Topic: Federated and privacy-preserving machine learning in support of drug discovery

All information regarding future IMI Call topics is indicative and subject to change. Final information about future IMI Calls will be communicated after approval by the IMI Governing Board.

Topic details

Action type	Research and Innovation Action (RIA)
Submission & evaluation process	2 Stages

Specific challenges to be addressed

Enabled by an ever-expanding arsenal of model systems, analysis methods, libraries of chemical compounds and other agents (like biologics), the amount of data generated during drug discovery programmes has never been greater, yet the biological complexity of many diseases still defies pharmaceutical treatment. Hand in hand with rising regulatory expectations, this growing complexity has inflated the research intensity and associated cost of the average discovery project. It is, therefore, imperative that the learnings from these data investments are maximised to enable efficient future research. This could be empowered by the big data analysis and machine learning approaches that are currently driving the digital transformation across industries. These approaches not only rely on data generated specifically for any given project to inform it (as more established machine learning approaches tend to do), they also evaluate all other available data from different data sources and types for relevance to the question at hand. Thus, more information is extracted, which in turns enables a gradual virtualisation of drug discovery processes and increases efficiency in bringing more and safer drug candidates entering clinical trials. The success of this digital transformation will depend on unlocking the maximal amount of data for the learning tasks at hand and make them amenable to the latest approaches in machine learning.

To realise this digital transformation, the following specific challenges need to be addressed:

1. Unlock data that is distributed across data owners without disclosure of the data and related assets itself. This can be achieved by adhering to the following two principles.

    - **Privacy preservation** denotes the strict protection of confidential and IP-sensitive data and assets. In discovery, examples of these include the activities of compounds in assays or assay annotations the predictive models are trained with, or the predictive models derived from these data. In the strictest

sense, privacy preservation implies the data or assets never leave the control of their respective owners.

- **Federated machine learning** here denotes the distribution of the learning effort over physically separated partners. This goes beyond the more established concept of federated databases where the data but typically not the data functionalisation (i.e. learning from the data), is distributed. It is key to enable owner control over data or assets during learning.

2. Unlock data volumes from data sources or types that have hitherto remained untapped. In discovery, examples include image or transcriptional profiles or primary data points acquired in high throughput screens, all of which provide hard-to-interpret biological annotation of chemical compounds.

3. Make recent advances in machine learning such as multi-task learning and deep learning, amenable with the above data expansion strategies.

Need and opportunity for public-private collaborative research

The digital transformation that is driven by ever more exhaustive data collection and exploitation, is disrupting the entire industrial landscape. Sectors and geographies that fail to embrace this transformation will find themselves challenged in their remit by newcomers with a strong footing in data sciences.

In this context, a collaboration among pharmaceutical partners offers the perspective of doubling economies of scale in bringing better and safer drugs to patients. Firstly, it enables cost sharing and thereby bolsters the position of the European pharmaceutical industry in the global competition for data science and ICT resources, which includes academia and knowledge partners from Small and Medium-sized Enterprises (SMEs) and/or other commercial organisations active in the relevant spaces. Secondly, it encourages data and method standardisation, thus expanding the volume of collective data that can fuel the big data revolution. Notably, these collective data should not be misinterpreted as a freely accessible and hence fully precompetitive resource. Privacy-preserving approaches enable the reconciliation of collaborative investment with healthy within-sector competition.

The concepts of federated and privacy-preserving machine learning apply beyond the pharmaceutical remit. Indeed, by providing data owners the confidence that their data and the corresponding predictive models will remain private, the methodologies developed will encourage the formation of data and model consortia in various commercial and non-commercial contexts where data and knowledge ownership is at play, yet learning benefits from (indirect) access to larger data volumes. This creates opportunities for SMEs or other commercial partners that offer front-end or back-end services in the areas of software-as-a-service products in big-data analytics, clouded high-performance computing and privacy-preserving solutions. The public-private partnership proposed enables such partners to get exposed to, on the one hand, a strong application field with relevant use cases and clear ICT and security requirements, and on the other hand, academia and other knowledge partners with deep expertise in rapidly evolving science and technology fields.

Scope

The topic aims for:

The delivery/generation of a validated federated privacy-preserving machine learning platform on publicly accessible data that is demonstrably safe enough (privacy-preserving in the face of legitimate and illegitimate (attempted) access and use) and scalable enough to be deployed to a significant representation of the private data in the actual preclinical data warehouses of at least six major pharmaceutical companies in yearly evaluation runs. This effort will be mainly driven by the applicant consortium and enabled by the EFPIA partners.

The anticipated collective private compound and activity data sets from the EFPIA partners (which are to be accommodated comprehensively in each of the at least yearly runs) that are used during the evaluations will include:

- at least 3 million chemical compounds annotated with dose-response quality activity data;
- at least 5 million chemical compounds annotated with some activity;
- at least 1 billion assay activity data points collected at single dose (low-complexity i.e. 1 to a few numerical values per compound, e.g. as from conventional primary screening at high-throughput);
- at least 100 million activity data points collected in dose response (over a range of doses, e.g. as from follow-up/secondary screening);
- several high-complexity activities collected at high-throughput (at least 100 thousand compounds in a standardised setting, e.g. high-resolution microscopy images or transcriptional profiles with 1000 readouts per well).

As a part of the effort, the EFPIA partners will standardise, format and normalise their private data for optimal interoperability. However, the EFPIA partners will not mutually share assay meta-information, such as specifying which drug target is tested. The objective is to learn to predict the activity of the compounds in the documented assays from descriptors of their chemical structure, while leveraging as much of the available side information as possible. Methods within the scope of this topic should be compatible with the scale, richness and limitations of the above data. For example, given the absence of assay meta-information, no predictive performance gains can be realised by constructing models across data columns with similar or shared annotations. Predictive performance improvements from federated learning are expected to stem from the multi-task effect across partners.

Note that the above data and their annotations will be considered confidential information and remain strictly private to their respective original owners at all times. Aggregated predictive and computational performance and evaluation results will be shared with the public to enable the evaluation of the platform and algorithms.

Upon establishment of proof-of-concept, the expected time and cost efficiency gains in a development context (using clinical data) will most likely far outweigh those in a discovery setting (using assay and preclinical data). Hence, the design of the platform must be compatible or readily extendable to the clinical setting. Nevertheless, the complexities of clinical data handling in terms of adequately addressing ownership and privacy legislation implications would distract from the core objectives. Hence, any involvement of clinical data falls out of scope.

It is crucial to understand that the preservation of privacy and confidentiality is a key component of the topic. Privacy preservation is interpreted to exclude any consolidation of assay data or annotations, or the corresponding predictive models (even encrypted) outside of architectures under direct and sole control of their respective owners. Confidentiality relates to the confidential treatment of data and protection from access to them by third parties. Consequently, the proposed project aims for federated machine learning which is not the same thing as machine learning on federated data. The difference lies in that in the former case, the machine learning effort itself is distributed over the parties involved, in the latter case, the machine learning is executed centrally over federated data, which is incompatible with the proposed interpretation of privacy preservation. Upon completion of a modelling exercise no data (derived or otherwise) should persist outside of those architectures. The pharma IT departments will consolidate their IT security requirements, based on current industry standards that aim to protect against illegitimate access to or use of the data or predictive models.

To further bolster the confidence in the proposed methods of the pharmaceutical partners (and of potential other future adopters), an intrinsic part of the proposal should focus on analysing the privacy preservation of the proposed methods in the case of legitimate use (targeting questions like “can a model owner reconstruct parts of the chemical or bioactivity data of individual other parties based on model components he can legitimately access”). Public data (prepared and processed by the pharma partners using the same protocols as for their own data) can be leveraged to this end.

Expected key deliverables

- A coherent, federated, privacy-preserving machine learning method that conforms with the following requirements should be delivered by month 12 and updated annually.
    - An early software prototype is delivered to allow the algorithm to be documented and to enable an analysis of privacy preservation by the use of legitimate modelling results [at project start, we assume the existence of an early software prototype of the first version, to ensure proper maturity of proposals;
    - A report of such privacy preservation analysis using public data, listing algorithmic or parameter options to navigate performance/privacy trade-offs. This includes evaluating vulnerabilities to e.g. differential attacks. This will enable conceptual sign-off to use on the massive pharma datasets;
    - Enterprise ready code, i.e. ready for independent code audit against joint pharma security requirements (that should preclude to reasonable standards illegitimate access to or use of data or models). A favourable audit report is a prerequisite for exposure of the massive pharma datasets;
    - Ability to be run on standalone pharma data and on IT architecture for federation;
    - From the 2nd year onwards, the solutions should be able to also incorporate image-derived or transcriptional descriptors of the compounds.
- Establishment of proof-of-concept of this platform, by deploying and evaluating it in an industrial setting:
    - Standalone and cross partner runs executed and their predictive performance compared with standalone results. The algorithmic, software and ICT infrastructure choices should enable a full cross-partner run to complete in maximally four weeks;
    - Evaluation of efficiency gains when deployed in life project settings as impact evaluation;
    - At least in one exercise the inclusion vs. exclusion of such features in a federated modelling run are compared head-to-head;
    - At least in one exercise the performance of the developed methodology is compared head-to-head to a credible established non-federated single-task method (minimally support vector machine (SVM), random forest or a comparably performant method);
    - As stated in the scope, predictive models will be generated to validate the platform. Such models will be derived from the private compound and activity data during this evaluation. In order to validate the effectiveness of the platform, the models will be evaluated on their (i) predictive performance gains in terms of chemical and biological applicability domain, (ii) accuracy (implying the generation of compound activity predictions from the model), and (iii) on their impact in actual discovery projects, e.g. how many lab experiments can be replaced by *in silico* predictions;
    - Although these predictive models (and their associated predictions and impact metrics) are directly derived from the private compound and activity data from the EFPIA participants, they are only of use for validating the platform, but are not required for the actual operation of the platform, and neither required for the further research use or direct exploitation of the platform. Therefore, the aforementioned predictive models, predictions and impact metrics created during and in the context of the validation of the platform should not be considered as actual project deliverables, but rather as data generated outside of project objectives, and will be treated as sideground for the purpose of the project. Because such predictive models are generated outside of the project objectives and are directly related to a specific private compound activity data-set, the applicant consortium should feel fully comfortable to establish in the consortium agreement that the ownership of these specific predictive models, predictions and metrics generated therefrom in the context of the validation of the platform, will be transferred to the initial owner of the contributed private compound and activity data, at no additional cost.

- Sustainability plans that detail how the applicant consortium intends to make the developed methodologies accessible to the pharmaceutical industry and to other future adopters after the project ends;

- Publication and dissemination of guidelines, advice, detailed processes (workflows and specific technical details) and ICT and security standards adhered to and of the predictive performance (at an aggregated level) to promote the uptake of the developed methodologies in the pharma and other) sectors;

- Identification and publication of any barriers to the uptake of the proposed methodology and publication of solutions to reduce those barriers.

Expected impact

The *in silico* predictions developed within the project will increasingly replace the costly and time-consuming *in vitro* testing, resulting in cost and time savings on compound synthesis and measurement in assays and preclinical studies and therefore increase the efficiency of pharmaceutical discovery research. Although out of the direct scope of the present topic, the application of similar concepts to clinical data to enable faster recruitment of more targeted patients holds the longer-term promise of reducing costs of development.

By providing data owners with the confidence that their data and the corresponding predictive models will remain private, this project will facilitate access to much larger data sets and therefore improve performance over that of conventional machine learning approaches.

The concepts developed within the project will be generic and will apply not only to the pharma setting, but also to multiple alternative industrial and other commercial settings where parties are interested in different predictive models that benefit from indirect access to the same volumes of private data.

For knowledge and ICT partners, federated learning presents a line of research and product development beyond that of data federation.

Applicants should indicate how they will strengthen the competitiveness and industrial leadership of Europe by, for example, engaging suitable SMEs.

Potential synergies with existing Consortia

Applicants may consider technologies and insights from earlier or ongoing national or EU funded projects with a focus on the use of machine learning approaches in support of pharmaceutical discovery.

For example, several IMI projects have already faced the challenge of facilitating research on private data, see http://www.sciencedirect.com/science/article/pii/S1359644615004249 and http://www.mdpi.com/1422-0067/15/11/21136/html

Another IMI project aims at the systematic FAIRification of data (the capture and management of data to make them Findable, Accessible, Interoperable and Reusable). The applicant consortium is encouraged to seek synergies with projects for the FAIRification of data (e.g. consider applying learnings and technologies from such projects), but should avoid replication of such efforts.

Industry Consortium

Key contributions from EFPIA partners:

- Agreement on protocols and solutions for processing data with the necessary and sufficient level of standardisation to enable the machine learning exercises. To encourage broader adoption the partners

will opt for open solutions where possible. Insights on data standards and technologies from ongoing EU funded projects (e.g. those in the context of the FAIRification IMI topic) will be considered;

- The anticipated collective industry datasets outlined under Scope, above;
- Data management;
- Formulation of joint security requirements in line with industry standards;
- Set up independent audit of all enterprise-readied code against those requirements;
- Evaluation of the analysis of privacy preservation based on legitimately accessed models;
- Expertise in cheminformatics and machine learning at scale in the context of this topic;
- Upon enablement by the consortium (access to secure software solutions), execute provided solutions on own data (standalone);
- Evaluate predictive performance in terms of accuracy and chemical and biological applicability domains;
- Extensive experience in drug discovery and development, including knowledge, of all *in vitro* and preclinical assays modelled;
- Follow-up of activity predictions in life projects to assess their impact, including aggregate level reporting;
- Expertise in image and omics analysis, to facilitate the accommodation of image or transcriptional information in the developed methods;
- Project management (centrally managed by a subcontracted project coordination office);
- Dissemination activities within the sector.

Indicative duration of the action

The indicative duration of the action is 36 months.

Future Project Expansion

Potential applicants must be aware that the Innovative Medicines Initiative 2 (IMI2) Joint Undertaking, may publish at a later stage another call for proposals restricted to those projects already selected under this call in order to enhance and progress their results and achievements by extending their duration and funding. Consortia will be entitled to open to other beneficiaries as they see fit. If proof-of-concept in terms of privacy and predictive performance is established, there is the possibility that it could be extended to clinical datasets.

Applicant Consortium

The applicant consortium will be selected on the basis of the submitted short proposals.

The applicant consortium is expected to address all the research objectives and make key contributions to the defined deliverables in synergy with the industry consortium which will join the selected applicant consortium in preparation of the full proposal for stage 2. Therefore, the applicant consortium should be able to demonstrate the full scope of experience and expertise needed to effectively address all goals outlined in this topic. The size of the applicant consortia should reflect the expertise needed to achieve the proposed objectives within the indicated budget while ensuring the ‘’manageability” of the consortium as well as efficient and effective team work. Therefore, the number of members of the applicant consortium needs to be thoroughly justified in the proposal and all partners involved should make a significant contribution to the project.

To meet the ambitions of the topic and ensure a first version can be deployed by the end of year one, the applicant consortium should describe the workhorse algorithms they intend to use in their short proposal,

convincingly demonstrating their compatibility with the type of data made available for this topic and with the proposed federated and privacy-preserving machine learning concepts, preferably with a (not necessarily secure or enterprise ready yet) software prototype.

Given the runs will involve the handling of private preclinical data sets at an unprecedented scale, the applicant consortium is expected to mobilise across academia, SMEs and other commercial organisations as appropriate, the following

- Demonstrated extensive hands-on expertise in solutions for big data handling at industrial scale and following;
- Demonstrated extensive hands-on expertise in ICT security and information leakage aspects;
- Demonstrated extensive hands-on expertise with deployment on high performance computing infrastructures;
- Demonstrated extensive hands-on expertise in software engineering;
- Demonstrated extensive hands-on expertise in machine learning technologies, including in the context of federated learning;
- Demonstrated hands-on expertise of deploying computational approaches in the context of drug design, drug discovery and development;
- Demonstrated hands-on expertise in general project management (ability to consistently set and achieve milestones on time and within budget; managing varying interests of multiple stakeholders) and professional communication (expertise in communication tools and systems for project management purposes), in the context of EU-funded projects.

The short proposal should include a description as to how the applicant consortium intends to make the developed methodologies accessible to the pharmaceutical and other industries after the project ends. To this end, it is suggested to include a party responsible for ensuring sustainability (including software, licensing, infrastructure options, potential broker services). While a broker role is acceptable, and could for example be filled by an SME, this role must be compatible with the outlined interpretation of federated and privacy-preserving machine learning, for instance the broker function will not have access to assay data, annotation or the corresponding models.

Suggested architecture of the full proposal

The applicant consortium should submit a short proposal which includes their suggestions for creating a full proposal architecture, taking into consideration the Industry consortium contributions and expertise provided below.

In the spirit of the partnership, and to reflect how IMI2 JU call topics are built on identified scientific priorities agreed together with EFPIA beneficiaries/large industrial beneficiaries, these beneficiaries intend to significantly contribute to the programme and project leadership as well as project financial management.

The final architecture of the full proposal will be defined by the participants in compliance with the IMI2 JU rules and with a view to the achievement of the project objectives. The allocation of a leading role within the consortium will be discussed in the course of the drafting of the full proposal to be submitted at stage 2. To facilitate the formation of the final consortium, until the roles are formally appointed through the consortium agreement, the proposed project leader from among EFPIA beneficiaries/large industrial beneficiaries shall facilitate an efficient negotiation of project content and required agreements.

All beneficiaries are encouraged to discuss the project architecture and governance and the weighting of responsibilities and priorities therein. The below architecture for the full proposal is a suggestion; different innovative project designs are welcome, if properly justified.

Work package 1 –Preprocessing of data up to a level of necessary and sufficient standardisation

- Select methodology for standardised preprocessing data and implement in scripts, including feature extraction, dimensionality reduction, weighted data integration;
- Enablement of participants to deploy scripts in standardised ways compatible with the architectures proposed for the exercise;
- Execute preprocessing of data and making them available (including public data for work package 3).

Industry consortium contribution:

Methodology selection, implementation and execution.

Expected applicant consortium contribution:

Enablement of architecture-compatible deployment, scientific advice on other bullets.

Work package 2 – Industrial IT technical scoping and deployment

- Joint pharma user requirements;
- Independent software audit of the resulting software (from work package 5);
- Enablement of runs on ICT infrastructure under pharma control.

Industry consortium contribution:

Formulation of user requirements, set-up of audit, enablement of runs.

Expected applicant consortium contribution:

Liaison between pharma driven work package 2 and consortium driven WP5 (software implementation), to ensure solutions match requirements and can be run on pharma controlled infrastructures.

Work package 3 – Federated Machine Learning Algorithms

- Development and scientific and software prototyping of the algorithm;
- Initial predictive performance estimation (on public data);
- Machine learning security analysis of algorithms (on public data), to enable security evaluation.

Industry consortium contribution:

Experts in machine learning applied to the domain of the topic.

Expected applicant consortium contribution:

Main drivers and executers of all the above.

Work package 4 – Evaluation of privacy and performance balance and of predictive performance of the versions up to implementation in discovery projects

- Evaluation of balance performance and privacy preservation balance (on prototypes);
- Retrospective and prospective evaluation of prediction performance (enterprise-ready product) and of impact of use of the models in discovery projects (up to metrics like ‘how many lab experiments were avoided/substituted by predictive models).

Industry consortium contribution:

Main drivers

Expected applicant consortium contribution:

Scientific support

Work package 5 – Software Implementation

- Initial software conceived and prototyped in work package, and evaluated for privacy/performance balance in WP4 (scientific) and WP2 (data privacy), to be readied to the point it can be securely deployed on the massive pharma datasets;
- This includes aspects of software engineering, ICT security, knowledge of ICT infrastructure to run on, with respect to software implications (high performance computation enablement, hardware acceleration, …).

Industry consortium contribution:

Industrial experts in ICT, security, machine learners and modelling.

Expected applicant consortium contribution:

Main drivers and executers of the implementation.

Work package 6 – Secure Federated Infrastructure

- Central ICT infrastructure that can connecting to the infrastructures under control of the partners involved, ensuring security and performance requirements;
- Operation Support.

Industry consortium contribution:

Industrial experts in ICT.

Expected applicant consortium contribution:

Selecting and setting up the secure federated infrastructure.

Work package 7 – Operations and Deployment

- Establish a detailed software and operating model with pharma organisations;
- Overseeing execution of runs upon initiation by pharma.

Industry consortium contribution:

Industrial experts in ICT and modelling.

Expected applicant consortium contribution:

Main drivers, may include partners involved in sustainability plans.

Work package 9 – Overall project governance, project management, dissemination and sustainability

- Grant administration;

- Strategic, operational, IP and financial management;
- Communication (within the consortium and with relevant external collaborators);
- Dissemination of scientific results and research data to the scientific community and within the pharma sector;
- Detailed sustainability plan to make results accessible beyond the duration of the action.

Industry consortium contribution:

Programme leadership with respect to application and valorisation aspects, project and financial management (setting up subcontract for a professional service provider), contribution to communication and dissemination.

Expected applicant consortium contribution:

Scientific and technical programme coordination, reporting to the commission (supported by the industry-sponsored project management service provider).