Federated Learning for Cross-Institutional Phytochemical Discovery: Collaborative Model Development Without Centralized Sensitive Data Sharing

Zsolt Kovács[1]*, Katalin Szabo[1], Gabor Nagy[2]

[1]Department of AI for Adaptogenic and Nootropic Plant Extracts, Faculty of Pharmacy, Semmelweis University, Budapest, Hungary.
[2]Department of Machine Learning for Cognitive Enhancement Targets, Faculty of Pharmacy, University of Pécs, Pécs, Hungary.

ABSTRACT

Cross-institutional phytochemical discovery increasingly depends on distributed chemical, biological, omics, safety, provenance, and proprietary datasets that cannot always be transferred to a central repository. Although federated learning may enable participating institutions to develop shared computational models while retaining source data locally, data locality alone does not resolve privacy leakage, cybersecurity, data heterogeneity, scientific validity, governance, or institutional-trust concerns. This article proposes a conceptual federated learning architecture for collaborative phytochemical discovery across natural product laboratories, academic institutions, pharmaceutical research groups, bioinformatics centers, and other authorized partners. The architecture links consortium formation and data-governance agreements with local data stewardship, metadata harmonization, feature and label alignment, local model training, governed model-update exchange, global aggregation, privacy and security safeguards, cross-site validation, bias assessment, expert review, model documentation, lifecycle governance, and structured feedback. It accommodates phytochemical records, chemical representations, bioactivity and assay data, target and pathway annotations, omics evidence where supported, ADMET and toxicity information, proprietary compound libraries, and carefully governed traditional knowledge, biodiversity, clinical, or pharmacovigilance signals where appropriate. The principal contribution is an architecture that treats federated learning as one component of a broader scientific and governance system rather than as an automatic privacy or discovery solution. The proposed framework supports research collaboration and hypothesis prioritization but does not establish privacy certification, security certification, therapeutic validity, clinical utility, regulatory readiness, or operational deployment.

Key Words: *Federated learning, Phytochemical discovery, Privacy-preserving AI, Cross-institutional collaboration, Sensitive data, Secure aggregation*

eIJPPR 2026; 16(3):100-116

HOW TO CITE THIS ARTICLE: Kovács Z, Szabo K, Nagy G. Federated Learning for Cross-Institutional Phytochemical Discovery: Collaborative Model Development Without Centralized Sensitive Data Sharing. Int J Pharm Phytopharmacol Res. 2026;16(3):100-16. https://doi.org/10.51847/y0yuNXQ0Gg

INTRODUCTION

Phytochemical discovery increasingly requires the integration of information that is institutionally distributed, scientifically heterogeneous, and subject to different stewardship conditions. Natural product laboratories may hold locally curated extracts, isolated-compound records, organism and sourcing metadata, or traditional-use information; pharmaceutical groups may maintain proprietary compound libraries and bioassay results; omics centers may generate high-dimensional molecular profiles; and clinical or pharmacovigilance groups may hold

Corresponding author: Zsolt Kovács
Address: Department of AI for Adaptogenic and Nootropic Plant Extracts, Faculty of Pharmacy, Semmelweis University, Budapest, Hungary.
E-mail: ✉ zsolt.kovacs@semmelweis.hu
Received: 12 March 2026; **Revised:** 08 June 2026; **Accepted:** 11 June 2026

downstream signals that cannot be freely transferred. Federated learning offers a collaborative modeling approach in which source records remain within participating institutions while coordinated model development occurs through the exchange and aggregation of approved computational updates. In biomedical settings, this arrangement may reduce routine centralized data transfer, but it does not independently establish privacy, security, scientific validity, or responsible institutional governance [1].

The relevance of this approach to phytochemical research extends beyond the protection of conventional patient records. Chemical structures, unpublished assay results, proprietary molecular libraries, target annotations, structure–activity relationships, natural product provenance, and negative experimental findings may have substantial scientific or commercial sensitivity. Molecular discovery has therefore emerged as a credible domain for federated learning because multiple organizations may benefit from broader training evidence without relinquishing direct control over confidential source datasets [2]. In phytochemical research, the same logic may apply to compound representations, bioactivity models, ADMET prediction, toxicity screening, phenotypic profiling, mechanism-oriented analyses, and candidate-prioritization tasks, provided that the resulting models remain subject to domain-specific quality controls and carefully defined interpretation boundaries.

Avoiding a central data repository is nevertheless only one element of responsible collaborative learning. Federated workflows can still expose information through model parameters, gradients, intermediate statistics, participation patterns, or inadequately controlled outputs. Privacy-preserving drug-discovery research consequently treats federation, secure aggregation, differential privacy, cryptographic protection, trusted execution environments, access controls, and governance procedures as complementary rather than interchangeable safeguards [3]. Their suitability depends on the data type, model architecture, adversary assumptions, permitted outputs, utility requirements, institutional capabilities, and scientific purpose. No single mechanism can be assumed to provide complete protection across all phytochemical, omics, safety, clinical, proprietary, or biodiversity-related contexts.

A federated architecture for phytochemical discovery must therefore address scientific interoperability and institutional responsibility at the same time. Data-locality constraints must be coupled with harmonized metadata, reproducible preprocessing, realistic treatment of non-identically distributed data, site-specific validation, documented model limitations, expert interpretation, and explicit control over model updates and outputs. Privacy and security risks also remain possible even when raw data never leave the originating institution, because collaborative training artifacts may disclose sensitive information under particular attack conditions [4]. The objective of this article is to propose a structured cross-silo federated learning architecture that integrates these technical, scientific, privacy, security, validation, and governance requirements while preserving clear no-privacy-guarantee and no-therapeutic-claim boundaries.

Cross-institutional phytochemical discovery

Cross-institutional phytochemical discovery can involve complementary data assets that are individually valuable but collectively difficult to centralize. One institution may possess chemical structures and standardized bioactivity measurements, another may contribute phenotypic screening data, and another may hold toxicity, omics, pathway, or target-association evidence. Industrial partners may additionally maintain proprietary structures, assay panels, and inactive compounds that would improve model calibration but cannot be disclosed to competitors or external repositories. Large-scale cross-pharmaceutical federated QSAR research demonstrates that collaborative molecular-property modeling can be conducted across confidential institutional datasets without pooling the underlying proprietary structure–activity records, although such collaboration still requires extensive task standardization, governance, and validation [5]. For phytochemical discovery, this precedent supports a model in which participating organizations contribute local evidence to approved predictive tasks while retaining control over compound identities, source records, assay observations, and sensitive provenance information.

The scientific difficulty is not limited to access. Institutional phytochemical collections may differ substantially in chemical-space coverage, taxonomic origin, extraction practices, assay platforms, endpoint definitions, target classes, positive-to-negative ratios, and data completeness. Federated molecular-property studies using graph neural networks have shown that heterogeneous local molecular distributions can influence the value of collaboration and the behavior of the aggregated model [6]. A phytochemical federation must therefore characterize site-level chemical distributions, assay contexts, label frequencies, missingness, and representation differences before training. Institutions with larger datasets should not automatically dominate the collaborative model, and locally rare chemical classes or biological endpoints should not be treated as noise merely because they are uncommon across the consortium.

Data partitioning and provenance are equally important. Randomly dividing a centralized chemical dataset into artificial clients may produce an unrealistically favorable

picture of federation because real institutions frequently hold compounds generated through distinct research programs, screening campaigns, organisms, geographic regions, or intellectual-property strategies. Research on privacy-preserving chemical dataset splitting indicates that the manner in which chemical records are allocated across sites materially affects federated evaluation [7]. The proposed architecture therefore requires institution-aware validation partitions, persistent compound and assay provenance, standardized structure curation, explicit treatment of stereochemistry and mixtures, and documentation of how local records were generated, transformed, filtered, or excluded. These measures are necessary to distinguish true cross-institutional generalization from performance driven by duplicated series, shared scaffolds, or hidden overlap.

Collaborative modeling must also accommodate non-identically distributed data without implying that every dataset should be forced into a single homogeneous representation. Federated drug-discovery analyses under non-IID conditions show that local distribution differences must be treated as an explicit modeling problem rather than as a minor implementation detail [8]. Depending on the task, an architecture may require stratified models, heterogeneity-aware aggregation, site-specific calibration, multitask learning, personalization, uncertainty estimates, or restrictions on the intended population of compounds and assays. Governance considerations operate in parallel: local stewards must determine which records are eligible, traditional knowledge or biodiversity-linked information may require additional authority and benefit considerations, proprietary fields may need exclusion from shared metrics, and clinical or pharmacovigilance signals should be incorporated only where separate authorization, relevance, and minimum disclosure safeguards exist.

Federated learning principles

Federated learning is a coordinated machine-learning arrangement in which multiple data-owning participants perform approved computation on local datasets and contribute model-related information to a collaborative training process. It differs from centralized learning, where source records are pooled into a common repository, and from ordinary distributed computing, where a single owner may divide one accessible dataset across computational resources. Foundational federated-learning taxonomies distinguish configurations according to how samples, features, and participants are distributed, while preserving the defining principle that separate data holders participate without routinely transferring their underlying records [9]. The architecture proposed here is primarily cross-silo: the participants are identifiable institutions with comparatively stable roles, governed infrastructure, persistent datasets, and formal responsibilities.

Federated learning does not remove the statistical and systems challenges that would exist in a pooled analysis and may introduce additional ones. Local sample sizes, compute capacity, network availability, preprocessing pipelines, assay distributions, feature spaces, and labeling practices may differ substantially. These differences can cause unstable optimization, unequal contributions, slow convergence, negative transfer, or a global model that performs well only for the largest participating sites [10]. Cross-silo federation must also be distinguished from cross-device federation, in which very large numbers of intermittently available personal devices contribute small local datasets. Institutional federation generally involves fewer participants, more stable connections, richer datasets, formal governance, and greater opportunity for site-specific validation, but it also involves complex contracts, proprietary interests, accountability requirements, and scientific dependencies [11].

Within a cross-silo workflow, local training and global aggregation are separate functions. Each institution applies a versioned preprocessing and feature-generation pipeline, trains or updates an approved model locally, evaluates it on designated local partitions, and prepares only the permitted model update and summary information for transfer. Multisite biomedical research has demonstrated that coordinated training can occur without directly exchanging source records, while also showing the importance of evaluating performance across participating institutions rather than relying exclusively on an aggregate result [12]. Distributed predictive-modeling studies further illustrate that local optimization can be coordinated through parameter exchange, but the presence of a distributed algorithm does not remove the need to verify local data quality, convergence behavior, update integrity, or site-level utility [13].

Secure aggregation may limit the coordinator's access to individual institutional updates, and differential privacy or encryption may be added where justified by the threat model and acceptable utility trade-offs. These mechanisms do not compensate for inconsistent endpoints, incompatible features, poor metadata, or invalid evaluation design. Federated benchmark research indicates that performance should be assessed under realistic institutional distributions because pooled-data or artificially balanced evaluations can conceal site-specific weaknesses [14]. Consequently, the architecture treats data harmonization, non-IID characterization, cross-site validation, model documentation, expert review, and decision boundaries as core federated-learning principles rather than as optional activities performed after aggregation. **Table 1** summarises federated learning

principles relevant to cross-institutional phytochemical discovery.

Table 1. Federated Learning Principles for Cross-Institutional Phytochemical Discovery: Data Locality, Local Training, Global Aggregation, Cross-Silo Collaboration, Non-IID Data, Secure Aggregation, Differential Privacy, Data Harmonization, Model Documentation, Cross-Site Validation, and Decision Boundaries

Federated learning principle	Core architecture question	Required input or condition	Collaborative modeling function	Potential output	Failure risk	Validation or governance need	Framework role
Data locality	Which records and attributes must remain within the originating institution?	Data classification, permitted-output policy, and local stewardship authority	Keeps raw phytochemical, assay, omics, safety, or proprietary records under local control	Approved local training view and restricted model update	Sensitive information may still be inferred from updates or outputs	Data-flow review, local enforcement, and residual-risk documentation	Defines the non-centralization boundary
Cross-silo collaboration	Which institutions may participate and under what responsibilities?	Consortium charter, participant criteria, roles, and withdrawal procedures	Coordinates persistent institutional clients around approved research tasks	Governed collaborative training network	Misaligned incentives, unqualified participants, or ambiguous accountability	Consortium oversight and institutional approval	Establishes the organizational form of federation
Local training	How is each site permitted to compute on its own data?	Approved model, preprocessing pipeline, local compute environment, and run configuration	Produces local model updates without routine export of source records	Local update, validation metrics, and run manifest	Overfitting, inconsistent preprocessing, or unauthorized local use	Local quality control and reproducibility checks	Converts site-controlled evidence into a federated contribution
Global aggregation	How are eligible local updates combined?	Validated updates, aggregation rule, contribution policy, and version control	Produces a candidate global model state	Aggregated research model and contribution report	Site dominance, negative transfer, corruption, or unstable convergence	Aggregation diagnostics, update review, and rollback capability	Coordinates shared model development
Non-IID data handling	How do chemical, assay, target, or endpoint distributions differ across sites?	Site-level distribution summaries, shift diagnostics, and predefined heterogeneity criteria	Applies weighting, stratification, personalization, multitask learning, or other justified methods	Heterogeneity-aware collaborative model	Poor transfer, hidden site failure, or majority-site bias	Site-wise performance and distribution-shift review	Addresses statistical diversity across institutions
Secure aggregation	Should the coordinator be prevented from viewing	Threat model, protocol assumptions, key-	Aggregates protected updates without routine	Aggregate update with reduced exposure of	Collusion, protocol failure, implementation error, or	Protocol testing and independent security review	Adds update-confidentiality protection

	individual updates?	management process, and dropout policy	inspection of each contribution	individual submissions	metadata leakage		where appropriate
Differential privacy	Is a formal bound on record or participant influence appropriate for the task?	Defined privacy unit, calibrated mechanism, clipping policy, and acceptable utility loss	Adds controlled perturbation to selected updates or outputs	Privacy-bounded computation under stated assumptions	Excessive utility loss, weak calibration, or unequal subgroup effects	Privacy accounting, attack testing, and fairness analysis	Optional privacy layer, not a universal guarantee
Encryption	Which data, updates, and artifacts require cryptographic protection?	Key lifecycle, transport and storage policy, and computational feasibility	Protects selected exchanges, retained artifacts, or approved computations	Encrypted transfer or processing channel	Key compromise, misconfiguration, side channels, or excessive overhead	Configuration, penetration, and key-management audits	Supports confidentiality of selected system components
Data harmonization	Are local variables and metadata semantically comparable?	Common data model, terminology mappings, unit rules, and conformance tests	Produces compatible local representations without centralizing the data	Harmonized local feature and metadata layer	False equivalence, mapping error, or loss of local context	Mapping review, version control, and discrepancy resolution	Creates preconditions for meaningful aggregation
Model documentation	Can collaborators understand what each model version represents?	Dataset summaries, code versions, safeguards, metrics, intended use, and limitations	Records the evidence, configuration, status, and boundaries of the model	Versioned model dossier	Misuse, irreproducibility, hidden limitations, or version confusion	Documentation audit and governance approval	Provides transparency and lifecycle traceability
Cross-site validation	Does the model function acceptably across participating institutional contexts?	Local test partitions, common metrics, subgroup definitions, and uncertainty criteria	Executes validation locally and combines permitted summaries	Site-wise and aggregate validation dossier	Aggregate performance may conceal local failure	Independent review and minimum site-level criteria	Serves as the scientific advancement gate
Decision boundaries	What conclusions may be drawn from collaborative model outputs?	Intended-use statement, claim taxonomy, uncertainty policy, and expert-review procedure	Restricts outputs to approved research interpretation	Hypotheses, model diagnostics, or research-prioritization signals	Unsupported privacy, safety, efficacy, or deployment claims	Publication, interface, and decision-log review	Maintains no-privacy-guarantee and no-therapeutic-claim boundaries

Sensitive data and collaborative modeling

Sensitive phytochemical data include more than personally identifiable information. Proprietary compound libraries, unpublished chemical structures, inactive series, assay protocols, screening images, target hypotheses, synthetic or isolation routes, supplier relationships, and negative results may reveal commercial strategies or institutional research directions. Federated learning has also been applied to cell-painting data for compound mechanism-of-action prediction, illustrating that collaborative discovery may involve high-dimensional phenotypic representations rather than only conventional molecular descriptors or scalar bioactivity endpoints [15]. Such data can be particularly revealing because images, embeddings, or learned features may encode compound-specific or laboratory-specific information. Local stewardship must therefore cover the source data, preprocessing artifacts, generated features, model updates, validation partitions, and any output that could enable reconstruction or competitive inference.

Collaborative modeling may incorporate chemical structures, bioactivity measurements, assay metadata, target and pathway annotations, omics evidence where supported, ADMET or toxicity data, and carefully authorized clinical or pharmacovigilance signals. Multi-party drug-discovery research supports the feasibility of jointly learning from institution-specific datasets while leaving the original records under local control, but it does not establish that every modality is compatible or that resulting candidates are scientifically validated [16]. Reproducibility also depends on transparent implementations, versioned configurations, and modular workflows that make local preprocessing, model architecture, aggregation, and evaluation choices inspectable across participating sites [17]. **Figure 1** illustrates how federated learning can support collaborative phytochemical discovery while keeping sensitive institutional data local.

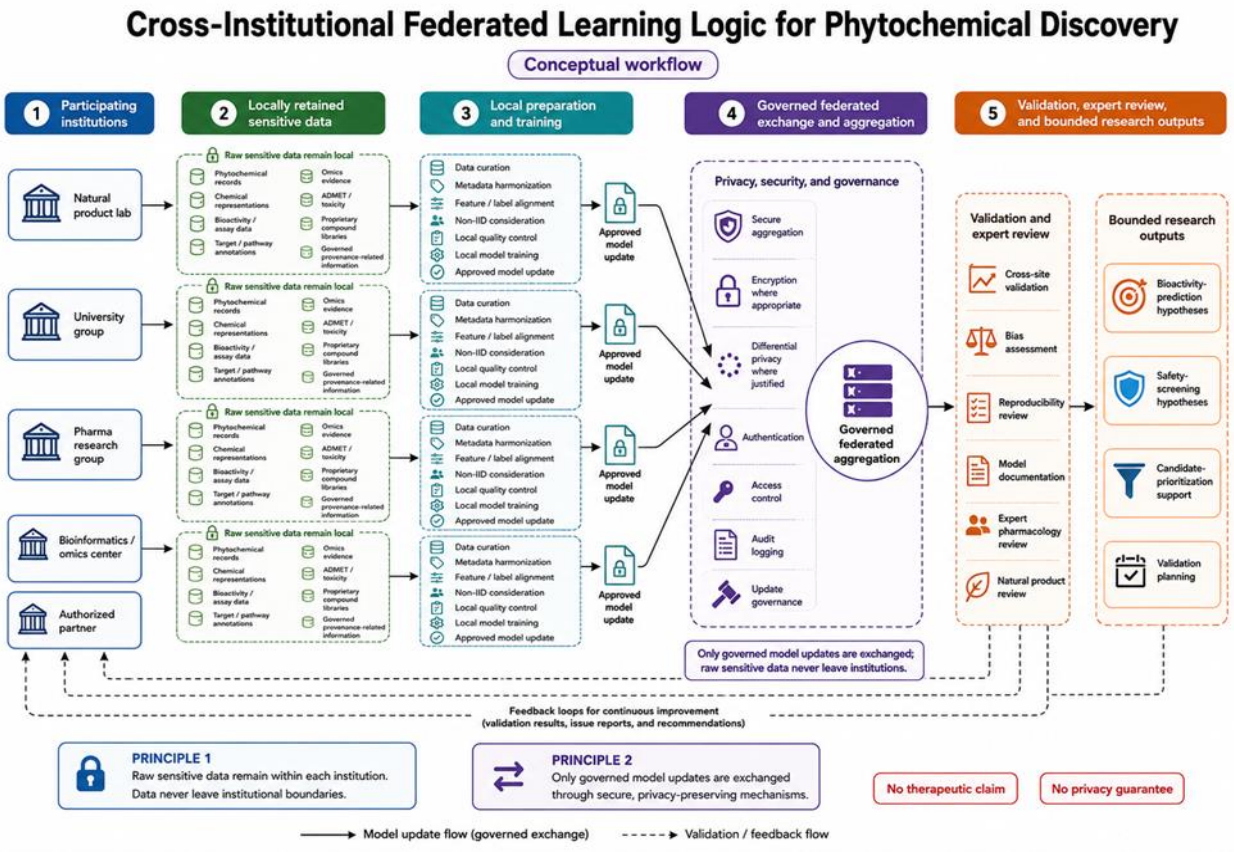

Figure 1. Cross-Institutional Federated Learning Logic for Phytochemical Discovery

Alt-text description

A conceptual federated learning diagram showing multiple institutions keeping sensitive data local while sending governed model updates to a central aggregation layer, followed by validation, expert review, and feedback loops.

A common data model is necessary when institutions use different compound identifiers, assay conventions, feature-generation pipelines, omics platforms, target vocabularies, or toxicity endpoints. Broader federated-healthcare research shows that multimodal collaboration creates

intertwined requirements for interoperability, communication, privacy controls, system resources, and participant coordination [18]. The proposed phytochemical architecture therefore requires each participating site to map approved local information into a shared semantic layer while retaining source-specific provenance and uncertainty. Feature alignment should specify how chemical graphs, descriptors, fingerprints, image embeddings, omics features, and contextual variables are generated. Label harmonization should define endpoints, units, censoring conventions, activity thresholds, species, assay systems, and uncertainty annotations. Where alignment would erase scientifically important differences, the architecture should preserve separate tasks or context-specific outputs rather than imposing artificial equivalence.

Collaborative updating must be embedded within a broader architecture that connects local curation, governance, technical safeguards, validation, and lifecycle review. Architecture-oriented federated healthcare research similarly indicates that aggregation should not be treated as an isolated component detached from interoperability, privacy, evaluation, organizational roles, and operational control [19]. In the proposed system, every local update would therefore be associated with an authenticated participant, approved model version, documented preprocessing pipeline, local quality report, and permitted-use statement. Encryption may protect selected exchanges or collaborative computations, as demonstrated in encrypted multisite biomedical learning, but it remains a partial safeguard that cannot independently prevent inference, poisoning, misconfiguration, or invalid scientific conclusions [20]. Sensitive traditional knowledge and biodiversity or sourcing information would additionally require explicit authority, purpose limitation, and potential exclusion from training, while candidate-level outputs would remain research hypotheses requiring independent experimental and expert evaluation.

Privacy, security, and governance requirements

Federated learning changes the location and form of information exchange, but it does not eliminate privacy or security exposure. Threats may arise from participating institutions, the coordinating service, external adversaries, compromised credentials, improperly configured infrastructure, or model outputs that reveal information about local training records. Privacy and security reviews have identified risks involving membership inference, model inversion, gradient leakage, update interception, malicious clients, and manipulation of the aggregation process [21]. The architecture therefore begins with a task-specific threat model that identifies protected assets, plausible adversaries, permitted information flows, expected participant behavior, and residual risks. Privacy risk assessment should cover raw records, local features, gradients, parameter updates, summary statistics, model outputs, logs, and participation metadata rather than focusing only on whether source datasets leave the institution.

Integrity risks require controls that are distinct from confidentiality protections. A compromised or unreliable participant could submit corrupted, strategically manipulated, or technically incompatible updates that degrade performance, create hidden behaviors, or shift the model toward an institutional interest. Federated-learning security taxonomies show that poisoning and related attacks can target data, local training, model updates, or aggregation procedures [22]. The proposed architecture consequently includes authenticated participants, signed and versioned updates, predefined eligibility checks, update-distribution monitoring, anomaly review, quarantine procedures, rollback capability, and investigation pathways. Robust aggregation may reduce the effect of selected abnormal updates, but it cannot establish that every contribution is trustworthy or that sophisticated manipulation is absent.

Privacy-enhancing technologies should be selected according to the specific data, task, adversary model, and acceptable loss of scientific utility. Secure aggregation may reduce coordinator access to individual updates, encryption may protect transmission or selected computations, and differential privacy may constrain information attributable to a defined privacy unit under stated assumptions. Privacy-preserving federated-learning scholarship emphasizes that these controls remain dependent on implementation quality, governance context, access conditions, and the precise threat model [23]. Differential privacy may also affect predictive utility differently across institutions, chemical classes, or underrepresented endpoints. Privacy and fairness must therefore be assessed jointly because a protection mechanism that lowers average accuracy slightly may produce larger errors for smaller sites or rare data distributions [24].

Technical controls require a governance structure that assigns responsibility for consortium membership, local authorization, model-update approval, incident handling, auditing, validation, publication, and lifecycle decisions. Governance research in federated healthcare indicates that accountability, stakeholder roles, data-use limits, institutional oversight, and model-lifecycle processes are frequently as important as the learning algorithm itself [25]. The proposed architecture therefore requires local institutional review, a consortium governance board, named data stewards, model owners, privacy and security leads, independent validation functions, documented

escalation procedures, and explicit no-privacy-guarantee and no-deployment-readiness boundaries. **Table 2** maps privacy, security, and governance requirements for federated phytochemical discovery.

Table 2. Privacy, Security, and Governance Requirements for Federated Phytochemical Discovery: Sensitive Data Stewardship, Secure Aggregation, Differential Privacy, Encryption, Access Control, Audit Logging, Membership Inference Risk, Model Inversion Risk, Gradient Leakage Risk, Poisoning Risk, Bias Assessment, and Model Update Governance

Requirement domain	Core risk question	Required safeguard	Architecture function	Failure risk	Validation or audit need	Responsible actor	Claim boundary
Sensitive data stewardship	Who controls local phytochemical, assay, omics, safety, provenance, proprietary, or culturally sensitive data?	Named local steward, data classification, permitted-use rules, and exclusion criteria	Maintains institutional responsibility while enabling approved local computation	Unauthorized use, poor curation, or ambiguous accountability	Local stewardship review and periodic eligibility audit	Participating institution and designated data steward	Participation does not transfer ownership or authority to the consortium
Privacy risk assessment	What information could be inferred from records, updates, outputs, logs, or participation?	Threat model, data-flow map, adversary assumptions, and residual-risk register	Determines which privacy controls are required for the approved task	Unrecognized leakage pathways and false reassurance	Prelaunch and recurring privacy review	Consortium privacy lead and local privacy officers	Assessment reduces uncertainty but does not establish privacy
Secure aggregation	Can the coordinating service inspect individual institutional updates?	Approved secure-aggregation mechanism, key management, threshold rules, and failure handling	Limits exposure of individual contributions during aggregation	Collusion, implementation failure, dropout problems, or metadata leakage	Protocol testing and independent security review	Security engineering team and external assessor where appropriate	Secure aggregation does not prevent all leakage, poisoning, or misuse
Differential privacy where appropriate	Can the influence of a defined record or participant be bounded without unacceptable scientific loss?	Defined privacy unit, clipping policy, calibrated perturbation, and privacy accounting	Adds controlled noise to selected updates or outputs	Utility loss, weak protection, or disproportionate site-level effects	Privacy accounting, attack testing, and fairness analysis	Privacy lead, model owner, and validation team	Differential privacy is conditional and not a universal guarantee
Encryption where appropriate	Which communications, stored artifacts, or computations require cryptographic protection?	Transport and storage encryption, key rotation, and justified advanced cryptographic controls	Protects selected exchanges and retained artifacts	Key compromise, misconfiguration, side channels, or excessive computational burden	Configuration audit, penetration testing, and key-lifecycle review	Consortium security operations	Encryption protects selected components, not the complete learning system

Access control	Can only authorized personnel and services access local clients, models, updates, and logs?	Least-privilege roles, separation of duties, and time-limited permissions	Restricts administrative and technical actions	Unauthorized access or privilege escalation	Scheduled entitlement review and access-log audit	Local and consortium identity administrators	Authorized access does not establish scientific validity
Authentication	Are participating institutions, services, and update sources genuine?	Strong identity management, mutual authentication, certificates, and credential rotation	Prevents unverified participants and services from joining the workflow	Impersonation, rogue clients, or stolen credentials	Credential and certificate lifecycle audit	Identity and access-management team	Authentication confirms identity, not correctness or trustworthiness of an update
Audit logging	Can actions, changes, approvals, and incidents be reconstructed?	Tamper-evident logs, synchronized timestamps, retention rules, and review procedures	Records access, training, aggregation, configuration changes, and governance decisions	Weak accountability or incomplete incident reconstruction	Log-completeness testing and scheduled review	Independent audit function and system operators	Logs support investigation but do not prevent harm by themselves
Membership inference risk	Could an observer infer whether a compound, assay record, image, or subject contributed to training?	Output minimization, controlled access, regularization, privacy controls, and empirical attack evaluation	Tests and limits disclosure of training participation	Exposure of proprietary compounds or sensitive records	Red-team testing under plausible access assumptions	Privacy and security validation team	Passing selected tests does not prove universal resistance
Model inversion risk	Could sensitive features or representative records be reconstructed from the model or updates?	Restricted outputs, update protection, access controls, and inversion testing	Reduces opportunities to reconstruct chemical, phenotypic, omics, or clinical attributes	Disclosure of proprietary or sensitive features	Scenario-based inversion assessment	ML security team and relevant data stewards	The architecture cannot claim that reconstruction is impossible
Gradient leakage risk	Can local data be reconstructed from gradients or parameter changes?	Secure aggregation, clipping, restricted update detail, and differential privacy	Protects high-information local updates	Reconstruction of rare or distinctive records	Model-specific gradient-leakage testing	ML security team	Risk reduction does not constitute zero-leakage assurance

		where justified					
Poisoning or integrity risk	Can malicious, corrupted, or low-quality updates alter the global model?	Signed updates, validation checks, anomaly detection, robust aggregation, quarantine, and rollback	Screens contributions and protects aggregation integrity	Backdoors, biased predictions, degraded performance, or convergence failure	Simulated integrity tests and continuous update monitoring	Security team, model owner, and local client operators	Robustness controls cannot guarantee the absence of manipulation
Data-use governance	Is every local computation consistent with the approved purpose and authority?	Purpose specification, permitted-use register, retention rules, and withdrawal procedures	Restricts federation to authorized research questions	Function creep, unauthorized secondary use, or violation of community or institutional conditions	Scheduled purpose-and-use audit	Consortium governance board and local data controllers	Technical capability does not create permission to use data
Institutional review	Has each institution independently approved its participation and local processing?	Applicable scientific, ethics, privacy, contractual, and security reviews	Preserves local authority over participation and processing	Inadequate authorization or inconsistent institutional obligations	Approval verification and renewal tracking	Each participating institution	Consortium governance does not replace local review
Bias assessment	Which sites, chemical classes, sources, assays, targets, species, or endpoints are underrepresented?	Representation analysis, error stratification, and distribution-shift diagnostics	Identifies systematic weaknesses in the collaborative model	Biased candidate prioritization or exclusion of rare chemical space	Site-level and subgroup bias report	Validation team and domain experts	Bias assessment does not establish fairness
Fairness assessment	Are model benefits and errors distributed acceptably across relevant sites or contexts?	Context-specific fairness criteria and privacy–fairness trade-off analysis	Evaluates whether aggregate improvement conceals unequal institutional outcomes	Disproportionate error or utility loss for smaller participants	Prespecified fairness review	Consortium fairness lead and affected stakeholders	No universal fairness guarantee is possible
Model update governance	Who may submit, approve, reject, roll back, or retire a model update?	Version control, signed provenance, approval workflow, rollback procedure, and change classification	Controls the lifecycle of local and global model versions	Unauthorized updates, silent drift, or irreproducible model changes	Update-provenance and regression audit	Model governance committee	Aggregated models remain research artifacts until separately validated

Reproducibility	Can an approved run be reconstructed from code, configurations, environments, and local transformation records?	Versioned code, environment manifests, model configuration, and run records	Supports repeatable local and global training	Inability to explain or reconstruct a model version	Independent rerun or configuration reconstruction	Technical operations and validation teams	Reproducibility does not prove generalizability or therapeutic validity
Model documentation	Are intended use, training context, metrics, safeguards, limitations, and status transparent?	Versioned model dossier or model card	Communicates evidence and restrictions across the consortium	Misuse, hidden limitations, or version confusion	Documentation completeness and consistency audit	Model owner and governance committee	Documentation must identify outputs as research-support tools
No privacy-guarantee boundary	Could privacy controls be misrepresented as proof that no information can leak?	Mandatory residual-risk language and prohibited-claim policy	Prevents overstatement in reports, interfaces, and publications	False assurance and unsafe expansion of access	Governance review of communications and documentation	Privacy lead and publication committee	The architecture may reduce selected risks but cannot guarantee privacy
No deployment-readiness boundary	Could feasibility or research performance be interpreted as operational approval?	Readiness levels, independent assessment, and separate authorization gates	Separates conceptual development and research pilots from deployment	Premature operational use and unmanaged risk	Formal readiness and change-impact assessment	Consortium governance board and institutional approvers	Architecture specification does not establish deployment readiness

Proposed federated learning architecture

The proposed architecture begins with consortium formation and an explicit data-governance agreement. Consortium formation defines the research purpose, eligible institutional roles, decision authorities, conflict-management procedures, withdrawal conditions, and the categories of data or knowledge that may or may not enter the federation. The data-governance agreement then specifies local stewardship, approved computations, model and update ownership, access conditions, publication review, incident responsibilities, audit obligations, and lifecycle decision rights. Multimodal federated-platform research indicates that collaboration across heterogeneous biomedical sources requires a shared semantic and structural data model before local information can contribute meaningfully to coordinated learning [26]. For phytochemical discovery, this shared layer would define common representations for compounds, assays, targets, pathways, sources, provenance, omics modalities, ADMET or toxicity endpoints, and uncertainty while retaining locally meaningful extensions.

The second architectural stage comprises local data curation, metadata harmonization, and feature and label alignment. Each institution would retain its source records and construct an approved local training view through versioned procedures for structure standardization, duplicate handling, unit conversion, assay mapping, missingness encoding, target identification, pathway mapping, and quality exclusion. Site-level documentation would describe provenance, composition, preprocessing, known limitations, permitted uses, and unresolved uncertainties. Translational AI research has emphasized that model documentation is necessary to communicate intended use, training context, validation status, performance evidence, limitations, and implementation conditions [27]. The architecture therefore requires a versioned model dossier for every local and global model state, rather than treating documentation as a final publication activity.

Local dataset documentation would complement model documentation. Participating sites should describe how records were collected, generated, curated, transformed,

maintained, and selected without exposing the sensitive records themselves. Structured dataset documentation frameworks support transparent reporting of dataset motivation, composition, collection processes, preprocessing, uses, distribution restrictions, and maintenance responsibilities [28]. Within the proposed architecture, each site would produce a controlled dataset dossier, a data-quality report, and a conformance report showing whether the local data meet the common model, metadata, feature, and label specifications. Non-IID characterization would be performed before training and repeated during the lifecycle to identify changes in chemical-space coverage, assay distributions, endpoints, or missingness that could alter the behavior of the global model.

The federated training stage would connect approved local models through authenticated and governed update exchange. Participating institutions would train locally, generate signed and versioned updates, apply required privacy and security controls, and transfer only permitted artifacts to the aggregation service. The aggregation layer would combine eligible contributions under a predefined policy, generate a candidate global model, and return the approved version for site-specific evaluation. Cross-site validation, bias assessment, fairness review, reproducibility checks, expert pharmacology review, natural product chemistry review, and safety-oriented review would then determine whether the model remains suitable for the stated research purpose. Traditional knowledge and biodiversity-linked information would be subject to an additional eligibility gate because responsible use may depend on collective benefit, authority to control, responsibility, and ethics rather than technical data locality alone [29]. **Table 3** presents the proposed federated learning architecture for cross-institutional phytochemical discovery.

Table 3. Proposed Federated Learning Architecture for Cross-Institutional Phytochemical Discovery: Consortium Formation, Data Governance, Local Data Curation, Metadata Harmonization, Local Training, Secure Update Transfer, Global Aggregation, Cross-Site Validation, Bias Review, Expert Oversight, Lifecycle Governance, Feedback, and Claim Boundaries

Architecture stage	Core purpose	Required input	Federated function	Potential output	Validation need	Failure risk	Feedback mechanism	Decision boundary
Consortium formation	Establish a legitimate scientific collaboration	Research objective, participant capabilities, stakeholder mapping, and role definitions	Defines eligible institutional clients and decision authorities	Consortium charter and scoped collaboration	Capability, representation, conflict, and trust assessment	Misaligned objectives, unequal power, or unsuitable participants	Periodic consortium review	No training before membership and purpose are approved
Data governance agreement	Define authority, permitted use, responsibilities, and withdrawal	Local policies, contractual conditions, ethics requirements, and rights-holder terms	Authorizes bounded local computation and model exchange	Approved governance framework	Institutional, scientific, privacy, and ethics review	Function creep, ownership disputes, or unauthorized reuse	Formal amendment and scheduled review process	Governance agreement does not prove legal sufficiency, privacy, or validity
Local data curation	Produce reliable institution-controlled datasets	Raw local records, provenance, quality criteria, and exclusion rules	Creates approved local training and validation views	Curated local dataset and quality report	Identity, duplicate, provenance, missingness, and outlier review	Poor-quality records or hidden duplication	Curator correction workflow	Critical unresolved defects exclude affected records or tasks
Metadata harmonization	Align scientific meaning and context across sites	Common vocabularies, mappings, units, assay fields, and provenance standards	Maps local metadata into interoperable representations	Harmonized local metadata layer	Semantic and mapping conformance tests	False equivalence or loss of local context	Mapping discrepancy log and terminology updates	Syntactic compatibility alone is insufficient
Feature and label alignment	Ensure that inputs and outcomes are	Feature dictionary, preprocessing	Produces compatible local feature	Conformant modeling matrices	Unit, pipeline, leakage, and	Incompatible updates or	Automated conformance	Training pauses when critical

	technically and scientifically comparable	specification, endpoint definitions, and missingness rules	pipelines and task labels		endpoint checks	misleading aggregate metrics	reports and remapping	definitions diverge
Non-IID characterization	Identify site-level differences before aggregation	Distribution summaries, chemical-space diagnostics, assay and endpoint profiles	Informs aggregation, personalization, weighting, stratification, or task separation	Heterogeneity report and modeling plan	Shift, representation, imbalance, and overlap assessment	Negative transfer or dominance by large sites	Reanalysis after data or participant changes	A single global model is not required when heterogeneity makes it unsuitable
Local model training	Learn from approved local data without exporting source records	Approved dataset, model version, hyperparameters, and local environment	Performs local optimization and generates an approved update	Local update, metrics, and run manifest	Local holdout evaluation and reproducibility check	Overfitting, unstable optimization, or unauthorized use	Local diagnostics and controlled retraining	Local success does not establish cross-site validity
Secure update transfer	Protect permitted collaborative artifacts in transit	Authenticated client, signed update, encryption, and output policy	Transfers governed model information to the aggregation service	Verified update available for review	Identity, integrity, confidentiality, and replay testing	Interception, tampering, metadata leakage, or rogue submission	Failure alerts, quarantine, and key rotation	No update is accepted without integrity and authorization checks
Global aggregation	Combine eligible institutional contributions	Validated updates, aggregation policy, contribution rules, and version controls	Produces a candidate global model state	Aggregated research model and contribution report	Convergence, contribution balance, anomaly, and robustness checks	Site dominance, negative transfer, or poisoned aggregation	Diagnostics, rollback, and revised aggregation policy	Aggregation is not scientific validation
Privacy and security safeguards	Reduce defined confidentiality, privacy, and integrity risks	Threat model, data classification, safeguard selection, and risk thresholds	Applies secure aggregation, encryption, differential privacy, access control, and monitoring where justified	Protected workflow with documented residual risk	Protocol, configuration, attack, and utility testing	Control failure, utility loss, or false assurance	Continuous monitoring and control recalibration	No privacy or security guarantee is permitted
Cross-site validation	Test performance in each institutional context	Prespecified test sets, metrics, uncertainty criteria, and subgroup definitions	Executes validation locally and combines approved summaries	Site-wise and aggregate validation dossier	Calibration, discrimination, uncertainty, error, and shift review	Aggregate success concealing local failure	Failure analysis and model revision	Failure of minimum site criteria prevents advancement
Bias and fairness review	Identify systematic representation and	Site, chemical-class, assay, source, species, and	Compares errors and model benefit across	Bias and fairness report	Metric suitability and privacy–fairness	Underrepresentation or unequal institutional utility	Data improvement, reweighting, personalizatio	Federation does not automatically

	performance disparities	endpoint summaries	relevant contexts		trade-off analysis		n, or scope restriction	create fairness
Expert pharmacology and natural product review	Assess biological plausibility, evidence quality, and domain limitations	Predictions, uncertainty, assay context, target and pathway evidence, and provenance	Adds structured human interpretation to computational outputs	Expert-reviewed research hypotheses	Expertise, conflict, traceability, and consistency review	Automation bias or mechanistic overinterpretation	Reviewer feedback and evidence requests	Expert review does not validate efficacy or safety
Candidate prioritization boundary	Restrict how model scores are used in discovery decisions	Ranked outputs, applicability domain, uncertainty, and supporting evidence	Produces a bounded research-prioritization list	Hypotheses for independent experimental evaluation	Stability, calibration, novelty, and evidence-coherence review	Treating predictions as proof of bioactivity, safety, or therapeutic value	Experimental feedback and reprioritization	Ranking is not candidate validation
Model documentation	Create an auditable account of each model version	Data summaries, code and environment versions, metrics, safeguards, and limitations	Packages technical, validation, governance, and intended-use information	Versioned model dossier	Completeness, consistency, and provenance audit	Hidden changes, misuse, or version confusion	Update documentation with every approved change	Undocumented models cannot advance
Lifecycle update governance	Control retraining, participant changes, drift, rollback, and retirement	Monitoring results, new data, incidents, and change requests	Authorizes controlled evolution of the federated system	Approved update, rollback, restriction, or retirement decision	Regression and change-impact testing	Silent drift, incompatible clients, or loss of reproducibility	Scheduled and incident-triggered review	Material changes require renewed validation
Feedback updating	Incorporate verified experimental or expert evidence into later versions	Curated experimental results, corrected metadata, and expert adjudication	Updates local labels or features through an approved process	Revised evidence base and future model version	Provenance, independence, temporal leakage, and label-quality checks	Circular validation or propagation of erroneous findings	Feedback ledger and curator review	Feedback enters only after quality and governance approval
No therapeutic-claim boundary	Prevent predictions from being represented as therapeutic evidence	Claim taxonomy, review checklist, and communication policy	Constrains interpretation of model outputs	Clearly bounded research statements	Report, interface, publication, and decision-log audit	Unsupported efficacy, safety, or clinical claims	Corrective review and withdrawal of overstated claims	The architecture supports research prioritization only
No privacy-guarantee boundary	Prevent safeguards from being represented as eliminating privacy risk	Residual-risk statement, threat-model scope, and prohibited-claim language	Embeds caution into documentation and governance	Transparent communication of remaining risks	Review of reports, interfaces, agreements, and publications	False reassurance or unsafe expansion of access	Reassessment after incidents, attacks, or architectural changes	Data locality and safeguards reduce selected risks but do not prove zero risk

The architecture operates as a controlled feedback system rather than as a linear pipeline ending with a ranked compound list. Validation failures, security findings, expert concerns, metadata discrepancies, distribution drift, or newly generated experimental evidence would return to the appropriate local curation, harmonization, training, or governance stage. No model output would cross the candidate-prioritization boundary without site-wise validation, uncertainty reporting, model documentation, and expert review. **Figure 2** presents the proposed

federated learning architecture for cross-institutional phytochemical discovery, integrating local stewardship, harmonization, federated training, privacy safeguards, validation, governance, expert review, and lifecycle feedback.

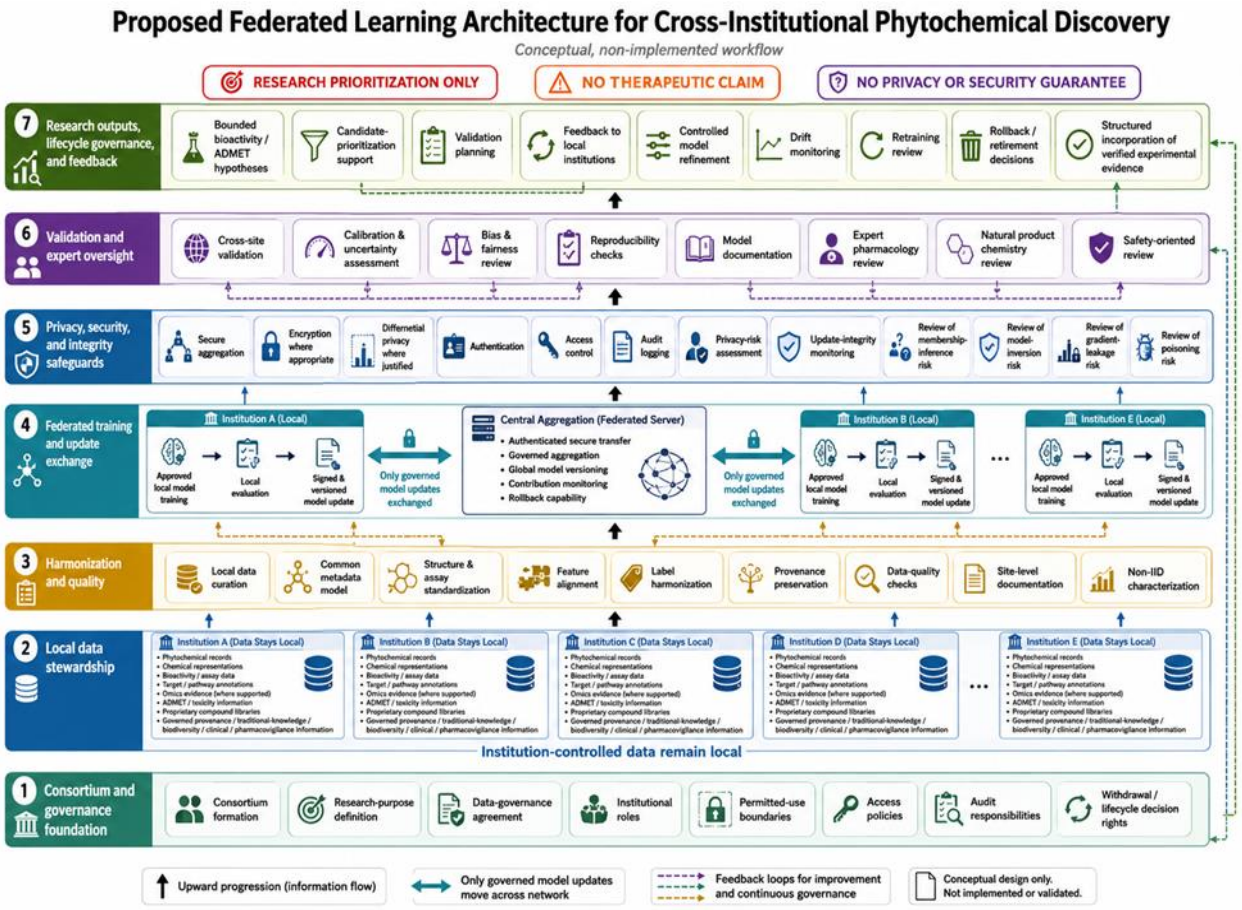

Figure 2. Proposed Federated Learning Architecture for Cross-Institutional Phytochemical Discovery

Alt-text description

A conceptual federated learning architecture showing consortium governance, local phytochemical data curation, harmonized metadata, local training, secure update transfer, global aggregation, validation, privacy and security review, expert oversight, lifecycle governance, and feedback loops.

Implementation implications

Implementation would require more institutional preparation than installing a common federated-learning package. Academic consortia, natural product laboratories, pharmaceutical partners, clinical pharmacology groups, and bioinformatics teams would first need to agree on scientific tasks, local eligibility rules, metadata requirements, endpoint definitions, structure-standardization procedures, validation partitions, and acceptable outputs. Multicohort federated research using a common data model demonstrates that local mapping, terminology alignment, extract-transform-load procedures, data-quality testing, and cross-site coordination can constitute a substantial part of the work required before collaborative learning begins [30]. Phytochemical consortia should therefore budget for data stewardship, ontology development, assay harmonization, provenance correction, documentation, and local conformance testing rather than treating these activities as incidental preprocessing.

A bounded feasibility phase would be appropriate before any broader research use. Such a phase could test whether each institution can execute the approved local pipeline, generate compatible features and labels, authenticate securely, transfer permitted updates, reproduce the intended model version, and return local validation summaries without exposing restricted information. Feasibility research over a distributed common-data-model network supports the value of testing local transformation, distributed execution, communication, and result consistency before inferring operational readiness [31]. For phytochemical discovery, feasibility criteria should include structure and assay conformance, non-IID diagnostics, update integrity, site-level model behavior,

documentation completeness, expert-review procedures, and the ability to suspend or roll back the workflow when validation or security conditions are not met.

Institutional trust would depend on continuing oversight rather than a one-time technical demonstration. Data stewards would monitor local eligibility and quality; privacy and security teams would review threat assumptions and controls; ethics committees or equivalent review bodies would assess sensitive data uses where applicable; governance boards would authorize model changes and resolve disputes; domain experts would evaluate biological and chemical plausibility; and funding bodies would need to support shared infrastructure, harmonization, auditing, and long-term maintenance. These implications describe a research implementation pathway, not legal, cybersecurity, regulatory, clinical, or deployment advice. Any operational transition would require separate context-specific assessment, authorization, validation, security testing, and institutional approval beyond the conceptual architecture presented here.

CONCLUSION

This article proposes a federated learning architecture for cross-institutional phytochemical discovery that connects local data stewardship with harmonized collaborative model development, governed update exchange, privacy and security safeguards, cross-site validation, bias and fairness review, expert oversight, model documentation, lifecycle governance, and structured feedback. Its central premise is that retaining sensitive data within participating institutions can support collaboration where centralization is inappropriate or impractical, but data locality does not automatically ensure privacy, security, fairness, scientific validity, therapeutic relevance, or institutional trust. Responsible use requires explicit management of heterogeneous chemical and biological data, realistic non-IID conditions, privacy leakage and integrity risks, common metadata and endpoint definitions, site-level validation, reproducibility, and carefully controlled interpretation. The proposed architecture therefore positions federated learning as an enabling component of collaborative research rather than as an autonomous discovery system, privacy guarantee, security certification, therapeutic-validation mechanism, or deployment-ready platform.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. doi:10.1038/s41746-020-00323-1

[2] Hanser T. Federated learning for molecular discovery. Curr Opin Struct Biol. 2023;79:102545. doi:10.1016/j.sbi.2023.102545

[3] Smajić A, Grandits M, Ecker GF. Privacy-preserving techniques for decentralized and secure machine learning in drug discovery. Drug Discov Today. 2023;28(12):103820. doi:10.1016/j.drudis.2023.103820

[4] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11. doi:10.1038/s42256-020-0186-1

[5] Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model. 2024;64(7):2331-44. doi:10.1021/acs.jcim.3c00799

[6] Zhu W, Luo J, White AD. Federated learning of molecular properties with graph neural networks in a heterogeneous setting. Patterns (N Y). 2022;3(6):100521. doi:10.1016/j.patter.2022.100521

[7] Simm J, Humbeck L, Zalewski A, Sturm N, Heyndrickx W, Moreau Y, et al. Splitting chemical structure data sets for federated privacy-preserving machine learning. J Cheminform. 2021;13(1):96. doi:10.1186/s13321-021-00576-2

[8] Huang D, Ye X, Zhang Y, Sakurai T. Collaborative analysis for drug discovery by federated learning on non-IID data. Methods. 2023;219:1-7. doi:10.1016/j.ymeth.2023.09.001

[9] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Trans Intell Syst Technol. 2019;10(2):12:1-12:19. doi:10.1145/3298981

[10] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50-60. doi:10.1109/MSP.2020.2975749

[11] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. 2021;5(1):1-19. doi:10.1007/s41666-020-00082-4

[12] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine:

Facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. doi:10.1038/s41598-020-69250-1

[13] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59-67. doi:10.1016/j.ijmedinf.2018.01.007

[14] Lee GH, Shin SY. Federated learning on clinical benchmark data: Performance assessment. J Med Internet Res. 2020;22(10):e20891. doi:10.2196/20891

[15] Ju L, Hellander A, Spjuth O. Federated learning for predicting compound mechanism of action based on image-data from Cell Painting. Artif Intell Life Sci. 2024;5:100098. doi:10.1016/j.ailsci.2024.100098

[16] Huang D, Ye X, Sakurai T. Multi-party collaborative drug discovery via federated learning. Comput Biol Med. 2024;171:108181. doi:10.1016/j.compbiomed.2024.108181

[17] Cozac R, Hasic H, Choong JJ, Richard V, Beheshti L, Froehlich C, et al. kMoL: An open-source machine and federated learning library for drug discovery. J Cheminform. 2025;17(1):22. doi:10.1186/s13321-025-00967-9

[18] Nguyen DC, Pham QV, Pathirana PN, Ding M, Seneviratne A, Lin Z, et al. Federated learning for smart healthcare: A survey. ACM Comput Surv. 2023;55(3):60:1-60:37. doi:10.1145/3501296

[19] Antunes RS, da Costa CA, Küderle A, Abdullahi Yari I, Eskofier BM. Federated learning for healthcare: Systematic review and architecture proposal. ACM Trans Intell Syst Technol. 2022;13(4):54:1-54:23. doi:10.1145/3501813

[20] Truhn D, Tayebi Arasteh S, Saldanha OL, Müller-Franzes G, Khader F, Quirke P, et al. Encrypted federated learning for secure decentralized collaboration in cancer image analysis. Med Image Anal. 2024;92:103059. doi:10.1016/j.media.2023.103059

[21] Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. Future Gener Comput Syst. 2021;115:619-40. doi:10.1016/j.future.2020.10.007

[22] Liu P, Xu X, Wang W. Threats, attacks and defenses to federated learning: Issues, taxonomy and perspectives. Cybersecurity. 2022;5:4. doi:10.1186/s42400-021-00105-6

[23] Truong N, Sun K, Wang S, Guitton F, Guo YK. Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Comput Secur. 2021;110:102402. doi:10.1016/j.cose.2021.102402

[24] Chen H, Zhu T, Zhang T, Zhou W, Yu PS. Privacy and fairness in federated learning: On the perspective of tradeoff. ACM Comput Surv. 2024;56(2):39:1-39:37. doi:10.1145/3606017

[25] Eden R, Chukwudi I, Bain C, Barbieri S, Callaway L, de Jersey S, et al. A scoping review of the governance of federated learning in healthcare. NPJ Digit Med. 2025;8(1):427. doi:10.1038/s41746-025-01836-3

[26] Cremonesi F, Planat V, Kalokyri V, Kondylakis H, Sanavia T, Mateos Resinas VM, et al. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. J Biomed Inform. 2023;141:104338. doi:10.1016/j.jbi.2023.104338

[27] Brereton TA, Malik MM, Lifson M, Greenwood JD, Peterson KJ, Overgaard SM. The role of artificial intelligence model documentation in translational science: Scoping review. Interact J Med Res. 2023;12:e45903. doi:10.2196/45903

[28] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723

[29] Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE Principles for Indigenous Data Governance. Data Sci J. 2020;19(1):43. doi:10.5334/dsj-2020-043

[30] Mateus P, Moonen J, Beran M, Jaarsma E, van der Landen SM, Heuvelink J, et al. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: A Netherlands consortium of dementia cohorts case study. J Biomed Inform. 2024;155:104661. doi:10.1016/j.jbi.2024.104661

[31] Lee GH, Park J, Kim J, Kim Y, Choi B, Park RW, et al. Feasibility study of federated learning on the distributed research network of OMOP common data model. Healthc Inform Res. 2023;29(2):168-73. doi:10.4258/hir.2023.29.2.168