Computational Toxicology 37 (2026) 100404

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

Full Length Article

Federation of toxicological data resources for *in silico* new approach methodologies (NAMs)

Nicoleta Spînu [a], Dimitris Stripelis [b], Mark T.D. Cronin [c], Gregory L. Warren [d], Andrew P. Worth [e,*]

[a] AI4Cosmetics, Amsterdam, Netherlands
[b] Flower Labs, Cambridge, UK
[c] School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK
[d] Osmo Labs, PBC, Cambridge, MA, USA
[e] European Commission, Joint Research Centre (JRC), Ispra, Italy

ARTICLE INFO

Keywords:
Federated analytics
Federated learning
Quantitative structure–activity relationship (QSAR)
Dermal permeability
Mutagenicity
Skin sensitisation
Chemical safety assessment

ABSTRACT

Next Generation Risk Assessment (NGRA) promotes animal-free, exposure-informed, and hypothesis-driven approaches to chemical safety assessment. *In silico* tools, such as quantitative structure–activity relationship (QSAR) models, are valuable new approach methodologies (NAMs) for use in NGRA. However, the practical implementation of *in silico* NAMs remains limited by challenges in data availability, heterogeneity, and regulatory acceptance. In this study, federated learning is introduced to advance chemical safety assessment while leveraging proprietary data domains. Federated learning is a decentralised machine learning approach where multiple organisations, devices or servers collaboratively train a model while keeping their data locally, sharing only model updates to preserve confidentiality and privacy. Three use cases were simulated with the Flower open-source federated learning framework, namely (i) federated analytics for dermal permeability (log Kp) screening; (ii) federated convolutional neural networks (CNNs) for mutagenicity prediction from SMILES strings, and (iii) federated eXtreme Gradient Boosting (XGBoost) models for predicting skin sensitisation potential using molecular fingerprints and descriptors. The results show that federated learning approaches can yield predictive performance comparable to centralised models while mitigating concerns over the visibility of, and access to, commercially sensitive data. Open challenges related to data curation, interpretability, and model governance, as well as future directions, are discussed. This work demonstrates that federated learning can facilitate secure collaboration across organisations, enhance the utility of distributed chemical datasets, and accelerate the adoption of *in silico* NAMs.

1. Introduction

Next Generation Risk Assessment (NGRA) is an emerging framework that advocates for exposure-informed, hypothesis-driven, and animal-free safety assessment of chemicals [1,2]. It relies heavily on new approach methodologies (NAMs), including *in vitro* assays and *in silico* tools. Despite increasing investment in non-animal alternatives, the uptake of NGRA approaches in regulatory settings has been slow across many regions and sectors, in part due to the current immaturity of the framework and the need for further development of underlying methodologies and case study evidence to demonstrate their robustness and protective capabilities [3]. However, developments and practical applications of NGRA, such as the Alternative Safety Profiling Algorithm (ASPA), offer a vision for its successful implementation [4].

Computational methods, such as quantitative structure–activity relationship (QSAR) models, are widely employed to address data gaps in chemical safety decision-making, particularly when experimental data are unavailable or inadequate. These models are valuable tools for hazard identification and prioritisation, but their use is not without limitations. The coverage of chemical space is often uneven, with certain compound classes overrepresented in publicly available sources due to historical testing practices [5]. This affects their applicability domain, which is not always clearly defined. In addition, the inherent variability in the data on which these models are trained can result in differences in

[*] Corresponding author.
E-mail addresses: nicoleta@ai4cosmetics.com (N. Spînu), andrew.worth@ec.europa.eu (A.P. Worth).

https://doi.org/10.1016/j.comtox.2026.100404
Received 25 November 2025; Received in revised form 15 January 2026; Accepted 28 January 2026
Available online 29 January 2026

predictions for the same endpoint. Furthermore, many QSAR models are not routinely updated to reflect newly available data, limiting their relevance and robustness in evolving regulatory and research contexts [6].

Federated learning (FL), a decentralised machine learning paradigm introduced by McMahan et al. [7], enables collaborative model training without requiring direct data sharing. Instead, parameters from locally trained models are aggregated to update a global model. Originally developed for different domains, such as mobile and edge device use cases [8], FL has emerged as a viable approach for drug discovery applications, with practical implementations such as MELLODY [9–11], and promising research results, such as data-centric federated knowledge distillation [12], drug-target binding affinity and drug-drug interaction [13], mechanism of action prediction based on cell painting [14] and other molecular discovery applications [15]. These studies have shown that federated models can outperform locally trained models, particularly in scenarios where individual datasets are small or incomplete, by leveraging patterns learned from decentralised data silos while at the same time preserving data privacy and confidentiality. Such collaboration can lead to improved model generalisability and expanded applicability domains. In an NGRA context, FL offers a mechanism to jointly learn from distributed safety-relevant datasets to support hazard identification, problem formulation, and prioritisation of NAM testing without compromising proprietary data ownership. Federated analytics extends this paradigm beyond model training by enabling the computation of aggregate statistics, feature distributions, and mechanistic signals across decentralised datasets without exposing individual-level data [16]. By facilitating privacy-preserving exploratory analysis and evidence synthesis, federated analytics can improve cross-organisation data diversity and clustering structure across decentralised chemical spaces [17], therefore, supporting more informed hazard hypothesis generation and prioritisation of NAM testing in early NGRA workflows.

Though a potentially valuable enabler of data-driven NAMs, FL remains largely underexplored in learning from local proprietary data and subsequent development of *in silico* NAMs for use in chemical safety assessment. Its relevance could be substantial in NGRA for cosmetic ingredients, where FL could unlock significant value. Unlike pharmaceuticals or biocides, cosmetics companies often operate within overlapping chemical spaces, in addition to facing stringent animal testing bans [18]. They also benefit from extensive historical data resources. However, safety-relevant data are often siloed, proprietary, and costly to generate, which restricts collaborative modelling efforts and the scope of QSARs for toxicological endpoints. Two recent examples of commercial datasets that are not otherwise available for modelling were from Skare et al. [19], who used a proprietary database of approximately 800,000 substances to identify analogues for PEG cocamines, while Gautier et al. [20] identified two out of 25 resorcinol analogues through proprietary data sources. Improved access to such datasets could significantly enhance the accuracy, relevance, and applicability of predictive models. A recent NGRA case study on coumarin in cosmetic products [21] demonstrated how mechanistic understanding of skin sensitisation (e.g., pro-hapten behaviour) can be used to formulate hazard hypotheses and guide the selection of appropriate NAMs, ultimately informing the derivation of a human-relevant point of departure and margin-of-exposure-based risk conclusions. FL could analogously strengthen these steps by improving the collective learning used to frame hypotheses, identify shared mechanistic signals, structural alerts and chemical grouping, and exposure trends, and thus, supporting evidence-based prioritisation of chemicals and pathways for targeted NAM testing. However, legal, financial, and logistical barriers remain.

This study applied FL approaches for three use cases relevant to the safety assessment of chemicals, including the cosmetics industry, i.e. dermal permeability, mutagenicity, and skin sensitisation. These endpoints were selected because they address complementary local and systemic toxicity. Dermal permeability determines the potential for systemic exposure following topical application, mutagenicity provides an early indication of genotoxic hazard relevant to long-term systemic exposure, and skin sensitisation evaluates the risk of local immune-mediated adverse effects associated with repeated use. Through this investigation, we demonstrate that FL-based approaches can effectively harness siloed knowledge from fragmented data, yielding substantial improvements in *in silico* toxicity prediction compared to traditional centralised and isolated, siloed methods.

2. Materials and methods

2.1. Problem formulation

Three use cases were designed to evaluate the application of federated learning approaches to assist in the development of models and analytics in chemical safety assessment:

1. Percutaneous dermal permeability potential (logarithm of the skin permeability coefficient of chemical substances, log Kp) across the epidermis and dermis using federated analytics for the skin penetration assessment, motivated by the limited size and fragmented nature of available experimental datasets, which constrain the development of robust predictive models at individual organisations In this setting, federated analytics enables privacy-preserving exploratory analysis across decentralised datasets, facilitating the identification of potential confounding factors, the assessment of inter-laboratory variability, and the characterisation of data heterogeneity. This could support causal interpretation and uncertainty assessment of skin permeability without the need for centralised data pooling.
2. Prediction of mutagenicity based on SMILES strings using a federated convolutional neural network (CNN), reflecting the limited availability of mechanistic information in the initial dataset that could otherwise inform model design. A CNN architecture was selected to demonstrate the applicability of federated learning to non-traditional molecular modelling approaches, in which a global model is initialised centrally and gradient information is communicated to local models, while locally learned latent embeddings derived from SMILES sequences are returned to the server to update the global model. This setup could support privacy-preserving representation learning while facilitating downstreamgenerative design and structural similarity analysis.
3. Identification of skin sensitisation hazards based on molecular fingerprints and descriptors using federated eXtreme Gradient Boosting (XGBoost), selected for its predictive performance and explanatory utility, which are critical for regulatory acceptance and decision-making. In contrast to the representation-learning approach adopted for mutagenicity, this use case employs parameter-level federation, whereby model parameters are updated and exchanged across organisations rather than latent embeddings or gradients, enabling interpretable feature attribution while preserving data privacy across decentralised datasets.

2.2. Data description

2.2.1. Dermal permeability

Three open-source datasets containing curated skin permeability coefficient (Kp) from *in vitro* studies using human skin were selected to simulate the federated analytics use case:

- HuskinDB [22] comprises 546 Kp values for 251 compounds, sourced from 94 publications, and includes detailed experimental metadata, such as skin source site, skin layer used, preparation technique, storage conditions, temperature, pH, and types of donor and acceptor solutions.
- SkinPiX [23] provides 202 Kp values for 109 compounds from 37 publications and includes additional parameters relevant to

cutaneous absorption, such as steady-state flux (Jss), maximum flux (Jmax), and lag time (tlag), along with comprehensive experimental conditions.

- The database compiled by Stevens et al. [24] contains Kp data for 73 compounds and includes, among data types, diffusion coefficients (D) and skin layer-specific information.

Compounds with log Kp values for the epidermis and dermis were selected for federated analysis of individual skin layers, resulting in 124 compounds from HuskinDB, 88 compounds from SkinPiX, and 60 compounds from Stevens et al. [24]. Where a compound had multiple log Kp per skin layer, the median value was calculated. Considerable overlap was found between Stevens et al. [24] and SkinPiX, with 81.7% of Stevens et al. compounds also present in SkinPiX, while HuskinDB showed minimal overlap with either dataset, and only eight compounds were common across all three.

2.2.2. *Mutagenicity*

A large open-source dataset of mutagenicity, compiled by Xu et al. [25], was used. It includes 8,348 compounds classified as mutagenic or non-mutagenic based on the *in vitro* Ames assay results collected from previously published studies. Specific details regarding test outcomes for individual *Salmonella typhimurium* strains or the presence/absence of S9 metabolic activation mix were not available within the large dataset utilised for model development. Although smaller datasets are available, this dataset was selected as the sole source to enable random sampling to simulate a federated learning use case involving non-independent and identically distributed (non-IID) data partitioning, in which data subsets differ in their underlying feature distributions and activity patterns across simulated organisations, thereby reflecting real-world institutional heterogeneity [26].

2.2.3. *Skin sensitisation*

Two proprietary data sources were selected to simulate the collaborative co-training of a global federated model for hazard prediction based on *in vivo* Local Lymph Node Assay (LLNA) results. The first dataset was developed in-house by AI4Cosmetics and was compiled and curated from 14 sources resulting in 370 compounds with LLNA classification [27–40]. The second dataset, Skin Doctor CP [41], was a curated source containing LLNA hazard classifications for 1,285 compounds. The overlap of compounds, determined based on chemical identity using standardised canonical SMILES, consisted of 240 common compounds, corresponding to 65.6% of the AI4Cosmetics dataset and 18.7% of the Skin Doctor CP dataset, out of a total of 1,411 unique compounds.

2.3. *Federated environment*

The three use cases were implemented with the open-source Flower v.1.23.0 framework, which provides built-in support for secure and private decentralised model training [42]. The Flower framework employs a hub-and-spoke topology wherein a server coordinates federated learning across multiple clients, with the server-side architecture comprising SuperLink, i.e., a long-running process for task forwarding, SuperExec, i.e., the process manager, and ServerApp, i.e., a short-lived, project-specific code for strategy implementation [43,44]. In total, Flower provides approximately 20+ built-in federated learning strategies allowing for further customisation [45]. Flower also supports both centralised and federated model evaluation: in centralised evaluation, aggregated model parameters are evaluated server-side on a test dataset, whereas during federated evaluation, model parameters are evaluated across federation clients using their locally held test data [46]. For example, the federated models for skin sensitisation and mutagenicity classifications are loaded server-side and evaluated on a centralised global shared test dataset. Lastly, tTo ensure no direct access to a client's data and prevent private information from being leaked from local models, Flower provides secure aggregation protocols, such as SecAgg+ [47] and Salvia [48]. Data visualisation was performed using Plotly v.6.1.2. The code and documentation are available at https://github.com/ai4cosmetics/fl-chemsafe.

2.3.1. *Dermal permeability*

Federated analytics were applied to dermal permeability, where each dataset was treated as a hypothetical client (i.e., organisation) without any pre-processing steps such as deduplication of records. Clients computed local histograms of log Kp values per skin layer type, epidermis and dermis. In the first round, raw histograms were shared; in the follow-up round, Gaussian noise was added to each bin count to ensure differential privacy [49]. The global aggregated histograms were obtained by summing bin counts across clients, revealing the combined distributions. Lower epsilon values indicated higher noise (larger sigma) and stronger privacy, for example, with $\varepsilon = 1$, $\sigma = 1.0$, introducing ±1 log unit of noise per data point. This trade-off between privacy and utility allowed data to remain close to their original values while reducing the risk of inferring sensitive information. The approach enabled the characterisation of the overall dermal permeability landscape across the three independently curated datasets without sharing raw data, increasing statistical power while maintaining data privacy.

2.3.2. *Mutagenicity*

The experimental setup involved two stages: (1) data partitioning of the initial data source into two training sets with different mutagenicity distributions, and (2) collaborative co-training via federated learning. After removing duplicates ($n = 174$), the dataset (8408 unique molecules) was split randomly into training (80%) and test (20%) sets. The training data were then partitioned into two hypothetical clients (i.e., organisations) with deliberately skewed distributions to simulate a non-IID setting. Organisation A ($n = 3137$, ~80% mutagenic, 2606 + ve and 531 −ve) and Organisation B ($n = 2776$, ~20% mutagenic, 651 + ve and 2125 −ve), with no shared molecules between them. A shared test set ($n = 1480$, 55% mutagenic) was prepared for evaluation.

Federated learning was implemented using the FedAvg algorithm [7] over 10 communication rounds, combining local stochastic gradient descent with server-side model averaging. Each client trained a local CNN using SMILES strings as input. SMILES strings were standardised using MolVS (v0.1.1) and validated with RDKit (v.2025.3.2). A character-level vocabulary was constructed from all unique characters across the entire list of SMILES. This shared vocabulary was applied consistently across both organisation training sets and the global test set. The CNN architecture consisted of a 32-dimensional character embedding layer followed by two 1D convolutional layers (64 and 128 filters, kernel size 3) with max pooling, global adaptive max pooling, and a fully connected output layer for binary mutagenicity classification of SMILES molecular sequences.

Model performance was evaluated using accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and false negative rate (FNR), comparing the federated model against individual local models to evaluate the performance and the mutagenicity distribution of the chemical space. This approach simulated a vertical federated learning where participants hold the same samples, i.e., the same chemical space sample, which may or may not be classified as mutagens, but a different feature space, i.e., SMILES strings as characters.

2.3.3. *Skin sensitisation*

The experimental setup involved one round of federated training using the FedXgbBagging aggregation method of the Flower framework [42] to combine two locally trained models on independent datasets. It applied a bagging technique that consisted of two stages: (i) a local learning stage, where each client (i.e., organisation) trains a local XGBoost model using its own dataset, and (ii) an aggregation stage, where the central server averages the prediction results from the local models to construct the global model. Each dataset was split into 80% training and 20% testing subsets locally without deduplication of compounds to simulate a real-world scenario. The global test set was formed by combining both local test sets. This scenario simulated a horizontal federated learning setup, where clients share the same feature space, i.e., molecular features, but hold different samples, i.e., a list of chemicals.

For centralised learning, training sets were merged after removing duplicate SMILES. SMILES strings were standardised using MolVS (v0.1.1), and a combination of MACCS key fingerprints and molecular descriptors was generated using RDKit v.2025.3.2. Preprocessing steps included handling infinities and extreme values, followed by filtering out constant, highly correlated, and low-variance features, resulting in 259 final features. The chemical space visualisation was performed using Uniform Manifold Approximation and Projection (UMAP), with hyperparameters n_neighbors = 15, min_dist = 0.1. The XGBoost model employed a gradient boosting framework with 200 estimators (max depth 6, learning rate 0.05).

The federated model was compared against two baselines: (i) Local Learning (LL), where each client trained a model solely on their private data to assess whether federated learning outperforms individual models, and (ii) Centralised Learning (CL), where all data were pooled to train a single model to evaluate how closely federated learning approximates compared to centralised performance while preserving data privacy.

Tukey's Honest Significant Difference (HSD) test [50] in statsmodels v0.14.4 with SciPy v1.16.3 was employed. This statistical method controls for family-wise error rate when performing multiple pairwise comparisons, making it appropriate for comparing more than two models simultaneously. For models trained on the global test set (FL and CL), performance metric distributions were generated using bootstrap resampling with 100 iterations. This approach provided robust estimates of the variance in model performance. For local models (AI4Cosmetics and SkinDoctorCP), which were evaluated on separate test sets, bootstrap distributions were simulated using a binomial approximation based on the point estimates and sample sizes. The standard error for each metric was calculated as $SE = \sqrt{(p(1\text{-}p)/n)}$, where p is the observed metric value and n is the test set size. Bootstrap samples were then drawn from a normal distribution with the observed mean and calculated standard error, subsequently clipped to the valid range [0,1]. Model performance was evaluated using accuracy, precision, recall, F1-score, AUC and FNR metrics.

3. Results and discussion

This study explored the potential of FL to support chemical safety assessment within the NGRA framework, especially as we move towards the regulatory adoption of NAMs. Three distinct applications were simulated to determine what may be achievable when applied to commercially sensitive data sets. These use cases demonstrate how FL could enable collaborative modelling while preserving data privacy, offering a viable path forward for unlocking distributed evidence across the cosmetics sector and beyond.

3.1. Federated analytics captures local data variability

Federated data analysis encompasses any analytical process performed across distributed datasets without moving the source data. This includes data exploration and statistical analyses, such as calculating the mean, sum, or histograms among the participating organisations. Thus, instead of training a model, statistics of the distributed datasets are computed without allowing explicit data sharing [16]. This could be particularly useful for the assessment of the training datasets underlying predictive models, such as QSARs.

Herein, histograms for dermal permeability coefficients across three independent data sources were aggregated to simulate a federated analytics scenario. This setup enabled a direct comparison between the original centralised distributions and the federated ones with and without the Gaussian noise (Fig. 1 vs. Fig. 2). The discrepancies between the datasets reflect non-IID effects commonly encountered in real-world data collaborations and highlight the importance of applying standardisation protocols when performing federated analytics. Rather than indicating redundancy, the presence of overlapping compounds underscores the need to incorporate dataset-specific metadata and experimental parameters that influence variations in permeability estimates. The permeability through the epidermis was more prevalent in the lower permeability range compared to the dermis. While direct comparison for single compounds is not possible with the current federated setup, i.e., histogram aggregation approach, federated analytics, in general, can help with the identification of typical permeability profiles for each skin layer and detection of compounds with unusually high or low permeability.

In chemical safety assessment, federated analytics have the potential to enable organisations to analyse a variety of datasets of varying size, e.g., high-throughput data, while gaining valuable insights without compromising the sensitive aspect of data. Such methodologies can facilitate the evaluation of data quality and consistency across sources, the comparison of local versus systemic toxicity indications, data bias, the retrieval of structural analogues to support read-across, the design of wet experiments, and consensus learning approaches, including the identification of reference compounds, learning about the overall data landscape. Collectively, these capabilities enhance the extraction of meaningful insights from distributed data, contributing to more robust and transparent safety evaluations. For example, Bujotzek et al. [17] demonstrated federated clustering as a practical federated analytics approach for molecular diversity analysis across decentralised PharmaBench datasets.

3.2. Federated learning overcomes dataset-specific biases

A vertical federated learning scenario was simulated for mutagenicity prediction, where different parties hold different features (columns, i.e., SMILES strings) of the same set of data samples (rows, i.e., compounds). Specifically, gradients of the CNN models were distributed to each client, and the resulting embeddings were collected to update the global federated CNN model. As shown in Fig. 3, the federated model had a balanced performance (i.e., maintaining consistent performance across clients with heterogeneous data) compared to the individual local models, which were trained on different mutagenic compound distributions. This performance stems from access to a broader, more diverse chemical space, yielding more reliable and confident predictions. Such an approach is particularly suited to scenarios where organisations train local models on different modalities, including proprietary molecular descriptors and *in vitro* data. For instance, Japanese regulations restrict access to Ames test data for class B (positive) and C (negative) chemicals, whilst class A (strong positive) chemicals are publicly available [52].

3.3. Federated learning has comparable performance to central and local models

Horizontal federated learning refers to a setting in which multiple parties collaboratively train a machine learning model on datasets that share the same feature space but contain different samples [8]. Herein, two QSAR XGBoost models were independently trained on molecular fingerprints and descriptors, using datasets with differing sample sizes. This simulated the exchange of model parameters related to the skin sensitisation endpoint across two organisations. The federated approach aggregated complete tree models from each client, constructed a global ensemble by combining the locally trained models, and returned an aggregated model for inference.

Given the simulated nature of this study, it allows for a direct comparison of prediction confidence between the federated global model and both local and centralised models. Fig. 4 illustrates the chemical space of the training datasets and the shared test set used for

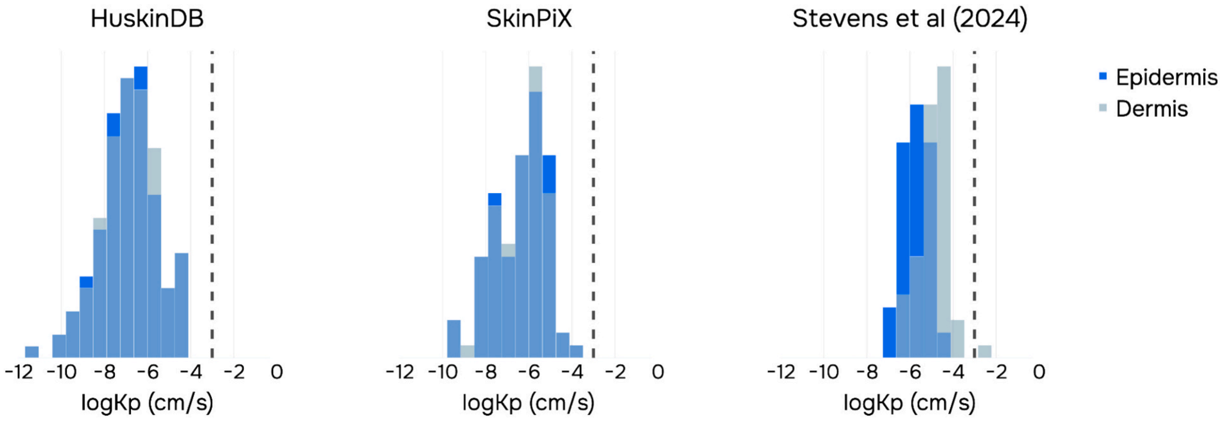

Fig. 1. Distribution of log Kp values for epidermis and dermis across the initial datasets. The dashed line at log Kp = −3 indicates the threshold for high permeability [51]. Most compounds fell below this value, suggesting low permeability. The Stevens et al. [24] dataset showed lower variability and clearer separation between skin layers. The non-IID distribution reflects the inherent heterogeneity of each dataset, representative of real-world scenarios: HuskinDB (−11.36 to −4.42), SkinPiX (−9.47 to −3.78), Stevens et al. [24] (−6.94 to −2.52).

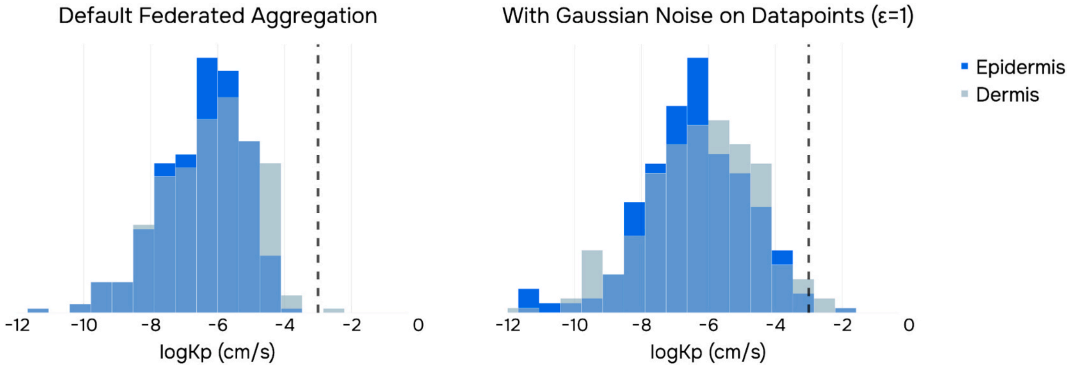

Fig. 2. Distribution of log Kp values for epidermis and dermis under different federated aggregation strategies. The left panel shows the baseline federated aggregation (−11.36 to −2.52), which accurately captures the variability in the local histograms of log Kp. The right panel demonstrates the impact of differential privacy, where Gaussian noise (ε = 1) was added during aggregation (−12.1 to −1.89). The inclusion of noise broadens the distribution, reflecting a trade-off between privacy and precision.

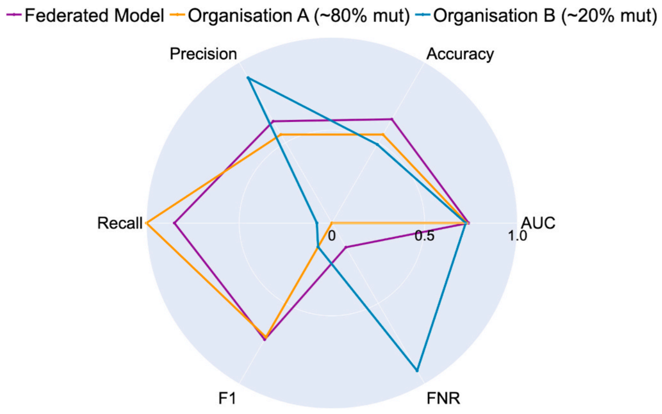

Fig. 3. Federated learning performance with heterogeneous client participation. The federated model demonstrated balanced performance across accuracy, precision, recall, F1-score, AUC and FNR metrics, effectively combining the complementary strengths of individual models while maintaining robust generalisation capabilities. Organisation A trained the model on a dataset with ~80% of 3,137 compounds being classified as mutagenic. Organisation B trained the model on a dataset with ~80% of 2,776 compounds being classified as non-mutagenic. Organisation A excels at finding known positive cases, being overly conservative (high recall, no FNR), while Organisation B misses almost all mutagenic compounds (high precision, high FNR).

performance evaluation. In real-world federated systems, however, training data remain undisclosed. For example, *in vitro* data typically do not leave the originating laboratories.

The comparative performance analysis based on the Tukey HSD tests (Fig. 5) showed that the CL model achieved the highest performance across most evaluation metrics, followed closely by the FL model, whereas the LL models exhibited greater variability and generally lower predictive capability. The FL model attained performance closely comparable to that of CL, with no statistically significant differences observed in several key metrics, including recall and FNR. Nonetheless, this marginal trade-off is outweighed by the intrinsic advantage of FL, which enables privacy-preserving, distributed model training without the need for centralised data aggregation.

Today, most open-source QSAR models are trained on the same publicly available datasets. They differ in architecture and optimisation methods. For example, Smajić et al. [53] highlighted a significant bias in models trained on public data, which tend to overpredict positive outcomes, while models using industrial data more frequently predict negative outcomes. Cronin et al. [54] highlighted several other reasons for differences in performance across QSAR models trained on the same data for mutagenicity, including a lack of appropriate descriptors related to the endpoint and mechanism of action. Similarly, it was observed herein that the AI4Cosmetic model prioritises finding known sensitisers (high recall) while the model trained on Wilm et al. [41] data excels at discrimination (high FNR). Thus, the horizontal FL approach can lead to a broader chemical space domain for both toxicity screening and

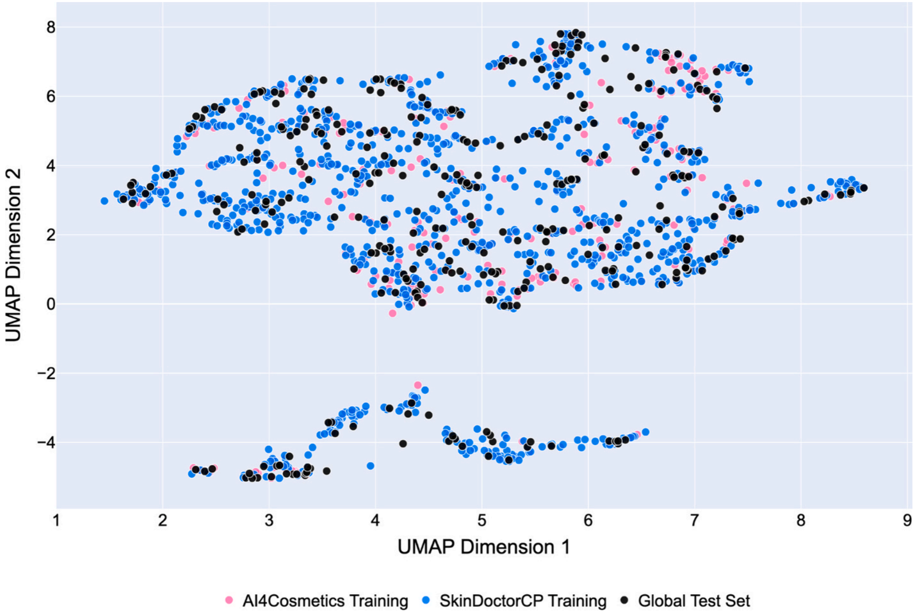

Fig. 4. Chemical space of the training and test sets. All datasets span a broad chemical space coverage, with the AI4Cosmetics training set (pink) presenting limited diversity due to the data size collected and curated from peer-reviewed open sources. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

regulatory applications.

3.4. Open challenges

How the training data underlying the local QSAR models are treated and preprocessed remains an open challenge in federated chemical safety assessment. The concern lies in the quality, heterogeneity, and interpretability of locally held data, often originating from assays with differing mechanistic relevance and inconsistently reported metadata [55]. For example, determining whether a compound is mutagenic, hepatotoxic, or neurotoxic may depend heavily on the assay design, e.g., bacterial strain, solvent, cytotoxicity, and biological coverage, which can vary significantly across organisations, especially in the case of non-standardised experimental protocols being used. Recent guidance, such as the OECD (Q)SAR Assessment Framework (QAF) [56], emphasises the importance of transparency, scientific rigour, and structured metadata in evaluating model predictions, principles which are equally critical for the federated modelling reporting to achieve regulatory acceptance. If the local QSARs have available the corresponding QSAR Model Reporting Format (QMRF) and/or provide those details, there should be no additional challenge in the regulatory acceptance of a federated QSAR. In principle, this is no different than the option to provide multiple QMRFs for the individual QSARs in a consensus model (such as CATMoS for acute toxicity prediction; [57] or the need to provide multiple QMRFs when multiple QSARs are integrated on an ad hoc basis to generate QSAR results following the OECD QSAR Assessment Framework [56].

Comprehensive metadata would increase the interpretability and confidence in predictions. This reinforces the need for organisations to engage in collaborative efforts to delineate applicability domain boundaries, such as through nearest-neighbour or similarity-based methods. For example, the choice between horizontal and vertical FL depends primarily on data alignment across clients, as well as the objective of the collaboration, to expand the number of samples to improve generalisation, as in horizontal FL, or to enrich the feature space for better predictions, as in vertical FL.

Since data preprocessing and standardisation are typically the responsibility of each data custodian, harmonised data cleaning protocols are essential for trustworthy federated analytics and model development. Heyndrickx et al. [9], as part of the MELLODY consortium, published a manual to ensure that all data of the contributing partners, mainly the input structure activity data, are prepared for the federated machine learning according to the same standards and principles. This leads to a consistent representation of chemical structures and the biological activity data. These protocols were executed by each data custodian individually on its own compute platform, and only the output of this process was made accessible to the federated machine learning. Thus, in the absence of this type of preparatory work, FL can aggregate poor data, which could result in increased noise and lower prediction rates.

Federated QSAR is constrained by the absence of standardised benchmarks in both datasets and model baselines, making it difficult to assess whether the performance of a global model is meaningful or simply reflective of well-explored chemical space. Applicability domains remain inconsistently defined, and neither being inside nor outside an applicability domain reliably predicts performance. Even consensus modelling can compound weaknesses when constituent models poorly cover chemical space.

Lastly, adopting federated learning across organisational boundaries necessitates rethinking how models are governed, maintained, and applied in diverse regulatory and scientific contexts. Unlike conventional modelling approaches, federated systems involve multiple stakeholders contributing data, infrastructure, or expertise. This imbalance can create structural inequities and introduce complexity into decision-making, particularly when models are applied in high-stakes regulatory uses. Some data holders may have more to gain from the collaboration, e.g., access to broader chemical space, or more to lose, e.g., potential commercial or regulatory sensitivity of their data. Effective governance must not only control access and permissions, for example, who can retrain, validate, or interpret the model, but also establish foundational principles such as Findable, Accessible, Interoperable, and Reusable (FAIR) data and model stewardship [58]. Incentives for participation may vary, from regulatory preparedness to scientific discovery or commercial advantage. Thus, aligning expectations early for a sustainable and effective federated collaboration is essential. Without such frameworks, the adoption and credibility of federated QSAR models will remain limited.

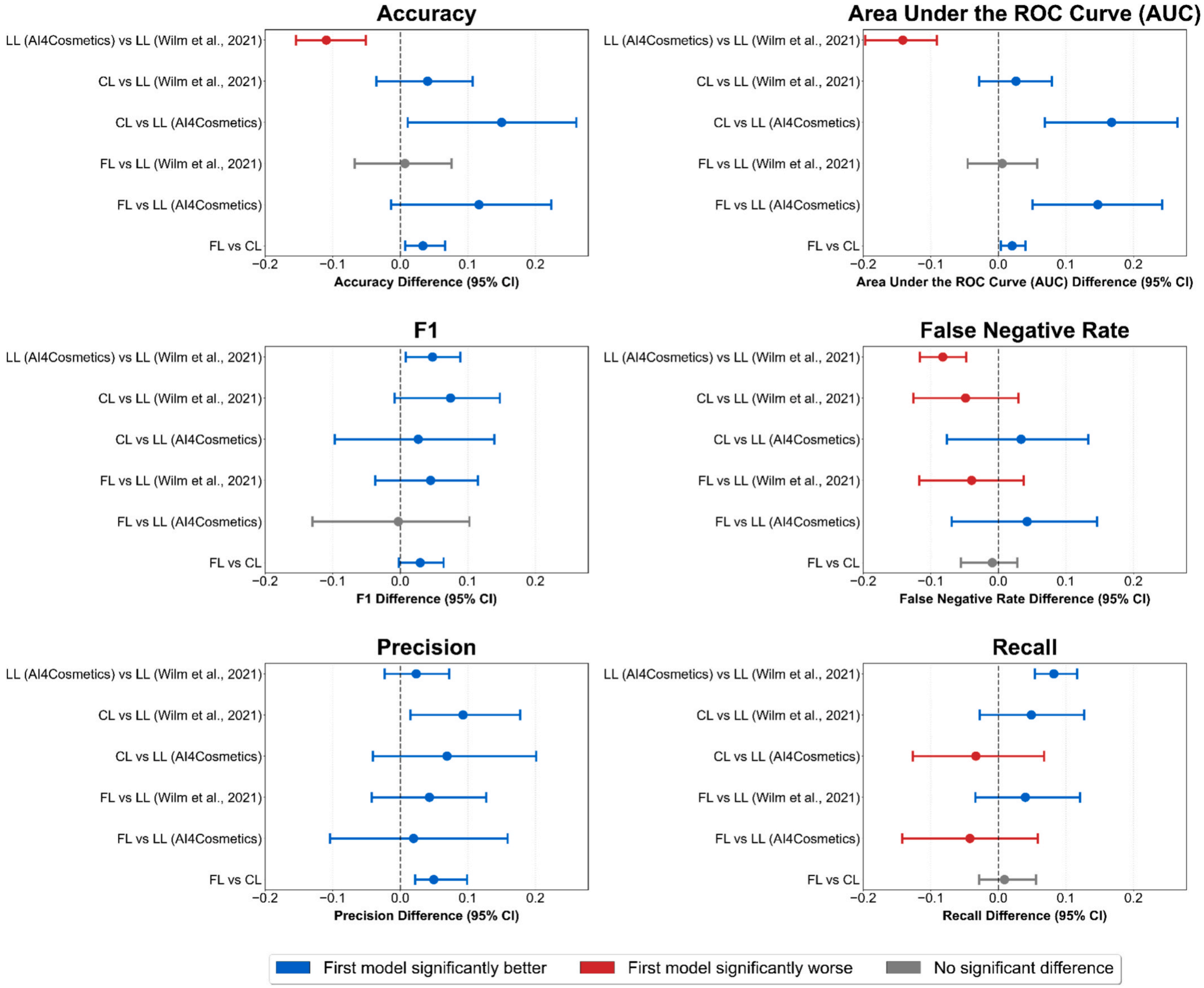

Fig. 5. Pairwise model comparison using Tukey's Honest Significant Difference (HSD) test. It shows the mean differences with 95% confidence intervals for all pairwise comparisons across the six performance metrics (accuracy, AUC, F1, FNR, precision, and recall). Comparisons were colour-coded: blue indicates the first model significantly outperformed the second ($p < 0.05$), red indicates the first model significantly underperformed, and grey indicates no statistically significant difference. Federated Learning achieves nearly the same performance as Centralised Learning, and both outperform Local Learning across all metrics. Local Learning (LL); Centralised Learning (CL); Federated Learning (FL); False Negative Rate (FNR). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.5. *Opportunities and future directions*

Federated decentralised architecture provides a pathway for collaborations in chemical safety assessment by enabling joint analysis and modelling across distributed datasets without requiring centralised data pooling. It reduces risks of data leakage (i.e., compromising data confidentiality) and supports compliance with data protection regulations. It facilitates secure, permission-based access, e.g., via application programming interfaces (APIs), promoting interoperability while preserving data ownership and local control.

A clear opportunity is to integrate public and proprietary datasets across organisations, where participating parties may specialise in distinct chemical domains, such as surfactants, fragrances, or industrial compounds. This thereby enhances both the diversity of the underlying evidence and the representativeness of the resulting predictive models. This broader integration can strengthen scientific understanding of the data landscape and reduce fragmentation across historically siloed datasets.

For academia, federated infrastructures may enable access to substantially larger and more diverse training corpora than are typically available in the public domain, supporting methodological innovation and more realistic benchmarking against industry-relevant chemical domains. For regulators, federated approaches could support more evidence-rich decision-making by enabling privacy-preserving interrogation of otherwise inaccessible datasets, facilitating independent validation across multiple data holders, and improving confidence in conclusions for chemicals that sit close to decision thresholds. For the industry, federated learning offers a practical route to collaborative model development that reduces duplication of experimental testing and strengthens safety substantiation by improving predictive performance of internal models and decision-making workflows without exposing proprietary assets. However, realising these benefits across all stakeholder groups depends on sufficiently broad participation and dataset diversity. Without contributions from large, well-resourced organisations, federated efforts may be constrained, limiting both regulatory relevance and the potential for scientific democratisation. In addition to following the OECD QSAR Assessment Framework [56], evaluation strategies may also benefit from existing guidance on best practices for machine learning in toxicological QSAR development [59]. Together, these resources support an opportunity to benchmark models and clearly define both toxicological and chemical applicability domains.

The absence of accessible datasets necessitates alternatives to standard validation approaches. This aligns with emerging efforts to establish scientific confidence in NAMs, such as the framework proposed by van der Zalm et al. [60], which, although focused on *in vitro* methods, offers relevant principles to be adopted by federated collaborations. In particular, for NAMs that contain intellectual property (protected elements in Test Guidelines), the OECD provides tools to maintain transparency, including reasonable and non-discriminatory terms ("RAND") for licensing commitments [61]. Protected elements can include, for example, computational algorithms and associated datasets. An example of how proprietary data can be protected while deriving new structural alerts for skin sensitisation is illustrated by MacMillan et al. [62]. As frameworks for the validation and regulatory acceptance of computational models continue to evolve, it is important to reconcile IPR considerations with the need for transparency during the validation and peer review processes.

Future research should explore other federated approaches, such as federated transfer learning and federated retrieval-augmented generation, for example, to support the scalable development and validation of knowledge constructs such as adverse outcome pathways. In parallel, reverse engineering methods will be essential to formally verify that local model execution does not result in unintended information leakage. Lastly, more research on all elements of acceptability, ranging from validation, verification, FAIRness, to reusability of these models, is vital.

4. Conclusions

Three use cases, namely federated analytics for dermal permeability assessment, vertical federated learning for mutagenicity assessment, and horizontal federated learning for skin sensitisation assessment, were simulated and implemented using the open-source federated learning framework Flower. Together, these examples illustrate that diverse molecular data modalities and modelling approaches, ranging from feature distributions to XGBoost and CNNs, can be implemented in a federated setting. The results demonstrate the feasibility of federated approaches for knowledge sharing, bridging critical information gaps between exposure and hazard. By enabling collaboration without direct data sharing, federated learning and analytics preserve data privacy while improving the performance and generalisability of local models and datasets for toxicity endpoints of critical regulatory interest. Decentralised learning can handle isolated datasets, such as for dermal permeability, and complex models, such as for mutagenicity and skin sensitisation, minimising the risk of data breaches and ensuring efficient, private, and robust training across distributed systems. This can strengthen the weight of evidence for chemicals prioritising targeted additional data generation where margins of safety are limited. Overall, decentralised approaches offer a promising route to advance NGRA and create new opportunities for regulatory science through more holistic and collaborative risk assessment across organisational boundaries.

Funding statement

This work was supported by the MKB innovatiestimulering topsectoren (MIT) Noord-Holland grant for the feasibility project FL-CHEMSAFE.

CRediT authorship contribution statement

Nicoleta Spînu: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dimitris Stripelis:** Writing – review & editing, Writing – original draft, Methodology. **Mark T.D. Cronin:** Writing – review & editing, Writing – original draft. **Gregory L. Warren:** Writing – review & editing, Writing – original draft, Methodology. **Andrew P. Worth:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Given his role as Co-Editor-in-Chief of Computational Toxicology, Mark T.D. Cronin had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. All the other authors, Nicoleta Spînu, Dimitris Stripelis, Gregory L. Warren, Andrew P. Worth, declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Flower Pilot Program is gratefully acknowledged. NS greatly acknowledges the feedback received from Arianna Bassan, James-Lucas Cairns, Stijn Denissen, Adrian Fowkes, Mojca Fuart-Gatnik, Caroline Gomes, Elaheh Jafarigol, Robert Landsiedel, Arto Maranjyan, Quinten Marcelis, Samaneh Mohammadi, Aleksandra Mostrag-Szlichtyng, Hirota Morihiko, Gladys Ouedraogo, Katarzyna Przybylak, Maria Sapounidou, and Clemens Wittwehr.

Data availability

Data and code publicly available here: https://github.com/ai4cosmetics/fl-chemsafe.

References

[1] S. Schmeisser, A. Miccoli, M. von Bergen, E. Berggren, A. Braeuning, et al., New approach methodologies in human regulatory toxicology - not if, but how and when!, Environ. Int. 178 (2023) 108082, https://doi.org/10.1016/j.envint.2023.108082.

[2] M.P. Dent, E. Vaillancourt, R.S. Thomas, P.L. Carmichael, G. Ouedraogo, H. Kojima, J. Barroso, et al., Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients, Regul. Toxicol. Pharmacol. 125 (2021) 105026, https://doi.org/10.1016/j.yrtph.2021.105026.

[3] C. Westmoreland, H.J. Bender, J.E. Doe, M.N. Jacobs, G.E.N. Kass, F. Madia, C. Mahony, I. Manou, G. Maxwell, P. Prieto, R. Roggeband, T. Sobanski, K. Schütte, A.P. Worth, Z. Zvonar, M.T.D. Cronin, Use of new approach methodologies (NAMs) in regulatory decisions for chemical safety: report from an EPAA Deep Dive Workshop, Regul. Toxicol. Pharmaco. 135 (2022) 105261, https://doi.org/10.1016/j.yrtph.2022.105261.

[4] M. Leist, S. Tangianu, F. Affourtit, H. Braakhuis, J. Colbourne, E. Cöllen, N. Dreser, et al., An alternative safety profiling algorithm (ASPA) to transform next generation risk assessment into a structured and transparent process, ALTEX (2025), https://doi.org/10.14573/altex.2509081.

[5] A. Smajić, T. Steger-Hartmann, G.F. Ecker, A. Hackl, Data exploration for target predictions using proprietary and publicly available data sets, Chem. Res. Toxicol. 38 (5) (2025) 820–833, https://doi.org/10.1021/acs.chemrestox.4c00347.

[6] J.C. Madden, S.J. Enoch, A. Paini, M.T.D. Cronin, A review of in silico tools as alternatives to animal testing: principles, resources and applications, ATLA 48 (2020) 146–172, https://doi.org/10.1177/0261192920965977.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics AISTATS. 2017, 54:1273-1282.

[8] P. Kairouz, H. McMahan, B. Avent, A. Bellet, M. Bennis, A. Bhagoji, et al., Advances and open problems in federated learning, arXiv (2019), https://doi.org/10.48550/arXiv.1912.04977.

[9] W. Heyndrickx, L. Mervin, T. Morawietz, N. Sturm, L. Friedrich, et al., MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information, J. Chem. Inf. Model. 64 (7) (2024) 2331–2344, https://doi.org/10.1021/acs.jcim.3c00799.

[10] M. Oldenhof, G. Ács, B. Pejó, A. Schuffenhauer, N. Holway, et al., Industry-scale orchestrated federated learning for drug discovery, Proc. AAAI Conf. Artif. Intell. 37 (13) (2024) 15576–15584, https://doi.org/10.1609/aaai.v37i13.26847.

[11] D. Bassani, A. Brigo, A. Andrews-Morger, Federated learning in computational toxicology: an industrial perspective on the Effiris hackathon, Chem. Res. Toxicol. 36 (9) (2023) 1503–1517, https://doi.org/10.1021/acs.chemrestox.3c00137.

[12] T. Hanser, E. Ahlberg, A. Amberg, et al., Data-driven federated learning in drug discovery with knowledge distillation, Nat. Mach. Intell. 7 (2025) 423–436, https://doi.org/10.1038/s42256-025-00991-2.

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

[14] L. Ju, A. Hellander, O. Spjuth, Federated learning for predicting compound mechanism of action based on image-data from cell painting, Artif. Intell. Life Sci. 5 (2024) 100098, https://doi.org/10.1016/j.ailsci.2024.100098.

[15] T. Hanser, Federated learning for molecular discovery, Curr. Opin. Struct. Biol. 79 (2023) 102545, https://doi.org/10.1016/j.sbi.2023.102545.

[16] D. Ramage, S. Mazzocchi. Federated analytics: Collaborative data science without data collection. Google Research, 2020, https://research.google/blog/federated-analytics-collaborative-data-science-without-data-collection/.

[17] M. Bujotzek, E. Trautmann, C. Hand, I. Hales. Insights into the unknown: Federated data diversity analysis on molecular data. 2025. arXiv preprint arXiv:2510.19535.

[18] P.L. Carmichael, M.T. Baltazar, S. Cable, S. Cochrane, M. Dent, H. Li, A. Middleton, I. Muller, G. Reynolds, C. Westmoreland, A. White, Ready for regulatory use: NAMs and NGRA for chemical safety assurance, ALTEX 39 (3) (2022) 359–366, https://doi.org/10.14573/altex.2204281.

[19] J.A. Skare, K. Blackburn, S. Wu, et al., Use of read-across and computer-based predictive analysis for the safety assessment of PEG cocamines, Regul. Toxicol. Pharmacol. 71 (3) (2015) 515–528, https://doi.org/10.1016/j.yrtph.2015.01.013.

[20] F. Gautier, F. Tourneix, V.H. Assaf, E. van Vliet, D. Bury, N. Alépée, Read-across can increase confidence in the next generation risk assessment for skin sensitisation: a case study with resorcinol, Regul. Toxicol. Pharmacol. 117 (2020) 104755, https://doi.org/10.1016/j.yrtph.2020.104755.

[21] G. Reynolds, J. Reynolds, N. Gilmour, R. Cubberley, S. Spriggs, A. Aptula, K. Przybylak, S. Windebank, G. Maxwell, M.T. Baltazar, A hypothetical skin sensitisation next generation risk assessment for coumarin in cosmetic products, Regul. Toxicol. Pharmacol. 127 (2021) 105075, https://doi.org/10.1016/j.yrtph.2021.105075.

[22] D. Stepanov, S. Canipa, G. Wolber, HuskinDB, a database for skin permeation of xenobiotics, Sci. Data 7 (426) (2020), https://doi.org/10.1038/s41597-020-00764-z.

[23] L. Chedik, S. Baybekov, F. Cosnier, et al., An update of skin permeability data based on a systematic review of recent research, Sci. Data 11 (224) (2024), https://doi.org/10.1038/s41597-024-03026-4.

[24] J.N. Stevens, A.K. Prockter, H.A. Fisher, et al., A database of chemical absorption in human skin with mechanistic modeling applications, Sci. Data 11 (755) (2024), https://doi.org/10.1038/s41597-024-03588-3.

[25] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P.W. Lee, Y. Tang, *In silico* prediction of chemical Ames mutagenicity, J. Chem. Inf. Model. 52 (11) (2012) 2840–2847, https://doi.org/10.1021/ci300400a.

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

[27] Y. Nukada, M. Miyazawa, S. Kazutoshi, H. Sakaguchi, N. Nishiyama, Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals, Toxicol. In Vitro 27 (2) (2013) 609–618, https://doi.org/10.1016/j.tiv.2012.11.006.

[28] M. Hirota, S. Fukui, K. Okamoto, S. Kurotani, N. Imai, et al., Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization, J. Appl. Toxicol. 35 (11) (2015) 1333–1347, https://doi.org/10.1002/jat.3105.

[29] J.S. Jaworska, A. Natsch, C. Ryan, J. Strickland, T. Ashikaga, M. Miyazawa, Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy, Arch. Toxicol. 89 (12) (2015) 2355–2383, https://doi.org/10.1007/s00204-015-1634-2.

[30] O. Takenouchi, S. Fukui, K. Okamoto, S. Kurotani, N. Imai, et al., Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals, J. Appl. Toxicol. 35 (11) (2015) 1318–1332, https://doi.org/10.1002/jat.3127.

[31] D. Urbisch, A. Mehling, K. Guth, T. Ramirez, N. Honarvar, et al., Assessing skin sensitization hazard in mice and men using non-animal test methods, Regul. Toxicol. Pharmacol. 71 (2) (2015) 337–351, https://doi.org/10.1016/j.yrtph.2014.12.008.

[32] D. Asturiol, S. Casati, A. Worth, Consensus of classification trees for skin sensitisation hazard prediction, Toxicol. In Vitro 36 (2016) 197–209, https://doi.org/10.1016/j.tiv.2016.07.014.

[33] J. Strickland, Q. Zang, M. Paris, D.M. Lehmann, D. Allen, N. Choksi, J. Matheson, A. Jacobs, W. Casey, N. Kleinstreuer, Multivariate models for prediction of human skin sensitization hazard, J. Appl. Toxicol. 37 (3) (2017) 347–360, https://doi.org/10.1002/jat.3366.

[34] M. Hirota, T. Ashikaga, H. Kouzuki, Development of an artificial neural network model for risk assessment of skin sensitization using human cell line activation test, direct peptide reactivity assay, KeratinoSens™ and in silico structure alert parameter, J. Appl. Toxicol. 38 (4) (2018) 514–526, https://doi.org/10.1002/jat.3558.

[35] N.C. Kleinstreuer, S. Hoffmann, N. Alépée, D. Allen, T. Ashikaga, et al., Non-animal methods to predict skin sensitization (II): an assessment of defined approaches, Crit. Rev. Toxicol. 48 (5) (2018) 359–374, https://doi.org/10.1080/10408444.2018.1429386.

[36] A. Natsch, G.F. Gerberick, Integrated skin sensitization assessment based on OECD methods (I): Deriving a point of departure for risk assessment, ALTEX 39 (4) (2022) 636–646, https://doi.org/10.14573/altex.2201141.

[37] S. Hoffmann, N. Alépée, N. Gilmour, P.S. Kern, E. van Vliet, et al., Expansion of the cosmetics Europe skin sensitisation database with new substances and PPRA data, Regul. Toxicol. Pharmacol. 131 (2022) 105169, https://doi.org/10.1016/j.yrtph.2022.105169.

[38] J. Reynolds, N. Gilmour, M.T. Baltazar, G. Reynolds, S. Windebank, G. Maxwell, Decision making in next generation risk assessment for skin allergy: using historical clinical experience to benchmark risk, Regul. Toxicol. Pharmacol. 134 (2022) 105219, https://doi.org/10.1016/j.yrtph.2022.105219.

[39] N. Alépée, F. Tourneix, A. Singh, N. Ade, S. Grégoire, Off to a good start? Review of the predictivity of reactivity methods modelling the molecular initiating event of skin sensitization, ALTEX 40 (4) (2023) 606–618, https://doi.org/10.14573/altex.2212201.

[40] F. Tourneix, L. Carron, L. Jouffe, S. Hoffmann, N. Alépée, Deriving a continuous point of departure for skin sensitization risk assessment using a Bayesian network model, Toxics 12 (8) (2024) 536, https://doi.org/10.3390/toxics12080536.

[41] A. Wilm, U. Norinder, M.I. Agea, C. de Bruyn Kops, C. Stork, J. Kühnl, J. Kirchmair, Skin Doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, Chem. Res. Toxicol. 34 (2) (2021) 330–344, https://doi.org/10.1021/acs.chemrestox.0c00253.

[42] D.J. Beutel, T. Topal, A. Mathur, X. Qiu, T. de Parcollet, P. Gusmão, N.D.F. Lane. Flower: A friendly federated learning research framework. arXiv. 2020, arXiv: 2007.14390.

[43] Flower Labs (2026a) 'Flower Architecture', Flower Framework Documentation, v1.25.0. Available at: https://flower.ai/docs/framework/explanation-flower-architecture.html (Accessed: 13 January 2026)..

[44] H.R. Roth, D.J. Beutel, Y. Cheng, M.J. Fernandez, H. Pan, et al., Supercharging Federated Learning with Flower and NVIDIA FLARE, in: Federated Learning in the Age of Foundation Models - FL 2024 International Workshops, Springer-Verlag, Singapore, 2025, pp. 36–45, https://doi.org/10.1007/978-3-031-82240-7_3.

[45] Flower Labs (2026b) 'Flower Architecture', Flower Framework Documentation, v1.25.0. Available at: https://flower.ai/docs/framework/tutorial-series-build-a-strategy-from-scratch-pytorch.html (Accessed: 13 January 2026).

[46] Flower Labs (2026c) 'Flower Architecture', Flower Framework Documentation, v1.25.0. Available at: https://flower.ai/docs/framework/explanation-federated-evaluation.html (Accessed: 13 January 2026)..

[47] J.H. Bell, K.A. Bonawitz, A. Gascón, T. Lepoint, M. Raykova, Secure single-server aggregation with (poly)logarithmic overhead, in: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020, pp. 1253–1269, https://doi.org/10.1145/3372297.3417885.

[48] L.K. Hei, P. Buarque, P. de Gusmão, D.J. Beutel, N.D. Lane. Secure aggregation for federated learning in flower. In Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning (DistributedML '21). Association for Computing Machinery, 2021, 8-14. doi: 10.1145/3488659.3493776.

[49] C. Dwork. Differential privacy. International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[50] S. Seabold, J. Perktold, Statsmodels: econometric and statistical modeling with Python, in: Proceedings of the 9th Python in Science Conference, 2010, pp. 92–96.

[51] G.P. Moss, J.C. Dearden, H. Patel, M.T. Cronin, Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption, Toxicol. In Vitro 16 (3) (2002) 299–317, https://doi.org/10.1016/s0887-2333(02)00003-6.

[52] A. Furuhama, A. Kitazawa, J. Yao, C.E. Matos Dos Santos, J. Rathman, et al., Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project, SAR QSAR Environ. Res. 34 (12) (2023) 983–1001, https://doi.org/10.1080/1062936X.2023.2284902.

[53] A. Smajić, I. Rami, S. Sosnin, G.F. Ecker, Identifying differences in the performance of machine learning models for off-targets trained on publicly available and proprietary data sets, Chem. Res. Toxicol. 36 (8) (2023) 1300–1312, https://doi.org/10.1021/acs.chemrestox.3c00042.

[54] M.T.D. Cronin, H. Basiri, G. Chrysochoou, S.J. Enoch, J.W. Firman, The predictivity of QSARs for toxicity: recommendations for improving model performance, Comput. Toxicol. 33 (2025) 100338, https://doi.org/10.1016/j.comtox.2024.100338.

[55] L.D. Burgoon, F.M. Kluxen, A. Hüser, M. Frericks, The database makes the poison: how the selection of datasets in QSAR models impacts toxicant prediction of higher tier endpoints, Regul. Toxicol. Pharmacol. 151 (2024) 105663, https://doi.org/10.1016/j.yrtph.2024.105663.

[56] A. Gissi, O. Tcheremenskaia, C. Bossa, C.L. Battistelli, P. Browne, The OECD (Q)SAR assessment framework: a tool for increasing regulatory uptake of computational approaches, Comput. Tox. 31 (2024) 100326, https://doi.org/10.1016/j.comtox.2024.100326.

[57] K. Mansouri, A.L. Karmaus, J. Fitzpatrick, G. Patlewicz, P. Pradeep, D. Alberga, N. Alepee, et al., CATMoS: collaborative acute toxicity modeling suite, Environ. Health Perspect. 129 (4) (2021) 47013, https://doi.org/10.1289/EHP8495.

[58] S.J. Belfield, H. Basiri, S. Chavan, G. Chrysochoou, S.J. Enoch, J.W. Firman, et al., Moving towards making (quantitative) structure-activity relationships ((Q)SARs) for toxicity-related endpoints findable, accessible, interoperable and reusable (FAIR), ALTEX (2025), https://doi.org/10.14573/altex.2411161.

[59] S.J. Belfield, M.T.D. Cronin, S.J. Enoch, J.W. Firman, Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs), PLoS One 18 (5) (2023) e0282924, https://doi.org/10.1371/journal.pone.0282924.

[60] A.J. van der Zalm, J. Barroso, P. Browne, W. Casey, J. Gordon, T.R. Henry, N. C. Kleinstreuer, A.B. Lowit, M. Perron, A.J. Clippinger, A framework for establishing scientific confidence in new approach methodologies, Arch. Toxicol. 96 (11) (2022) 2865–2879, https://doi.org/10.1007/s00204-022-03365-4.

[61] OECD (2019). Guiding Principles on Good Practices for the Availability/Distribution of Protected Elements in OECD Test Guidelines. OECD Series on Testing and Assessment, No. 298. OECD Publishing, Paris, https://doi.org/10.1787/2b290577-en.

[62] D.S. Macmillan, M.L. Chilton, J. Hillegass, Improvements to in silico skin sensitisation predictions through privacy-preserving data sharing, Regulatory Toxicology and Pharmacology 137 (2023) 105292, https://doi.org/10.1016/j.yrtph.2022.105292.