IMPROVING ADMET MODEL PERFORMANCE BY COLLABORATING ON PROPRIETARY DATA

Challenge: Limited training data and a lack of diversity limit the performance and applicability domain of ADMET prediction models

Predicting ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) endpoints is essential for drug optimization, yet the limited availability of chemically diverse, high-quality datasets constrains the development and application of QSAR (Quantitative Structure-Activity Relationship) models. Recent advancements in foundation models based on graph-based deep learning have enhanced structure-based drug discovery. However, accurately predicting ADMET endpoints remains challenging due to the inherent complexity of these processes, uneven experimental scalability across endpoints, and limitations in the size, diversity, and chemical relevance of available datasets, particularly in early-stage research. These factors constrain the robustness and transferability of QSAR models across chemical spaces, resulting in variable predictive performance. Consequently, while ADMET modelling approaches are widely adopted, their reliability and impact on decision-making remain uneven and highly context-dependent.

Solution: Federated network for collaborative training of ADMET models while protecting data confidentiality

There is an alternative to training models only on public and your own data: secure, federated learning on the data of multiple parties. Pharmaceutical companies form a network to enable training on their proprietary chemical data without the need for data sharing. Apheris has launched a network of pharmaceutical companies to improve model performance by training secure federated models on members' ADMET data, while keeping the confidentiality and IP of the data sets fully protected.

Objectives

The ADMET Network seeks to:

- Expand the applicability domain and performance of state-of-the-art ADMET models
- Enable members to further fine-tune these federated models on their proprietary chemical space

Endpoints

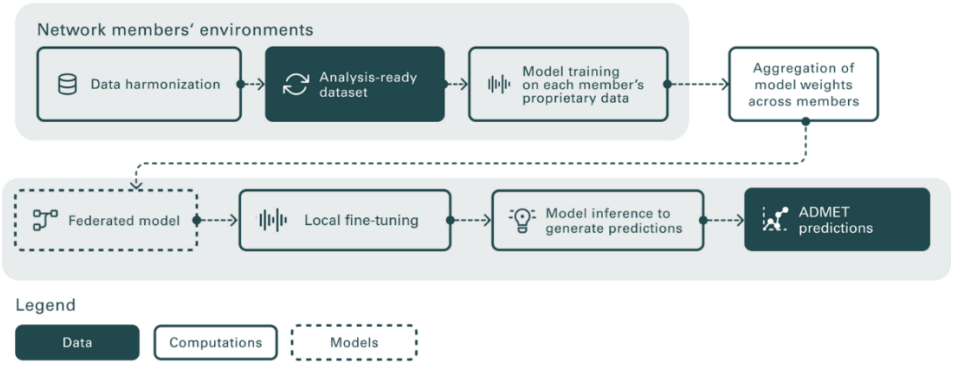
The network is launched with five founding members and a clearly defined starting set of ADMET endpoints (see exemplary endpoints below), established as a practical baseline for initial development and benchmarking.

- Aqueous solubility
- Permeability
- Metabolic stability & intrinsic clearance
- hERG liability
- Lipophilicity
- Tissue binding
- CYP inhibition of major isophorms

In parallel, discussions with network participants on expanding this scope are already underway, with multiple additional endpoints actively being considered based on expressed scientific interest.

The first federated training runs take place with combined datasets of at least 30,000 data points for common ADMET endpoints. The network is open to prospective members with a broad range of potential data contributions in terms of volume and diversity.

End-to-end workflow



- First, each network member's proprietary SMILES data linked to assay results and descriptions is pre-processed and aligned to each member's environment to create analysis ready datasets.
- Next, a deep learning ADMET model is trained across partner datasets via the Apheris federated computing product to produce a joint network model.
- Then the privacy of the network model will be assessed before weights will be shared with the network members.
- Typically, the federated weights are then further fine-tuned for individual network members on further data to optimize performance in their programs.
- Finally, the federated model or the fine-tuned models can be accessed for running model inference and benchmarking against in-house models.

Data and IP protection

Apheris has worked on facilitating collaborations among companies, while ensuring privacy and security, since 2019. In previous collaborations, we worked with molecular data of different parties to train graph- and transformer-based deep learning models. Here we conducted sophisticated privacy assessments (particularly running attacks geared at reverse engineering the trained models) and checked that models were not susceptible to such attacks. We have thorough expertise in both assessing privacy risks of machine learning models and mitigating them.

For the end-2-end workflow above, this means:

- Data: Data always stays under the full control of the network member that contributes it. Data is never directly shared with others, and the federated computing product ensures that only permitted computations can be carried out on it.
- Model weights: Only model weights leave a network member's environment following local training; the underlying data never does. Model parameters from all participating members are aggregated to form a shared network model. This aggregated model is then systematically evaluated for privacy risks, including exposure to known privacy and inference attacks, to assess the likelihood of sensitive training data leakage and to define appropriate mitigation measures where required.
- Inference: The resulting model is made available for local inference, while limiting the attack surface for potential reverse engineering. Importantly, the prompts submitted by network members to the model remain local and private.

Interested in learning more?

We invite you to an initial conversation to discuss the network proposal in more detail and clarify any open questions. Please reach out Julian Schönauer (j.schoenauer@apheris.com).