platform guarantees data privacy through verifiable cryptographic methods, meeting regulatory standards and addressing privacy concerns of oversight bodies
Transparency and verifiability	Data ownership, transactions, and usage are recorded immutably on a public blockchain, providing transparency and allowing for verifiable audit trails
Fair incentives	Market driven incentives to provide funding and compensation to data providers, to enhance data quality and to promote data usage
Automated credit sharing	Smart contracts automate transactions, ensuring that credits and benefits are shared equitably with data providers, enhancing incentives for data sharing

this choice should be entirely voluntary, reflecting their autonomy, rather than being imposed by the limitations of traditional systems that fail to protect their rights and interests. This approach strikes a balance between promoting open science and safeguarding the rights of data contributors.

3.4 Features of COS framework

Implementing a Cryptographic Open Science framework offers transformative potential for biomedical research (Table 2). This framework could significantly accelerate the advancement of precision medicine. Secure data sharing will enable researchers to access large-scale, diverse datasets, driving progress in scientific research, drug development, and personalized treatment plans. For example, researchers studying rare diseases often face difficulties in reaching statistically significant conclusions due to limited sample sizes. Encrypted data sharing can overcome geographical and institutional barriers, allowing researchers to pool sufficient case data, thus speeding up the discovery of disease mechanisms and the development of targeted therapies. Additionally, this framework will promote global scientific collaboration, break down data silos, and accelerate the pace of scientific discovery.

4 Challenges and future directions

The implementation of COS framework also faces several technical and ethical challenges.

From a technical perspective, while FHE technology has made significant strides recently, it is still evolving and requires further optimization to handle large-scale biological data analysis effectively. The primary challenge for FHE is its computational overhead, with operations being several orders of magnitude slower than plaintext computations (Zhang et al., 2022). Addressing this challenge requires an integrative approach that includes advanced scheme and algorithm development, optimized bootstrapping techniques, and hardware acceleration. Selecting the appropriate FHE scheme based on application needs can significantly improve performance (Marcolla et al., 2022). For instance, the Brakerski-Fan-Vercauteren (BFV, 2012) scheme is optimized for integer arithmetic

operations. The Torus FHE (TFHE, 2016) scheme focuses on high-speed bootstrapping and is ideal for binary computations, while the Cheon-Kim-Kim-Song (CKKS, 2017) scheme is designed for approximate computations on floating-point numbers, making it particularly useful for machine learning applications. Both academic and industry players are actively developing FHE libraries for practical applications, including IBM's HELib, Microsoft's SEAL, PALISADE, and TFHE (Gouert et al., 2023). Hardware acceleration is another critical factor in enhancing FHE efficiency (Zhang et al., 2022). The development of Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) can greatly speed up FHE operations. ASICs, for example, are estimated to accelerate computations by more than 10,000 times compared to single-core CPU (Ahmad A et al., 2023). Additionally, application-specific optimizations, such as pre-computed lookup tables, can reduce computational requirements, and layered security models can be used to categorize data according to privacy levels, optimizing processing based on the sensitivity of the data. Additionally, transitioning existing bioinformatics algorithms to encrypted computing platforms remains a massive engineering challenge, necessitating the redesign of many classic algorithms. This task demands substantial investments in manpower, time, and close collaboration between biologists, computer scientists, and cryptographers. To make Fully Homomorphic Encryption (FHE) more accessible to non-experts, the community can focus on three key areas of development. First, we can create FHE compilers that allow users to write code in high-level languages like Python, which the compiler then automatically maps to encrypted computations. This approach enables developers to leverage FHE without needing specialized cryptographic knowledge. Second, FHE Software Development Kits (SDKs) can be developed to provide user-friendly, low-level APIs that simplify FHE implementation. While this requires developers to have a basic understanding of homomorphic encryption, they are shielded from its most complex aspects. Finally, we can build bioinformatic algorithm libraries specifically designed for FHE, effectively transferring fundamental algorithmic functions into an encrypted environment. Hosting hackathons and programming competitions could further facilitate collaborations between FHE experts and bioinformaticians to develop such libraries.

While the Cryptographic Open Science (COS) framework offers a technical resolution to the Open Science Dilemma, it introduces a complex set of cyber-ethical challenges that must be navigated to avoid exacerbating existing inequities. A primary ethical concern cited in critical literature is the “hyper-financialization” of health data. By tokenizing genomic information, the COS framework risks transforming patients from altruistic research participants into speculative assets. Additionally, establishing and maintaining a global encrypted biological data sharing platform requires significant investment and sustained international cooperation. Balancing the interests of diverse stakeholders and establishing a sustainable operational model are key challenges. Traditional data sharing initiatives like United Kingdom Biobank and China Kadoorie Biobank rely heavily on public funding and donations. However, the introduction of decentralized Web 3.0 technologies, such as blockchain and token-based incentives, could revolutionize the funding model by encouraging participation and unlocking market-driven investment from the public. This approach has the potential to raise funds at levels far exceeding traditional methods. Critics argue that market-driven incentives might skew research priorities toward commercially viable conditions (e.g., those affecting wealthy populations) at the expense of rare diseases or public health crises affecting marginalized groups. Furthermore, the volatility of crypto-assets could introduce instability into scientific funding, where the value of a research project fluctuates with token market speculation rather than scientific merit. Thirdly, a fundamental tension exists between the immutability of blockchain and the “Right to Erasure” (Right to be Forgotten) mandated by GDPR (Article 17) (Regulation, 2016). Strict interpretations of GDPR suggest that even hashed or encrypted personal data on a blockchain might still qualify as “personal data” if it can technically be linked back to an individual. A possible solution is to utilize “crypto-shredding” (destroying decryption keys) as a compliance mechanism. However, if a blockchain contains immutable references to personal data that a patient has revoked consent for, the entire protocol could still face significant regulatory risk, potentially forcing a “hard fork” or abandonment of the network. Lastly, the complexity of Web 3.0 interfaces poses a significant barrier to equitable access. Expecting patients to manage private keys, wallets, and gas fees for “self-sovereign identity” creates a high technical threshold that may exclude elderly, low-income, or less tech-savvy populations. This could lead to biased datasets that over-represent tech-literate demographics, subsequently training AI models that generalize poorly to the broader population. While FHE offers robust privacy protection, large-scale public education and outreach will be necessary to build understanding and trust in these technologies. Additionally, an international effort is required to harmonize legal and ethical standards across countries and regions, creating a flexible global regulatory framework that respects local requirements while upholding the core principles of data sharing. An international governing body would be needed to ensure that global projects adhere to these standards.

5 Conclusion

The Cryptographic Open Science framework provides a promising approach to tackling the privacy and security

challenges associated with biomedical big data sharing. By integrating advanced cryptographic technologies, decentralized mechanisms, and international collaboration, this framework can facilitate true “open science” while safeguarding individual privacy and data sovereignty. This approach will not only accelerate the development of precision medicine and public health initiatives but also establish a solid data foundation for the application of artificial intelligence in the biomedical field and individual-centric fields like metaverse and digital twins (El Saddik et al., 2025; Saad et al., 2023).

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#), further inquiries can be directed to the corresponding authors.

Author contributions

YW: Conceptualization, Funding acquisition, Writing – original draft, Writing – review and editing. JF: Investigation, Methodology, Resources, Writing – original draft, Writing – review and editing. ZB: Investigation, Resources, Writing – original draft, Writing – review and editing. SZ: Writing – original draft, Investigation. LL: Investigation, Resources, Writing – original draft, Writing – review and editing. MY: Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review and editing. GC: Conceptualization, Investigation, Resources, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by Institute of Digital Medicine at City University of Hong Kong.

Conflict of interest

Authors YW, ZB, and SZ were employed by Zhiyu DAO, KanDaoShanWei Technology Inc.

Author JF was employed by Open Security Research.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. GPT 4.0 was used to improve the readability and language of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure

accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product

that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbloc.2025.1631217/full#supplementary-material>

References

- Ahmad Al Badawi, D. B. C., Polyakov, Y., and Rohloff, K. (2023). Hardware acceleration of fully homomorphic encryption: making privacy-preserving machine learning practical
- Almadhoun, N., Ayday, E., and Ulusoy, Ö. (2020). Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics* 36, 1696–1703. doi:10.1093/bioinformatics/btz837
- Azaria, A., Ekblaw, A., Vieira, T., and Lippman, A. (2016). *2nd international conference on open and big data (OBD)*. IEEE, 25–30.
- Barona López, L. I., and Borja Saltos, T. (2025). Heterogeneity challenges of federated learning for future wireless communication networks. *J. Sens. Actuator Netw.* 14, 37. doi:10.3390/jsan14020037
- Behnke, R. (2023). What is a trusted execution environment (TEE)? Available online at: <https://www.halborn.com/blog/post/what-is-a-trusted-execution-environment-tee>.
- Berentsen, A., Lenzi, J., and Nyffenegger, R. (2023). An introduction to zero-knowledge proofs in blockchains and economics. *Fed. Reserve Bank St. Louis Rev.* 105, 280–294. doi:10.20955/r.105.280-94
- Blatt, M., Gusev, A., Polyakov, Y., Rohloff, K., and Vaikuntanathan, V. (2020). Optimized homomorphic encryption solution for secure genome-wide association studies. *BMC Med. Genomics* 13, 1–13. doi:10.1186/s12920-020-0719-9
- Bonomi, L., Huang, Y., and Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nat. Genetics* 52, 646–654. doi:10.1038/s41588-020-0651-0
- Cho, H., Wu, D. J., and Berger, B. (2018). Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnology* 36, 547–551. doi:10.1038/nbt.4108
- Commission, E. (2024). The EU's open science policy. Available online at: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.
- Crosby, M., Pattanayak, P., Verma, S., and Kalyanaraman, V. (2016). Blockchain technology: beyond bitcoin. *Appl. Innovation* 2, 71. Available online at: <https://scc.berkeley.edu/wp-content/uploads/AIR-2016-Blockchain.pdf>.
- Crowson, M. G., Moukheiber, D., Arévalo, A. R., Lam, B. D., Mantena, S., Rana, A., et al. (2022). A systematic review of federated learning applications for biomedical data. *PLoS Digit. Health* 1, e0000033. doi:10.1371/journal.pdig.0000033
- El Saddik, A., Ahmad, J., Khan, M., Abouzahir, S., and Gueaieb, W. (2025). Unleashing creativity in the metaverse: generative AI and multimodal content. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 1–43. doi:10.1145/3713075
- Gouert, C., Mouris, D., and Tsoutsos, N. (2023). Sok: new insights into fully homomorphic encryption libraries via standardized benchmarks. *Proc. Privacy Enhancing Technologies* 2023, 154–172. doi:10.56553/popets-2023-0075
- Grishin, D., Obbad, K., and Church, G. M. (2019). Data privacy in the age of personal genomics. *Nat. Biotechnology* 37, 1115–1117. doi:10.1038/s41587-019-0271-3
- Grishin, D., Raisaro, J. L., Troncoso-Pastoriza, J. R., Obbad, K., Quinn, K., Misbach, M., et al. (2021). Citizen-centered, auditable and privacy-preserving population genomics. *Nat. Comput. Sci.* 1, 192–198. doi:10.1038/s43588-021-00044-9
- He, Z., and Zhou, J. (2020). Inference attacks on genomic data based on probabilistic graphical models. *Big Data Min. Anal.* 3, 225–233. doi:10.26599/bdma.2020.9020008
- Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., et al. (2023). Melloddy: cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *J. Chemical Information Modeling* 64, 2331–2344. doi:10.1021/acs.jcim.3c00799
- IBM (2024). Cost of a data breach report. Available online at: <https://www.ibm.com/reports/data-breach>.
- Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., et al. (2020). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research* 48, W395–W402. doi:10.1093/nar/gkaa434
- Jauernig, P., Sadeghi, A.-R., and Stapf, E. (2020). Trusted execution environments: properties, applications, and challenges. *IEEE Secur. and Priv.* 18, 56–60. doi:10.1109/msec.2019.2947124
- Kaaniche, N., and Laurent, M. (2017). *2017 IEEE 16th international symposium on network computing and applications (NCA)*. IEEE, 1–5.
- Kairouz, P., and McMahan, H. B. (2021). Advances and open problems in federated learning. *Found. Trends® Machine Learning* 14, 1–210. doi:10.1561/22000000083
- Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., and Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: techniques and applications. *Comput. Biol. Med.* 158, 106848. doi:10.1016/j.combiomed.2023.106848
- Khan, S. N., Loukil, F., Ghedira-Guegan, C., Benkhelifa, E., and Bani-Hani, A. (2021). Blockchain smart contracts: applications, challenges, and future trends. *Peer. Peer. Netw. Appl.* 14, 2901–2925. doi:10.1007/s12083-021-01127-0
- Khan, M., El Saddik, A., and Gueaieb, W. (2023). “In 2023 IEEE international conference on metaverse computing, networking and applications (MetaCom),”. IEEE, 622–626.
- Kim, D., Son, Y., Kim, D., Kim, A., Hong, S., and Cheon, J. H. (2020). Privacy-preserving approximate GWAS computation based on homomorphic encryption. *BMC Med. Genomics* 13, 1–12. doi:10.1186/s12920-020-0722-1
- Kun, J. A. (2024). High-level technical overview of fully homomorphic encryption. Available online at: <https://www.jeremykun.com/2024/05/04/fhe-overview/>.
- Li, M., Xu, P., Hu, J., Tang, Z., and Yang, G. (2025). From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare. *Med. Image Anal.* 101, 103497. doi:10.1016/j.media.2025.103497
- Liu, W. (2024). The 1% gift to humanity: the human genome project II. *Cell Res.* 1-4. doi:10.1038/s41422-024-01026-y
- Mammen, P. M. (2021). Federated learning: opportunities and challenges. arXiv preprint arXiv:2101.05428
- Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H. P., and Aaraj, N. (2022). Survey on fully homomorphic encryption, theory, and applications. *Proc. IEEE* 110, 1572–1609. doi:10.1109/jproc.2022.3205665
- Micheli, M. (2022). Public bodies' access to private sector data: the perspectives of twelve European local administrations. *First Monday*. doi:10.5210/fm.v27i2.11720
- MyHeritage (2018). Cybersecurity incident: june 10 update. Available online at: <https://blog.myheritage.com/2018/06/cybersecurity-incident-june-10-update/>.
- NASA (2025). NASA open-source science initiative. Available online at: <https://nasa-impact.github.io/ossi-website/>.
- Ness, R. B., and Committee, J. P. (2007). Influence of the HIPAA privacy rule on health research. *Jama* 298, 2164–2170. doi:10.1001/jama.298.18.2164
- Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S. H., Reina, G. A., et al. (2022). Federated learning enables big data for rare cancer boundary detection. *Nat. Communications* 13, 7346. doi:10.1038/s41467-022-33407-5
- Personal Information Protection Law (2021). Personal information protection law of the people's Republic of China. Available online at: <https://personalinformationprotectionlaw.com/#:~:text=The%20PIPL%20came%20into%20effect,legal%20basis%20and%20disclosure%20requirements>.
- Pinto, R. P., Silva, B. M., and Inacio, P. R. (2022). A system for the promotion of traceability and ownership of health data using blockchain. *IEEE Access* 10, 92760–92773. doi:10.1109/access.2022.3203193
- Regulation Regulation (EU) 2016/679, Article 17 (right to erasure). (2016).
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *NPJ Digital Medicine* 3, 119. doi:10.1038/s41746-020-00323-1
- Saad, M., Khan, M., Saeed, M., El Saddik, A., and Gueaieb, W. (2023). “IEEE international smart cities conference (ISC2),”. IEEE, 1–5.

- Sabt, M., Achemlal, M., and Bouabdallah, A. (2015). *IEEE trustcom/bigDataSE/ispa*. IEEE, 57–64.
- Sarmah, S. S. (2018). Understanding blockchain technology. *Comput. Sci. Eng.* 8, 23–29. doi:10.5923/j.computer.20180802.02
- Schelig, K., and Friesike, S. (2014). Putting open science into practice: a social dilemma? *First Monday*. doi:10.5210/fm.v19i9.5381
- Shah, S. T., Ali, Z., Waqar, M., and Kim, A. (2025). *Healthcare*, 2760. MDPI.
- Sun, X., Yu, F. R., Zhang, P., Sun, Z., Xie, W., and Peng, X. (2021). A survey on zero-knowledge proof in blockchain. *IEEE Network* 35, 198–205. doi:10.1109/mnet.011.2000473
- Tidy, S. M. J. (2023). 23andMe: profiles of 6.9 million people hacked. Available online at: <https://www.bbc.com/news/technology-67624182>.
- Tran, A.-T., Luong, T.-D., Karnjana, J., and Huynh, V.-N. (2021). An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation. *Neurocomputing* 422, 245–262. doi:10.1016/j.neucom.2020.10.014
- UNESCO (2025). Open science: making science more accessible, inclusive and equitable for the benefit of all. Available online at: <https://www.unesco.org/en/open-science>.
- Voigt, P. (2017). “Von dem Bussche, A. The EU general data protection regulation (gdpr),” in *A practical guide, 1st Ed., Cham*. 10. Springer International Publishing, 10–5555.
- Weber, G. M., Murphy, S. N., McMurry, A. J., Macfadden, D., Nigrin, D. J., Churchill, S., et al. (2009). The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inf. Assoc.* 16, 624–630. doi:10.1197/jamia.M3191
- Yang, M., Zhang, C., Wang, X., Liu, X., Li, S., Huang, J., et al. (2022). TrustGWAS: a full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation. *Cell Syst.* 13, 752–767. doi:10.1016/j.cels.2022.08.001
- Zastrow, M. (2020). Open science takes on the coronavirus pandemic. *Nature* 581, 109–111. doi:10.1038/d41586-020-01246-3
- Zhang, P., White, J., Schmidt, D. C., Lenz, G., and Rosenbloom, S. T. (2018). FHIRChain: applying blockchain to securely and scalably share clinical data. *Comput. Structural Biotechnology Journal* 16, 267–278. doi:10.1016/j.csbj.2018.07.004
- Zhang, J., Cheng, X., Yang, L., Hu, J., Liu, X., and Chen, K. (2022). Sok: fully homomorphic encryption accelerators. *ACM Comput. Surv.* 56, 1–32. doi:10.1145/3676955
- Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C. Z., Li, H., et al. (2019). Secure multi-party computation: theory, practice and applications. *Inf. Sci.* 476, 357–372. doi:10.1016/j.ins.2018.10.024
- Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. *Adv. Neural Information Processing Systems* 32. Available online at: https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf.

Glossary

Blockchain	A decentralized, distributed ledger technology that creates immutable records of data ownership, provenance, and usage
Crypto-shredding	The act of destroying decryption keys to comply with data erasure mandates
Cryptographic open science (COS)	A framework integrating blockchain, fully homomorphic encryption (FHE), and crypto tokens to enable secure, privacy-preserving, and incentivized biomedical data sharing
Dataset tokens (dTokens)	Individual fungible tokens created by fractionalizing a dataset NFT, allowing direct value capture and profit-sharing for data providers
Decentralized autonomous organization (DAO)	A community-driven organization where token holders manage curation, funding, and decision-making through decentralized governance
DePIN	Decentralized physical infrastructure devices (e.g., DNA sequencers) that perform encryption and blockchain recording at the point of data generation
Differential privacy (DP)	A method of limiting information leakage by injecting calibrated noise into released statistics or model updates
Federated learning (FL)	A paradigm where a global model is trained across decentralized devices without exchanging the local data itself
Fully homomorphic encryption (FHE)	An encryption form that allows computations to be performed directly on ciphertexts without decryption, ensuring end-to-end data confidentiality
Non-fungible token (NFT)	A digital certificate of ownership and provenance used to represent unique datasets within a decentralized network
OMICS token	A platform-level utility token used for infrastructure payments, staking, governance, and ecosystem incentives
Open science dilemma	A situation where researchers hesitate to share data due to privacy concerns, intellectual property risks, and lack of recognition, prioritizing personal interest over the collective good
Secure multi-party computation (SMPC)	Cryptographic protocols that allow multiple parties to jointly compute a function while keeping their individual inputs private
Smart contracts	Self-executing agreements that automatically enforce predefined terms, such as data access controls and credit sharing
Soul-bound token (SBT)	A non-transferable digital token that binds data control to an immutable digital identity to establish verifiable ownership
Trusted execution environments (TEE)	Hardware-based secure areas within a processor that isolate sensitive computations and data from the rest of the system
Zero knowledge proofs (ZKP)	Protocols that allow one party to prove a statement is true without revealing any information beyond the validity of the statement itself