Aligning data from public and proprietary sources to develop federated QSAR models

Adrian Fowkes, Andrea Sartini, Jeffrey Plante, Robert Davies, Stephane Werner, Thierry Hanser
Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK

Lhasa Limited

Abstract #200

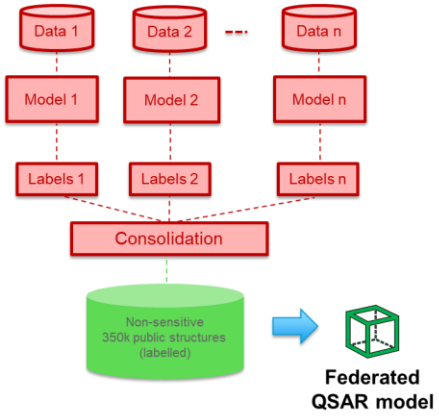
1. Background

Problem: Large sources of high-quality and relevant data are inaccessible for modelling as they reside in proprietary data silos

Outcome: User's QSAR models may struggle to generate confident predictions as they venture into new areas of chemical space

OUTSIDE DOMAIN ☹

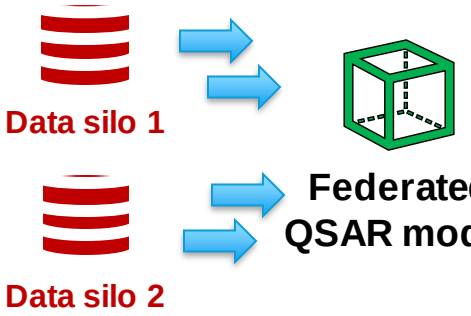
2. Federated learning can distil knowledge from silos



Experiment: 8 pharmaceutical companies federated data and trained a new QSAR model for hERG inhibition

Result: The federated model outperformed every other model trained on a single source of data

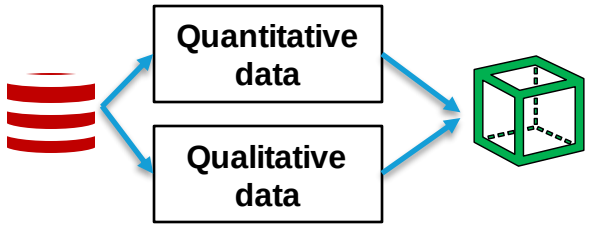
3. Data for generating federated models needs to be aligned



***Q**. Which endpoints and assays?*
***A**. Survey of consortium performed*

***Q**. Which thresholds to use?*
***A**. Target-specific thresholds*

4. Maximising data



Models can be improved by combining different types of bioactivity data

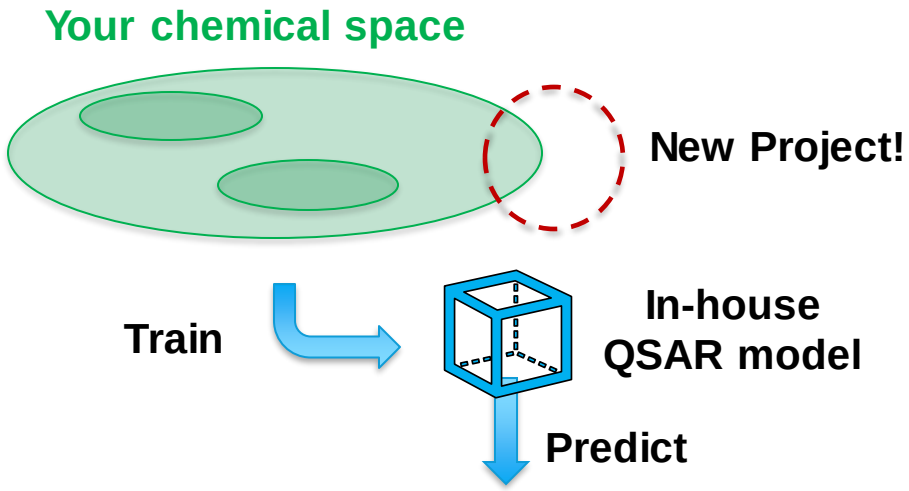
Future Perspectives

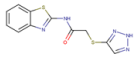
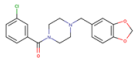
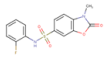
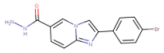
Lhasa has developed an application called Effiris to enable federated learning

Effiris

QSAR models benefit from relevant data

- The applicability domain of QSAR models is in part governed by the training set
- Therefore, decision making based on QSAR methods is limited to knowledge built on in-house data

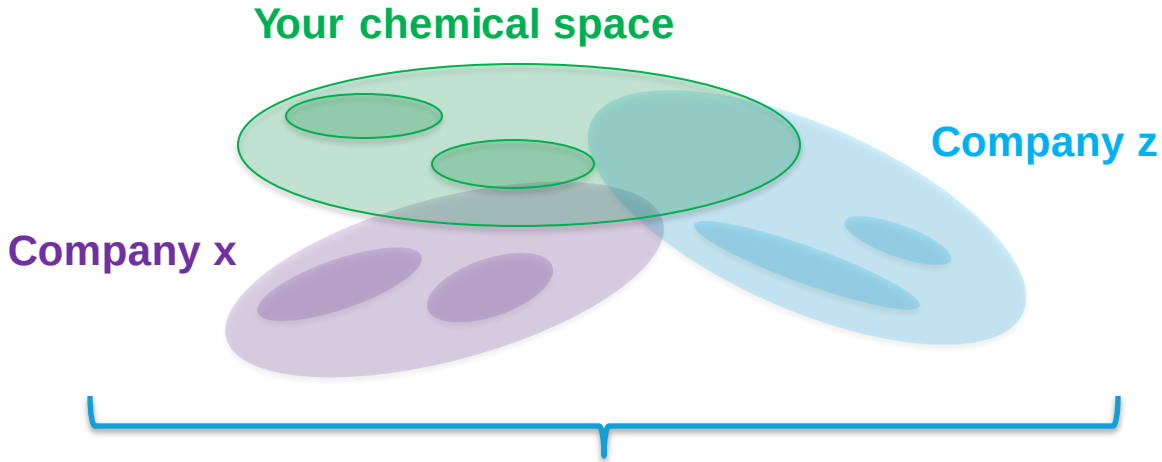


Structure	Primary target	Secondary pharmacology profile			LogP
	8.2	?			3.6
	7.6	?	?	?	3.8
	8.1		?	?	4
	7.2	?	?		3.7

In silico prediction

Positive
Negative
? Low confidence

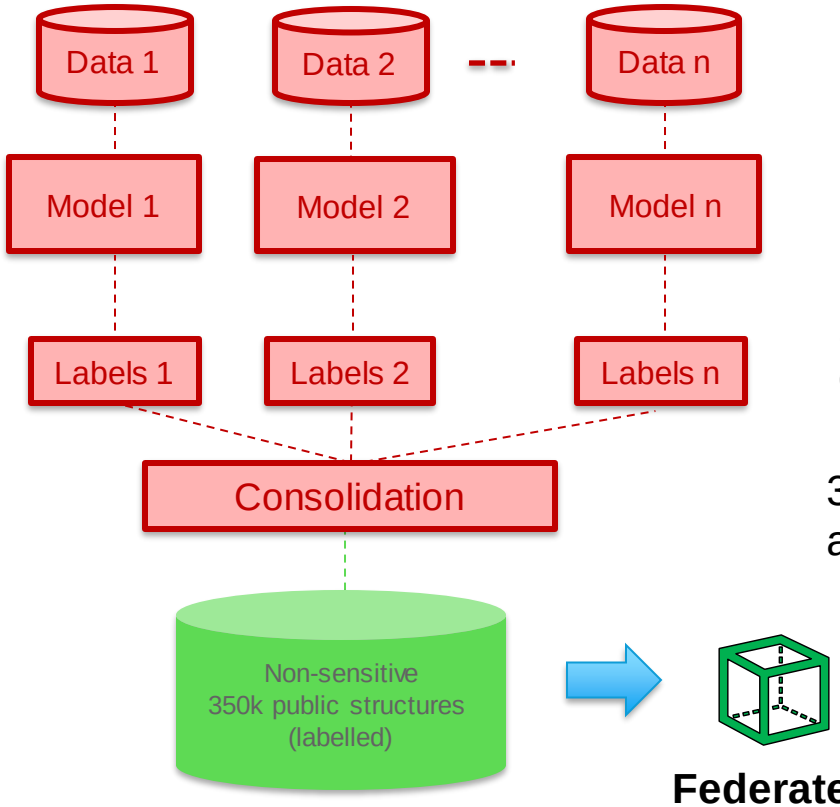
High-quality and relevant data resides in proprietary data silos



Federated learning is a privacy-preserving approach to distil and share knowledge between organisations in a pre-competitive manner

hERG Inhibition: Proof-of-Concept Study

8 pharmaceutical companies were involved in a proof-of-concept study modelling hERG inhibition



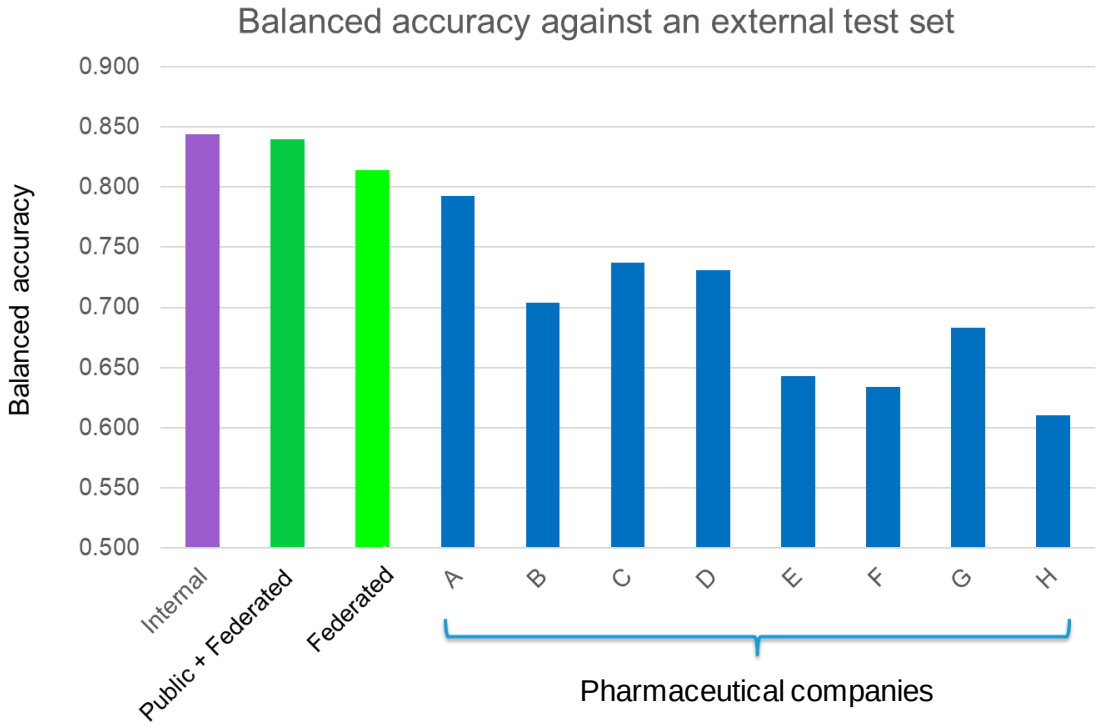
Federated QSAR model

Method

1. Train models on proprietary data

2. Predict activity on public structures

3.Consolodate labels and train model



Models trained from multiple sources through federated learning (green bars) outperformed every single model trained on a single source of data (blue bars). The performance of the federated models is similar to the internal validation of the test set (purple bar).

1) N. Papernot *et al. 2017* https://arxiv.org/abs/1610.05755 2) R. Preissner *et al. 2018* https://doi.org/10.1021/acs.jcim.8b00150

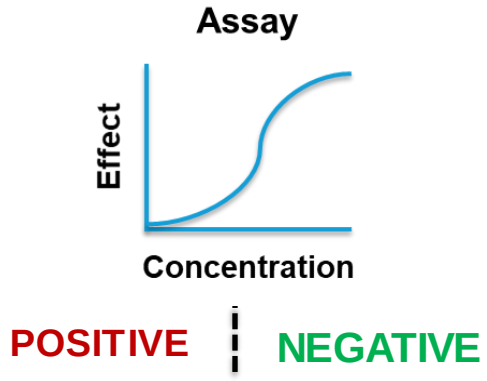
Aligning data for secondary pharmacology profiling

- Given the success of the proof-of-concept study, the approach has been extended to federate data across multiple partners focusing on secondary pharmacology profiling
- To enable federation of data across multiple endpoints, alignment is required across the consortium to establish what to model and how to model it:

Industry survey to identify priority endpoints

Established target-specific thresholds to provide meaningful outputs. Thresholds informed by:

1. Assay survey identifying the protocols used and the subsequent decisions made



2. Potency of known ligands

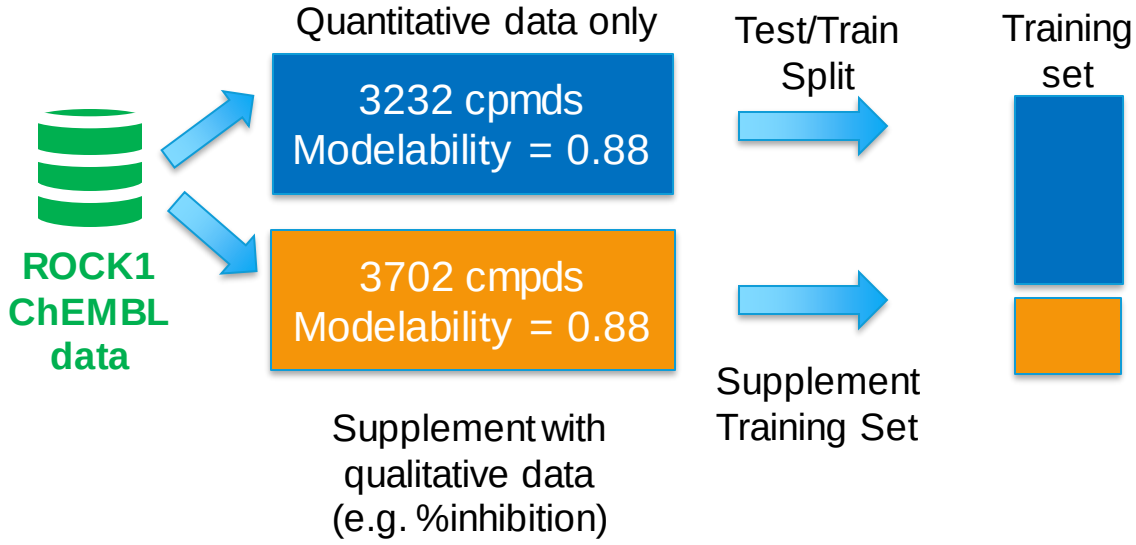
Example receptor	pX_{50}
Median potency for known drugs	6.5
Endogenous ligand	8.5

1) Bowes *et al*. *2012* https://doi.org/10.1038/nrd3845 2) Lynch *et al. 2017* https://doi.org/10.1016/j.vascn.2017.02.020
3) Bofil *et al. 2019* https://dx.doi.org/10.1016%2Fj.drudis.2019.06.007

Lhasa Limited

Maximising data & Future perspectives

Models can be improved by combining different types of bioactivity data



	Quantitative data only		Supplemented training set	
Test	MCC	Coverage	MCC	Coverage
Random split	0.75	0.65	0.76	0.68
Temporal split	0.13	0.52	0.33	0.56

Combining different datatypes produced a model with more knowledge of bioactivity compared to a model solely trained on quantitative data

Future Perspectives

- Lhasa Limited continues to develop an application called Effiris which enables the federation of data to train new models with greater knowledge of bioactivity whilst preserving the privacy of user's data
- Effiris will be developed to further support secondary pharmacology use-cases
- Future research will examine the potential to build federated regression models using this technology

Please get in touch if you would like to know more… adrian.fowkes@lhasalimited.org