

ORIGINAL RESEARCH

Open access

Position: Federated Learning for Proprietary Materials Databases — Privacy-Preserving Training without Centralized Data

Samuel Boateng^{1*}, Kwesi Mensah¹, Kojo Asante², Linda Owusu¹

Abstract

Federated learning represents a paradigm-shifting solution for unlocking proprietary materials databases in artificial intelligence-driven materials science. This position paper asserts that federated learning enables privacy-preserving collaborative training across decentralized sites—industry laboratories, corporate R&D facilities, and academic institutions—without ever centralizing sensitive raw data. By allowing each participant to train locally on its own proprietary datasets while sharing only model updates, federated learning overcomes the longstanding impasse created by competitive, legal, and privacy barriers that have historically prevented meaningful data sharing in the field. Industry holds vast, high-value proprietary materials databases encompassing real-world processing conditions, rare compositions, failure modes, and multi-fidelity experimental results that are simply unavailable in public repositories. Academic models, by contrast, remain constrained by biased, narrow public datasets that limit generalization and predictive power. Federated learning bridges this divide, delivering three core benefits: genuine privacy preservation that satisfies export-control regulations and intellectual-property safeguards; unprecedented data diversity that spans industrial-scale variability; and competitive pre-training of foundation models that every participant can subsequently fine-tune for proprietary tasks. Despite acknowledged challenges—heterogeneous data distributions across clients, communication overhead, and residual security risks—the technology has already demonstrated feasibility in closely related domains such as pharmaceutical chemistry and additive manufacturing. This perspective argues that the materials science community must now prioritize infrastructure investment, standardized protocols, and cross-sector consortia to realize federated learning's potential. Only through deliberate collective action can the field harness the full power of proprietary data while preserving each organization's competitive edge and legal compliance. The time for pilot projects and community standards is immediate.

Keywords Materials informatics, Federated learning, Privacy-preserving machine learning, Proprietary materials databases, Data sharing barriers, Distributed training

*Correspondence:

Samuel Boateng
samuel.boateng@gmail.com

¹ Department of Computational Materials Engineering, Faculty of Engineering, University of Ghana, Accra, Ghana

² Department of Materials Data Science, Faculty of Science and Technology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Introduction

The position

Proprietary materials databases held by industry are a vast, untapped resource for machine learning. But legal, competitive, and privacy concerns prevent data sharing. Federated learning offers a solution: train models across multiple

sites without ever centralizing data. Each site trains locally; only model updates are shared. This position paper argues that federated learning can unlock proprietary data for materials ML, enabling larger, more diverse training sets while preserving privacy and competitive advantage.

Figure 1 presents the hierarchical logic of the argument, showing how proprietary data barriers in materials science are translated through federated-learning infrastructure into shared model gains, managed risk controls, and phased field-level implementation.

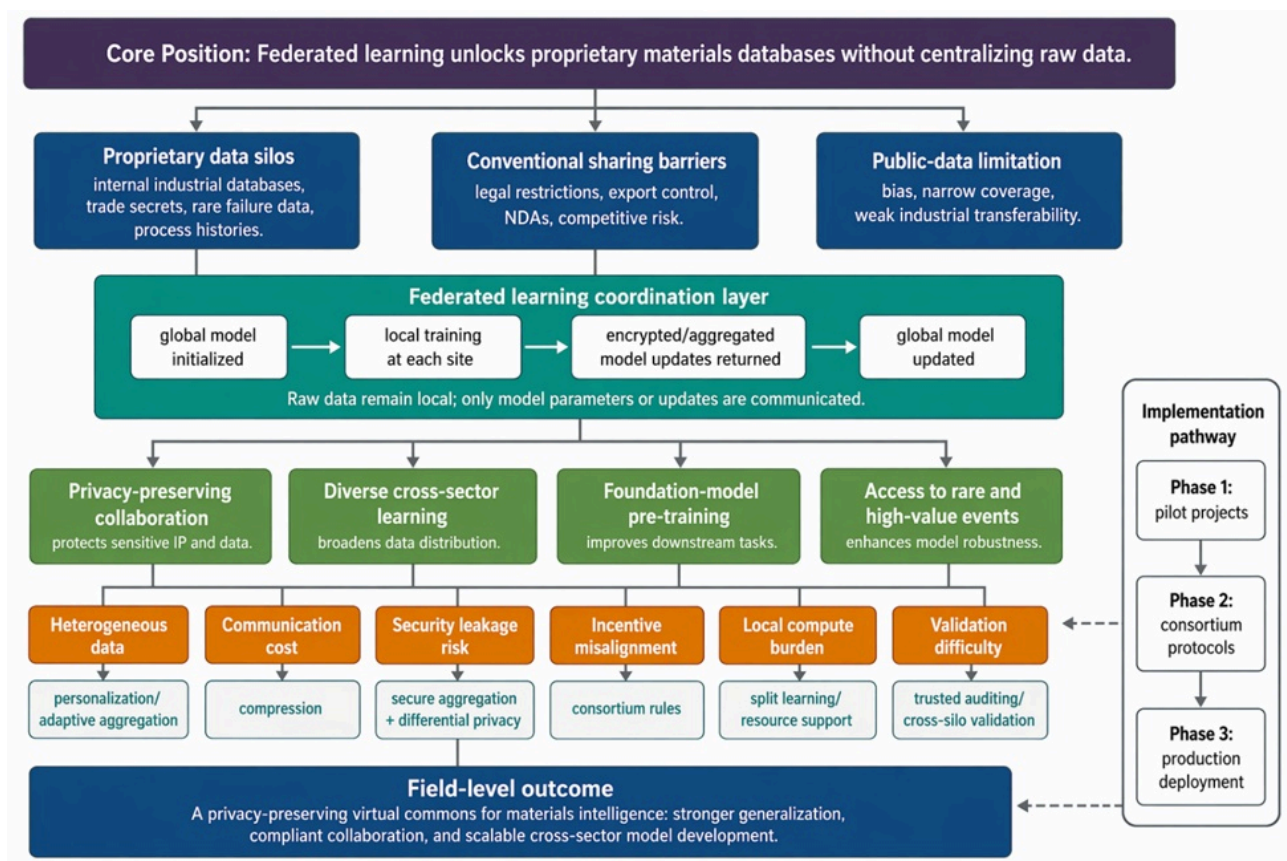


Figure 1. Hierarchical architecture of federated learning as a privacy-preserving coordination model for proprietary materials databases

The materials science community stands at a critical juncture. Machine-learning models for property prediction, materials discovery, and process optimization have become unmistakably data-hungry. Yet the most valuable data—detailed experimental records, proprietary synthesis routes, long-term performance logs, and failure-case analyses—reside almost exclusively behind corporate firewalls. Public databases, while invaluable, suffer from well-documented biases toward easily synthesized compounds, common elements, and idealized laboratory conditions [1-3]. Academic models trained exclusively on such data therefore exhibit limited transferability to real-world industrial settings. Industry, meanwhile, develops powerful but narrowly scoped models that cannot benefit from the breadth of knowledge distributed across competitors and research partners.

This position matters now because the gap between available public data and the scale required for next-generation foundation models is widening rapidly. Recent analyses of data-driven materials science have repeatedly highlighted how dataset bias and limited diversity constrain model performance and slow innovation [2-4]. Federated learning directly addresses this structural problem by enabling collaborative training without any transfer of raw data [5]. It is not a theoretical proposal; practical implementations already exist in pharmaceutical research and additive manufacturing,

where cross-organizational model training has produced measurable gains in predictive accuracy while fully respecting data sovereignty [6-9].

The core mechanism is straightforward yet powerful: a central coordinating server initializes a global model and distributes it to participating clients. Each client—whether a battery manufacturer, a specialty chemicals firm, or a national laboratory—performs local training on its proprietary database. Only the resulting model updates (gradients or weight deltas) return to the server, where they are aggregated into an improved global model. No raw data ever leaves its original secure environment. Variants such as secure aggregation and differential privacy further harden the system against potential information leakage [10, 11].

By adopting this approach, the materials science community can finally integrate the complementary strengths of industrial and academic data ecosystems. The result will be models that generalize across broader composition spaces, processing conditions, and operating regimes—precisely the capability needed to accelerate discovery of novel materials for energy storage, catalysis, aerospace, and sustainable manufacturing [12-15]. This paper therefore takes a clear stance: federated learning is not merely an incremental technical improvement but a necessary enabler for the next decade of progress in computational and data-driven materials engineering [16, 17]. The community must move from discussion to implementation, beginning with targeted pilot projects and culminating in shared infrastructure that respects the realities of proprietary data ownership.

The Data Sharing Barrier in Materials

Data sharing in materials science and chemistry has long been recognized as exceptionally difficult. Proprietary databases constitute genuine competitive assets. A company's internal repository of experimental outcomes on solid electrolytes, for example, encodes years of costly trial-and-error, optimized process parameters, and trade-secret formulations that directly influence product performance and market position. Releasing such data—even in anonymized form—risks eroding that advantage.

Legal and regulatory restrictions compound the problem. Export-control regulations, International Traffic in Arms Regulations (ITAR), and corporate non-disclosure agreements frequently prohibit cross-border or even inter-company data transfers. Privacy concerns extend beyond intellectual property; detailed process parameters can inadvertently reveal manufacturing know-how or supply-chain vulnerabilities. Intellectual-property ownership of any jointly trained model adds further complexity: who owns the resulting weights when multiple organizations contribute data? These barriers are not abstract; they are routinely cited by both industrial and academic stakeholders as primary obstacles to broader adoption of machine learning in materials [2].

The consequences of non-sharing are severe and self-reinforcing. Academic research groups rely almost exclusively on open repositories that, while large, are biased toward computationally accessible or historically popular materials [3, 18]. Industry models, trained solely on internal data, achieve high accuracy within narrow domains but fail to extrapolate to new chemistries or processing routes. Neither side benefits from the other's strengths. The combined dataset that could emerge from cross-sector collaboration—spanning rare-earth-free magnets, high-entropy alloys, solid-state battery formulations, and sustainable polymers—remains fragmented and underutilized.

Table 1 clarifies why federated learning is not simply another data-sharing option but a distinct governance model that combines distributed data control with collective model formation.

Table 1. Comparative governance logics for learning from materials data: isolated local modeling, centralized sharing, open repositories, and federated learning

Governance model	Raw data movement	Protection of proprietary know-how	Ability to capture cross-organization diversity	Legal/regulatory feasibility	Model generalizability	Coordination burden	Strategic impact for non-science
Isolated local modeling	None	Very high	Very low	Very high	Low to moderate; domain-constrained	Low	Prevents cross-organizational fragmentation and extra
Centralized inter-organizational data pooling	Full transfer to shared repository	Low	Very high	Low in many industrial settings	High in principle	Very high	Technically attractive, often hindered by IP constraints and barriers
Open repositories only	Public release	Low to moderate	Moderate, but skewed toward publishable/open data	Moderate	Moderate; often biased toward idealized or historically popular systems	Moderate	Supports reproducibility but access to high-quality data is distributed
Federated learning	No raw data transfer; only model updates	High	High	High relative to centralized pooling	High; improved breadth without direct sharing	Moderate to high initial setup, lower recurring sharing friction	Reduces collaboration friction, excludes domain knowledge, integrates

This fragmentation represents a profound missed opportunity. Rare materials, extreme-condition data, and long-term degradation studies are statistically scarce within any single organization. Pooling such information, even indirectly, would dramatically improve model robustness and enable extrapolation into previously inaccessible regimes. Federated learning resolves the impasse by eliminating the need for raw-data exchange. Participants retain full control of their databases, yet the global model learns from the collective statistical distribution. Early demonstrations in related chemical domains have already shown that federated approaches can match or exceed the performance of centralized training while satisfying every legal and competitive constraint [8, 9].

The barrier is therefore not technological immaturity but institutional inertia and outdated assumptions about data collaboration. Industry perceives data sharing as zero-sum; federated learning reframes it as positive-sum. Academia recognizes the limitations of public datasets but lacks mechanisms to incorporate proprietary knowledge safely. By explicitly acknowledging these realities and proposing a practical, privacy-first alternative, this position paper reframes the data-sharing problem from an insurmountable obstacle into a solvable coordination challenge. The materials community now possesses both the motivation and the technical means to act.

How Federated Learning Works (For Materials Scientists)

Federated learning is a distributed machine-learning paradigm in which models are trained across multiple decentralized clients—companies, laboratories, or research groups—each holding local proprietary data, without any exchange of the underlying datasets [16, 17, 19, 20]. The defining characteristic is that raw data never leaves its original secure environment.

The basic federated learning loop proceeds as follows. A central server first initializes a global model architecture—commonly a graph neural network suited to crystal structures or a transformer adapted for composition-property mapping. This initial model is broadcast to all participating clients. Each client then trains the received model on its own local proprietary database using standard optimization techniques. After a fixed number of local epochs, the client computes and returns only the model updates (gradients or weight differences) to the server. The server aggregates these updates—most commonly via the FedAvg algorithm, a weighted average based on the size of each client's dataset—and produces an improved global model. The process repeats for dozens to hundreds of communication rounds until convergence.

Crucially, all communication can be protected by secure aggregation protocols and differential privacy mechanisms, ensuring that individual updates reveal no identifiable information about local data. Materials scientists need not master the cryptographic details; existing open-source frameworks already handle these safeguards transparently [10, 11, 21].

Two variants are especially relevant to materials science. Cross-silo federated learning involves a small number of clients (typically 3–10 large organizations), each contributing massive, high-quality datasets. This setting matches the industrial landscape where a handful of battery manufacturers or specialty-alloy producers hold complementary proprietary repositories [8]. Cross-device federated learning, by contrast, envisions many smaller clients—individual academic laboratories or smaller enterprises—each contributing modest but diverse datasets. Hybrid approaches are also feasible.

Materials data introduce realistic heterogeneity: clients may focus on different chemical families (oxides versus sulfides), temperature regimes, or characterization methods. Standard federated learning algorithms were originally designed under the assumption of independent and identically distributed data, yet materials datasets are inherently non-i.i.d. Recent advances in heterogeneous federated learning—personalized models, client clustering, and adaptive aggregation—directly address this challenge [16, 20].

It is equally important to clarify what federated learning does not solve. It does not reduce the local computational burden; each client must still train a full-scale model on its data. Communication costs remain non-trivial for large neural architectures. Heterogeneity must be actively managed rather than ignored. Nevertheless, these limitations are engineering problems with known mitigation strategies, not fundamental barriers.

For materials scientists accustomed to centralized databases, the shift to federated learning requires only a change in workflow mindset: data stay put, knowledge travels in the form of aggregated model intelligence. The approach has already been demonstrated successfully in molecular property prediction, chemical structure handling, and additive manufacturing quality control [6, 22–24]. The same principles translate directly to crystal-graph networks, composition-based descriptors, and multi-fidelity experimental records that define modern materials informatics.

Benefits for Proprietary Materials Databases

Federated learning delivers six interlocking benefits that collectively justify its adoption for proprietary materials databases.

First, privacy preservation is absolute at the data level. Raw experimental records, synthesis protocols, and proprietary process parameters remain entirely within each organization's secure infrastructure. Competitors collaborate without

exposing trade secrets. Legal compliance with export controls, ITAR, GDPR, and corporate NDAs is maintained because no data transfer occurs [21, 25].

Second, data diversity increases dramatically. Models trained through federation incorporate composition spaces, processing histories, and performance metrics drawn from multiple industrial partners and academic sources. This breadth mitigates the well-documented biases inherent in public repositories and produces models that generalize across realistic industrial variability [1-3].

Third, competitive pre-training becomes feasible. A shared foundation model can be trained on the union of all proprietary datasets without any party revealing its data. Each participant then fine-tunes the global model on its own specialized tasks—yielding higher accuracy for internal applications while retaining full ownership of downstream models.

Fourth, access to rare data improves. Single organizations rarely encounter enough examples of exotic materials, failure modes, or extreme operating conditions to train robust predictors. Federated aggregation pools these scarce events statistically, enabling reliable extrapolation without centralized storage.

Fifth, continuous learning emerges naturally. As new proprietary data are generated daily in production environments, clients can periodically contribute fresh updates to the global model. The system evolves without requiring historical data migration or repeated negotiations.

Sixth, regulatory compliance is simplified. Data never cross jurisdictional boundaries or leave approved secure environments, satisfying the strictest legal and ethical standards.

A concrete scenario illustrates the power of these benefits. Three competing battery manufacturers each maintain proprietary databases on solid-electrolyte performance. None can share raw cycling data or formulation details. Through federated learning they jointly train a global ionic-conductivity predictor. Each company subsequently fine-tunes the shared model on its specific chemistry and cell architecture. All three obtain superior predictive capability; none compromises its intellectual property. Similar collaborations have already succeeded in pharmaceutical quantitative structure–activity relationship modeling and additive manufacturing process qualification, demonstrating that the same workflow scales to materials property prediction [6, 8, 9].

Collectively, these benefits transform proprietary databases from isolated corporate assets into a virtual, privacy-preserving commons that accelerates the entire field. The competitive advantage of any single participant is not diminished; rather, every participant gains from the collective intelligence while safeguarding its core data.

Challenges and Open Problems

Federated learning, while promising, introduces several technical and organizational challenges that the materials community must address proactively.

The first challenge is heterogeneous data distributions. Industrial partners naturally focus on different material classes, characterization protocols, and performance metrics. Standard aggregation algorithms can suffer from client drift when local data distributions diverge sharply. Personalized federated learning, client clustering, and adaptive weighting schemes offer practical remedies, but materials-specific implementations—accounting for crystal symmetry and periodic boundary conditions—require further algorithmic development [16, 20].

Communication cost constitutes the second challenge. Training large graph neural networks or transformers demands hundreds of communication rounds, each transmitting megabytes to gigabytes of updates. Techniques such as gradient compression, quantization, and one-shot or few-shot federated learning can reduce overhead substantially, yet bandwidth and latency remain practical considerations for smaller academic partners.

Security and privacy risks form the third challenge. Although raw data never leave the client, model updates can theoretically leak information through gradient inversion or membership inference attacks. Differential privacy, secure multi-party computation, and homomorphic encryption provide strong theoretical protections, but their computational overhead must be balanced against model utility [10, 11, 21, 25].

Incentive structures represent the fourth challenge. Companies may hesitate to contribute model updates if they perceive others as free-riders. Consortium agreements that reward participation proportionally to data contribution, combined with transparent auditing of model improvements, can align interests.

Computational cost is the fifth challenge. Each client must train a full model locally, potentially straining smaller organizations. Split learning, where clients train only the lower layers of a model while the server handles higher-level representations, offers a resource-efficient alternative.

Finally, validation and testing pose a sixth challenge. Without access to proprietary test sets, assessing the global model's true performance is difficult. Cross-silo validation protocols and trusted third-party auditing mechanisms provide workable solutions.

These challenges are not insurmountable; they are active research frontiers already being tackled in adjacent fields. The materials science community can accelerate progress by prioritizing materials-tailored algorithms, standardized benchmarks, and open-source toolkits designed explicitly for crystal-structure data and non-i.i.d. experimental distributions [22, 26]. Addressing them head-on will convert federated learning from a promising concept into a robust, production-ready platform for proprietary materials collaboration.

Table 2 consolidates the paper's central implementation claim by linking each major challenge in federated learning to its technical remedy, governance requirement, and expected field-level payoff.

Table 2. Challenge–mechanism–governance alignment for federated learning in proprietary materials informatics

Challenge in materials federated learning	Why the challenge is especially acute in materials science	Technical mechanism	Governance or organizational requirement	Expected payoff if resolved
Non-i.i.d. client data / client drift	Firms and labs specialize in different chemistries, process windows, instruments, and fidelity levels	Personalized FL, client clustering, adaptive aggregation, domain-aware weighting	Agreement on metadata standards and task comparability	Better cross-domain generalization without forcing artificial data uniformity
High communication overhead	Large graph neural networks, transformers, and repeated communication rounds can strain collaborators	Compression, quantization, fewer rounds, partial update sharing	Infrastructure support and minimum compute/network commitments across partners	Scalable collaboration with lower participation friction
Residual privacy leakage from updates	Even without raw data transfer, update-level inference may expose sensitive process signatures	Secure aggregation, differential privacy, encrypted update pipelines	Auditable privacy policies, third-party verification, compliance documentation	Greater trust, stronger legal defensibility, broader industry participation

Weak participation incentives / fear of free-riding	Firms may fear that competitors benefit disproportionately from their proprietary databases	Contribution-aware reward schemes, performance attribution, access-tiered model benefits	Consortium contracts, transparent rules on model use and benefit distribution	More stable long-term participation and stronger collective investment
Unequal local computational capacity	Large firms and small labs differ sharply in hardware, engineering support, and ML readiness	Split learning, lightweight clients, resource-adaptive training	Shared tooling, common onboarding support, grant-backed participation	Inclusion of smaller but scientifically valuable data holders
Validation without pooled test sets	Benchmarking is difficult when no party can reveal proprietary holdout data	Cross-silo validation, secure evaluation, trusted auditing nodes	Pre-agreed evaluation protocol and neutral oversight	Credible evidence of model quality for both science and deployment
Ambiguous ownership of resulting models	Jointly trained models can blur rights over weights, derivatives, and downstream fine-tuning	Modular model governance, checkpoint access controls, license structures	Clear IP agreements defining base-model rights and local fine-tuning rights	Reduced legal ambiguity and faster transition from pilot to production

Objections and Responses

A recurrent concern from industry leaders is the perception that proprietary data constitutes an irreducible competitive asset that cannot be exposed in any form. Federated learning directly circumvents this tension by exchanging only model updates rather than raw datasets, where these updates remain inherently noisy, aggregated, and can be further protected through differential privacy, making reconstruction of proprietary information practically infeasible [8-11]. Under this paradigm, organizations preserve full ownership of their databases while still benefiting from the collective representational capacity of a shared global model.

Closely related concerns often target efficiency, with the claim that federated learning is prohibitively slow and resource-intensive. While communication overhead is non-trivial, it must be weighed against the recurring burden of centralized data curation, legal coordination, and repeated retraining cycles. In cross-silo materials science settings, where participant numbers remain limited, compression strategies substantially mitigate bandwidth constraints, shifting the balance toward feasibility rather than inefficiency [6]. Empirical pharmaceutical deployments further indicate that, under strict regulatory conditions, federated training can in practice be more economical than centralized aggregation workflows [8].

Another recurring assumption is that limited local datasets are of marginal value. In contrast, federated optimization is structurally designed to leverage heterogeneity across clients, where even small datasets can capture rare compositional regimes or failure modes absent elsewhere. Although aggregation mechanisms naturally weight contributions by dataset size, all participants ultimately receive the same global model, allowing smaller academic or industrial actors to meaningfully shape the learned representation space [16, 17].

Trust-related objections focus on the risk of information leakage among participants. Secure aggregation protocols combined with differential privacy guarantees ensure that individual updates cannot be reverse-engineered into identifiable data, even under adversarial conditions [10, 11, 21, 25]. Such cryptographic protections, already validated in multi-institutional drug discovery collaborations, translate directly to distributed materials databases without compromising confidentiality.

While some advocate for fully open data as the epistemic ideal, such a position often collides with the realities of proprietary ownership, regulatory constraints, and export-controlled materials datasets [27]. Federated learning emerges as a pragmatic alternative under these constraints, enabling structured collaboration without violating confidentiality. In this sense, it reframes participation not as a concession of competitive advantage but as a controlled mechanism for collective model improvement under realistic industrial conditions.

Relation to Other Positions

This position builds directly on prior calls for improved data practices in materials science. The roadmap on data-centric materials science emphasized the urgent need for better sharing mechanisms and infrastructure to overcome current fragmentation [27]. Federated learning operationalizes that vision precisely when traditional open-data approaches encounter insurmountable legal and competitive barriers.

It also extends the dataset-bias critique that has repeatedly highlighted how public repositories limit model generalization [1-3]. Rather than lamenting bias, federated learning actively mitigates it by incorporating diverse proprietary distributions—real-world processing conditions, rare compositions, and industrial-scale variability—that public datasets cannot provide. The result is not incremental improvement but a qualitative leap in model robustness.

The position aligns naturally with established transfer-learning workflows in materials informatics [4]. Federated pre-training on the pooled proprietary knowledge base produces a powerful foundation model that each participant can then fine-tune on its internal tasks, exactly as recommended in recent surveys of machine-learning adoption barriers. This pre-train-and-adapt paradigm maximizes value while maintaining data sovereignty.

Privacy-preserving machine learning in adjacent fields offers further precedent. Drug-discovery consortia have already scaled federated learning to hundreds of thousands of proprietary compounds without compromising intellectual property [7-9, 28]. Materials science can adopt the same governance frameworks, client-clustering strategies, and secure-aggregation protocols that proved successful in pharmaceutical quantitative structure–activity relationship modeling. Even additive-manufacturing pilots have shown that cross-factory federated training improves process qualification without data centralization [6].

Collectively, these related positions converge on a single conclusion: the era of isolated datasets is ending. Federated learning is the missing technical and organizational bridge that converts long-standing aspirations for collaborative, privacy-aware data science into a practical reality for proprietary materials databases.

Implementation Roadmap for Materials

A phased, pragmatic roadmap offers a viable pathway to transition the community from conceptual exploration toward production deployment of federated learning in materials science. Initial efforts, spanning one to two years, center on carefully scoped pilot projects in which two or three organizations sharing overlapping interests—such as solid-state electrolytes or high-entropy alloys—initiate a minimal cross-silo trial focused on a single target property, for instance formation energy or ionic conductivity. Under these conditions, existing open-source frameworks, including TensorFlow Federated or PyTorch-based federated libraries, can be deployed directly alongside materials-specific adaptations for graph neural networks that accommodate crystal structures [16, 22]. Model convergence speed, privacy audit outcomes, and accuracy relative to isolated training then serve as critical indicators of early viability.

Building upon demonstrated feasibility, the subsequent phase, extending from two to four years, shifts toward structured consortium formation. A dedicated materials federated-learning consortium comprising five to ten members drawn from industry, national laboratories, and academia would operate under neutral governance to develop tailored protocols addressing heterogeneous data, periodic boundary conditions, and multi-fidelity inputs. This institutional scaffolding also

enables the open publication of standardized benchmarks that rigorously compare federated against centralized performance, thereby cultivating broader community trust [18, 26]. Legal templates for consortium agreements and contribution-based reward mechanisms further emerge as essential instruments for aligning incentives across participants.

Beyond these foundational steps, the final stage, spanning four to six years, envisions seamless integration of federated learning into routine industrial workflows. At this juncture, the approach supports simultaneous modeling of multiple properties while enabling continuous learning from live production data streams, with trusted third-party auditing nodes facilitating validation without compromising test-set confidentiality. The resulting infrastructure naturally accommodates both cross-silo and hybrid cross-device configurations, thereby allowing smaller academic groups to contribute meaningfully. Throughout this progression, sustained development of open-source toolkits and public case studies remains instrumental in accelerating adoption, ensuring that advancement proceeds incrementally from low-risk pilots only after privacy guarantees and tangible value have been firmly established.

Recommendations for Stakeholders

Industry partners should initiate pilot projects with one or two trusted collaborators using existing federated frameworks, then join or form a materials-specific consortium. Investment in secure aggregation hardware and differential-privacy toolkits will yield rapid returns through improved models without data-sharing risk [8-11].

Academic groups must develop materials-tailored federated algorithms that handle non-i.i.d. crystal data, periodic structures, and heterogeneous experimental distributions. Creating public FL benchmarks for formation energy, band-gap prediction, and mechanical properties will accelerate community progress [16, 22]. Academia should also actively seek industry partnerships for pilot validation.

Funding agencies are urged to support dedicated infrastructure grants for federated-learning platforms, consortium formation, and open-source tool development. Targeted calls that require cross-sector participation and explicit privacy guarantees will catalyze adoption.

Journals and conferences should welcome federated-learning case studies and require authors to disclose privacy mechanisms when proprietary data are involved. Special issues on privacy-preserving materials informatics would further legitimize the approach.

Collective action by all four stakeholder groups—industry, academia, funders, and publishers—will transform federated learning from a promising idea into the standard collaboration model for proprietary materials databases.

Conclusion

Proprietary materials databases remain an untapped resource of immense value. Federated learning provides the only practical mechanism to unlock them without centralized data exchange, preserving privacy, competitive advantage, and regulatory compliance. The benefits are clear: genuine privacy preservation, unprecedented data diversity, competitive pre-training of foundation models, access to rare events, continuous learning, and simplified legal adherence.

The challenges—heterogeneous data, communication costs, security risks, incentives, computational burden, and validation—are real but solvable through targeted research and consortium governance. The three-phase implementation roadmap—pilots, consortium formation, and production deployment—offers a concrete path forward.

The materials science community now faces a choice: continue operating with fragmented, biased datasets or embrace federated learning to integrate the complementary strengths of industrial and academic knowledge. This position paper

calls on industry to launch initial pilots, on academia to develop materials-specific algorithms, on funding agencies to invest in infrastructure, and on journals to champion transparent privacy-preserving research.

Only through deliberate, coordinated action can we build the federated-learning infrastructure that will define the next decade of data-driven materials discovery. The technology exists. The need is urgent. The time to act is now.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Dec 2024

Revised: 08 Mar 2025

Accepted: 16 May 2025

Published online: 18 July 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Omee SS, Fu N, Dong R, Hu M, Hu J. Structure-based out-of-distribution (OOD) materials property prediction: A benchmark study. *NPJ Comput Mater.* 2024;10(1):144.
<https://doi.org/10.1038/s41524-024-01316-4>.
- Boyce B, Dingreville R, Desai S, Walker E, Shilt T, Bassett KL, et al. Machine learning for materials science: Barriers to broader adoption. *Matter.* 2023;6(5):1320-3.
<https://doi.org/10.1016/j.matt.2023.03.028>.
- Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. *Adv Sci (Weinh).* 2019;6(21):1900808.
<https://doi.org/10.1002/advs.201900808>.

Jain A. Machine learning in materials research: Developments over the last decade and challenges for the future. *Curr Opin Solid State Mater Sci.* 2024;33:101189.

<https://doi.org/10.1016/j.cossms.2024.101189>.

Huang W, Barnard AS. Federated data processing and learning for collaboration in the physical sciences. *Mach Learn Sci Technol.* 2022;3(4):045023.

<https://doi.org/10.1088/2632-2153/aca87c>.

Mehta M, Bimrose MV, McGregor DJ, King WP, Shao C. Federated learning enables privacy-preserving and data-efficient dimension prediction and part qualification across additive manufacturing factories. *J Manuf Syst.* 2024;74:752-61.

<https://doi.org/10.1016/j.jmsy.2024.04.031>.

Moein M, Heinonen M, Mesens N, Chamanza R, Amuzie C, Will Y, et al. Chemistry-based modeling on phenotype-based drug-induced liver injury annotation: From public to proprietary data. *Chem Res Toxicol.* 2023;36(8):1238-47.

<https://doi.org/10.1021/acs.chemrestox.2c00378>.

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.

<https://doi.org/10.1021/acs.jcim.3c00799>.

Oldenhof M, Ács G, Pejó B, Schuffenhauer A, Holway N, Sturm N, et al. Industry-scale orchestrated federated learning for drug discovery. *Proc AAAI Conf Artif Intell.* 2023;37(13):15576-84.

<https://doi.org/10.1609/aaai.v37i13.26847>.

Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson Ú. Scalable private learning with PATE. *arXiv Preprint arXiv:1802.08908.* 2018.

<https://doi.org/10.48550/arXiv.1802.08908>.

Choquette-Choo CA, Dullerud N, Dziedzic A, Zhang Y, Jha S, Papernot N, et al. CaPC learning: Confidential and private collaborative learning. *arXiv Preprint arXiv:2102.05188.* 2021.

<https://doi.org/10.48550/arXiv.2102.05188>.

Ghalkhani M, Habibi S. Review of the Li-ion battery, thermal management, and AI-based battery management system for EV application. *Energies.* 2023;16(1):185.

<https://doi.org/10.3390/en16010185>.

Yang L, Yang X, Xia F, Gong Y, Li F, Yu J, et al. Recent progress on natural clay minerals for lithium-sulfur batteries. *Chem Asian J.* 2023;18(16):e202300473.

<https://doi.org/10.1002/asia.202300473>.

Ozcan A, Coudert FX, Rogge SM, Heydenrych G, Fan D, Sarikas AP, et al. Artificial intelligence paradigms for next-generation metal-organic framework research. *J Am Chem Soc.* 2025;147(27):23367-80.

<https://doi.org/10.1021/jacs.5c08214>.

Mishra RK, Mishra D, Agarwal R. An artificial intelligence-powered approach to material design. In: *Cutting-Edge Research in Chemical and Material Science*. Vol. 1. Kolhapur: Bhumi Publishing; 2024. p. 61-89.

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.

<https://doi.org/10.1016/j.patter.2022.100521>.

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.
<https://doi.org/10.1186/s13321-021-00576-2>.

Evans ML, Bergsma J, Merkys A, Andersen CW, Andersson OB, Beltrán D, et al. Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange. *Digit Discov.* 2024;3(8):1509-33.
<https://doi.org/10.1039/d4dd00039k>.

Dutta S, Leal de Freitas I, Maciel Xavier P, Miceli de Farias C, Bernal Neira DE. Federated learning in chemical engineering: A tutorial on a framework for privacy-preserving collaboration across distributed data sources. *Ind Eng Chem Res.* 2025;64(15):7767-83.
<https://doi.org/10.1021/acs.iecr.4c03805>.

Mammen PM. Federated learning: Opportunities and challenges. *arXiv Preprint arXiv:2101.05428.* 2021.
<https://doi.org/10.48550/arXiv.2101.05428>.

Kucur EN, Buyuktanir T, Ugurelli M, Yildiz K. Privacy-preserving machine learning techniques: Cryptographic approaches, challenges, and future directions. *Appl Sci.* 2026;16(1):277.
<https://doi.org/10.3390/app16010277>.

Hanser T. Federated learning for molecular discovery. *Curr Opin Struct Biol.* 2023;79:102545.
<https://doi.org/10.1016/j.sbi.2023.102545>.

Bai Q, Liu S, Tian Y, Xu T, Banegas-Luna AJ, Pérez-Sánchez H, et al. Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdiscip Rev Comput Mol Sci.* 2022;12(3):e1581.
<https://doi.org/10.1002/wcms.1581>.

Gallios G, Tsakiridis N, Tziolas N. Federated learning applications in soil spectroscopy. *Geoderma.* 2025;456:117259.
<https://doi.org/10.1016/j.geoderma.2025.117259>.

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.
<https://doi.org/10.1016/j.drudis.2023.103820>.

Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. *Comput Ind Eng.* 2020;149:106854.
<https://doi.org/10.1016/j.cie.2020.106854>.

Bauer S, Benner P, Bereau T, Blum V, Boley M, Carbogno C, et al. Roadmap on data-centric materials science. *Model Simul Mater Sci Eng.* 2024;32(6):063301.
<https://doi.org/10.1088/1361-651X/ad4d0d>.

Leeson PD. Impact of physicochemical properties on dose and hepatotoxicity of oral drugs. *Chem Res Toxicol.* 2018;31(6):494-505.
<https://doi.org/10.1021/acs.chemrestox.8b00044>.