Abstract # 3412

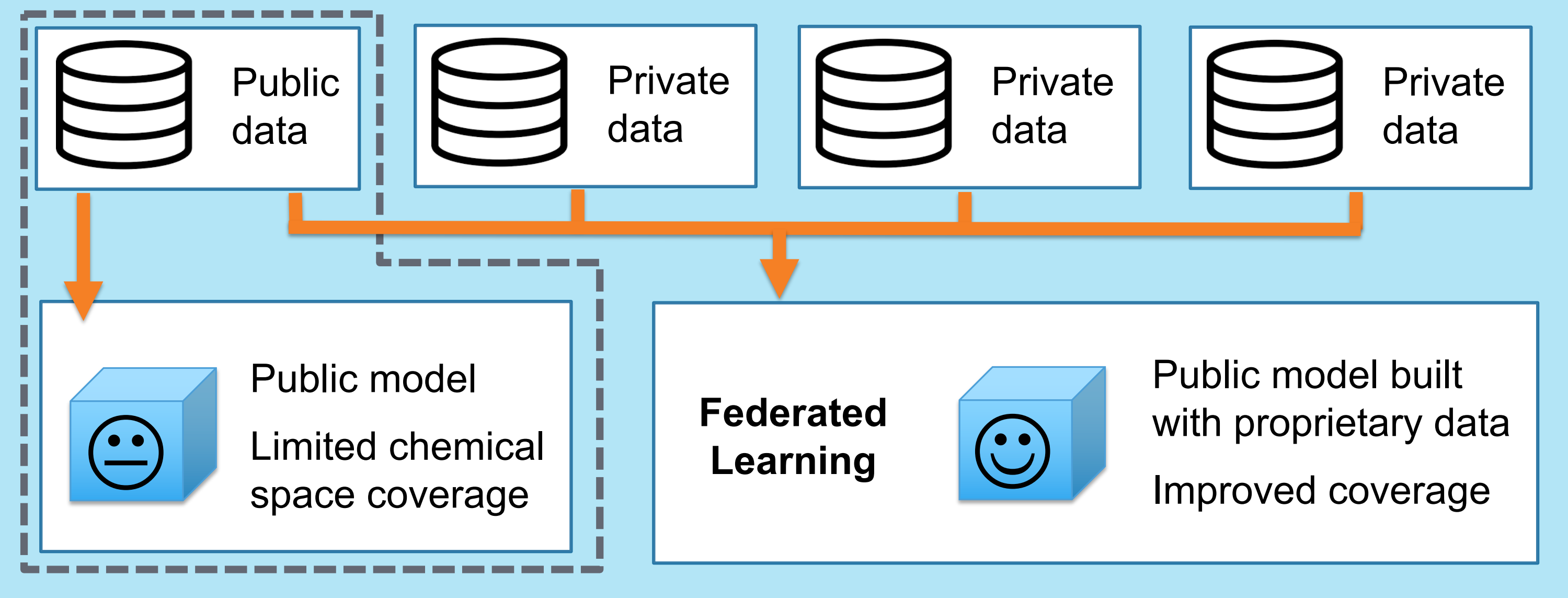
Consortium-led Federated QSAR Models for Secondary Pharmacology – Preparing the Data

Robert Davies, Adrian Fowkes, Richard Williams, Laura Johnston.
Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS

QSAR Models for Secondary Pharmacology

Challenge: Quantitative structure-activity relationship (QSAR) models trained on a single data source tend to have limited coverage against data from other sources. This is a particular issue when predicting proprietary data using models trained on public data.

Approach: Federated QSAR models trained on multiple data sources will produce public models covering a wider area of chemical space (Figure 1).



Public data

Private data

Private data

Private data

Public model

Limited chemical space coverage

Federated Learning

Public model built with proprietary data

Improved coverage

Figure 1. Overview of the generation of federated QSAR models trained on multiple data sources.

This work: Defining model endpoints and the preparation of datasets from the public domain, to support the training of public federated models which have also learnt from proprietary datasets from Lhasa Limited members.

Target Prioritisation & Data Sources

Numerous targets associated with adverse outcomes have been used to design robust *in vitro* secondary pharmacology screens. These targets of concern for drug development have been published in reviews including Bowes *et al.*[1] and Lynch III *et al.*[2]. In addition to these targets, Lhasa Limited members were surveyed for targets of high interest with respect to secondary pharmacology screening (Figure 2). To extract relevant data for these targets, a semi-automated workflow was built to retrieve data from publicly available bioactivity databases (ChEMBL[3] and ExCAPE-DB[4]) and build preliminary models (Figure 3).

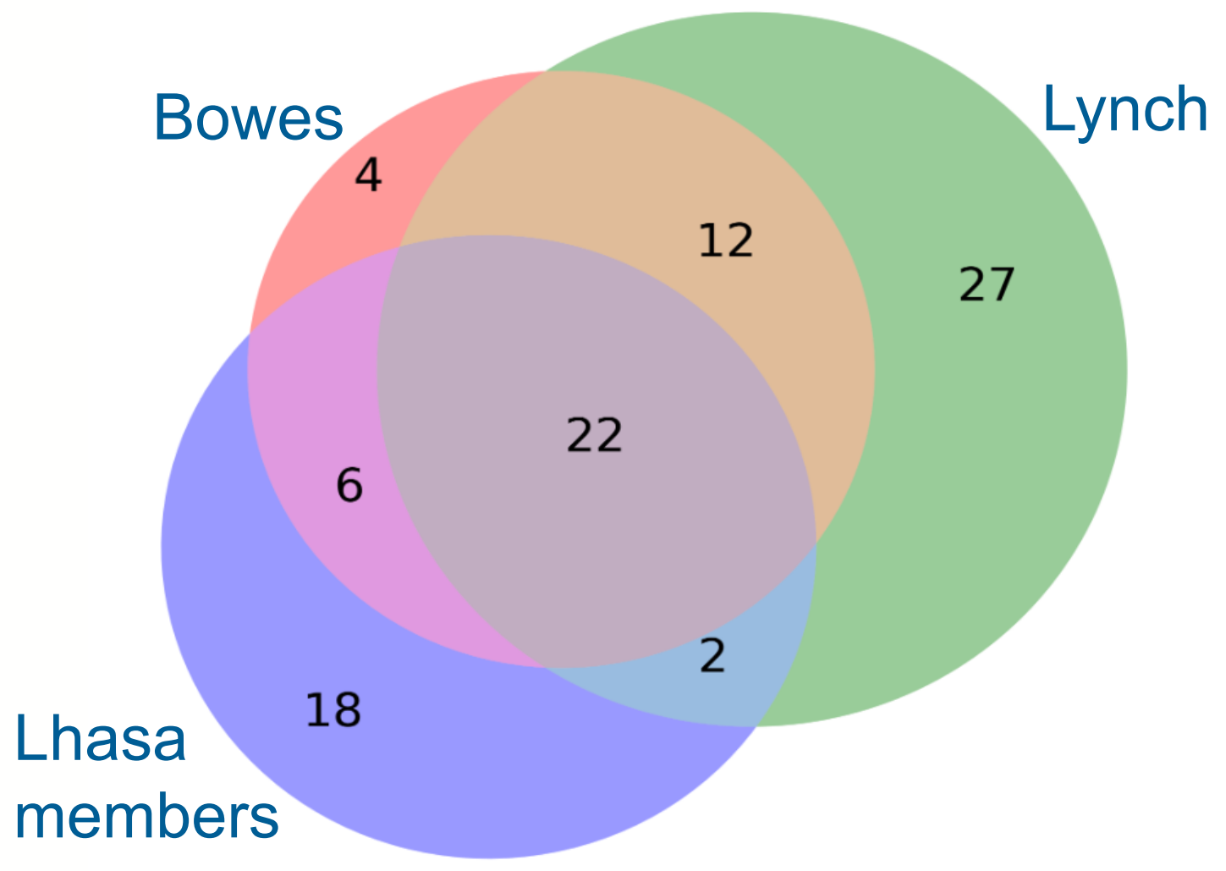
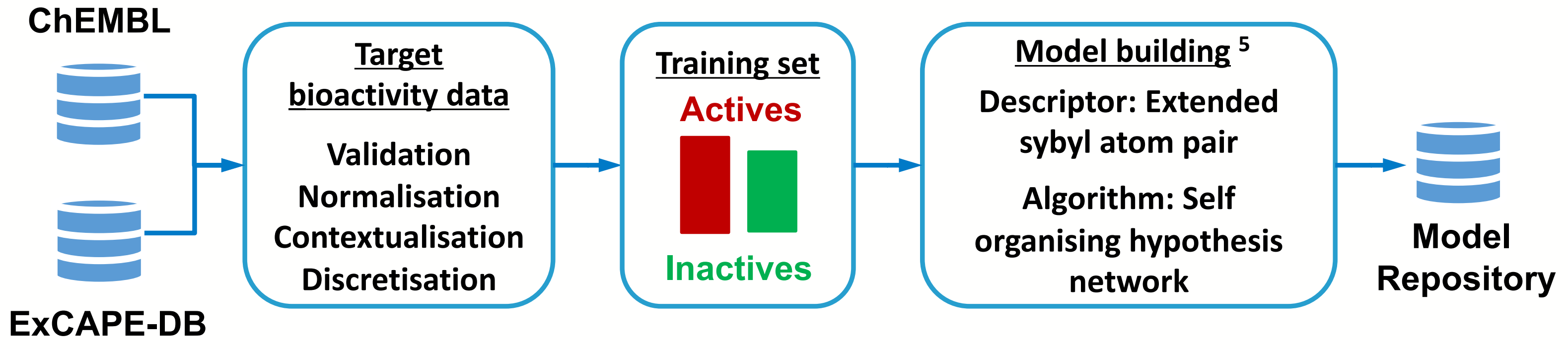


Figure 2. Overlap of prioritised targets between literature sources and Lhasa Limited members.



ChEMBL

ExCAPE-DB

Target bioactivity data

Validation

Normalisation

Contextualisation

Discretisation

Training set

Actives

Inactives

Model building [5]

Descriptor: Extended sybyl atom pair

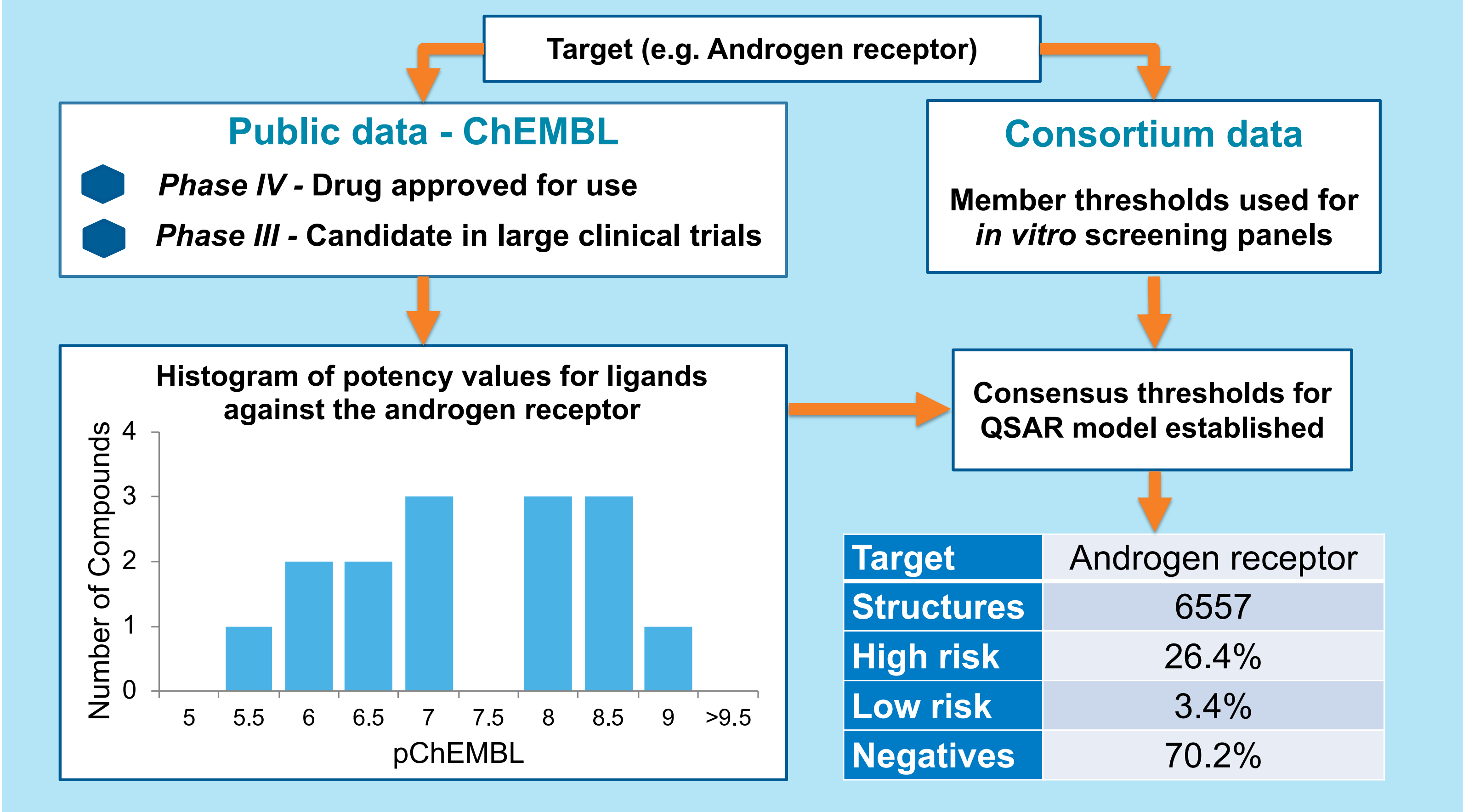
Algorithm: Self organising hypothesis network

Model Repository

Figure 3. Overview of a semi-automated workflow for handling data and building public models.

Classifying Compounds for Modelling

Qualitative models require thresholds to distinguish between different compound classes. Ideally, these thresholds should be relevant to decision making. The establishment of thresholds can be influenced by the potency of reference compounds and assay sensitivity. To define thresholds for federated models, data from the public domain and knowledge from Lhasa members were used to define inactive, low-risk and high-risk compounds (Figure 4).



Target (e.g. Androgen receptor)

Public data - ChEMBL

Phase IV - Drug approved for use

Phase III - Candidate in large clinical trials

Consortium data

Member thresholds used for *in vitro* screening panels

Histogram of potency values for ligands against the androgen receptor

Consensus thresholds for QSAR model established

Target	Androgen receptor
Structures	6557
High risk	26.4%
Low risk	3.4%
Negatives	70.2%

Figure 4. Workflow to define thresholds for federated QSAR models.

Training Set Composition and Initial Models

Five targets from different protein classes were selected for further investigation. The public datasets and initial models trained on the datasets are presented below (Table 1 and Table 2).

Table 1. Composition of the training sets

Target	Threshold	Data sources	Structures	Actives
Androgen receptor	Low risk	ChEMBL & PubChem	4872	50.0%
Adenosine A2a receptor	Low risk	ChEMBL	5637	69.8%
Cyclooxygenase-II	High risk	ChEMBL	3896	52.4%
Dopamine D2 receptor	Low risk	ChEMBL & PubChem	19440	50.0%
Serotonin transporter	High risk	ChEMBL	7506	64.9%

Table 2. Performance of the models assessed by 4:1 cross-validation.

Target	BA	SENS	SPEC	PPV	NPV	COV
Androgen receptor	0.93	0.97	0.90	0.92	0.96	0.73
Adenosine A2a receptor	0.76	0.94	0.58	0.86	0.79	0.84
Cyclooxygenase-II	0.81	0.85	0.78	0.85	0.78	0.75
Dopamine D2 receptor	0.96	0.99	0.93	0.94	0.99	0.90
Serotonin transporter	0.80	0.92	0.70	0.85	0.82	0.87

BA = Balanced accuracy, SENS = Sensitivity, SPEC = Specificity, PPV = Positive predictivity, NPV = Negative predicitivity, COV = Coverage.

Improving Model Performance

Many factors influence the performance of a model, including the composition of the dataset, the descriptors used, and the modelling algorithm deployed. To assess the impact of the training data on the performance of the model, the composition of the dataset was varied, and the model performance was assessed by cross-validation (Table 3 and Figure 5). The analysis shows that datasets generated from the public domain can produce performant models, indicating that increased feasibility of producing a federated QSAR model.

***Table 3.** Composition of training sets for different adenosine A2a receptor models.*

Dataset	Description of Dataset	Compounds	Actives
v1.0	Dataset generated by original workflow (Figure 3).	5637	69.8%
v1.1	Negatives in ChEMBL obtained from single concentration screens added to v1.0	8177	59.6%
v1.2	Dataset v1.1 balanced by under sampling major class	6604	50.0%

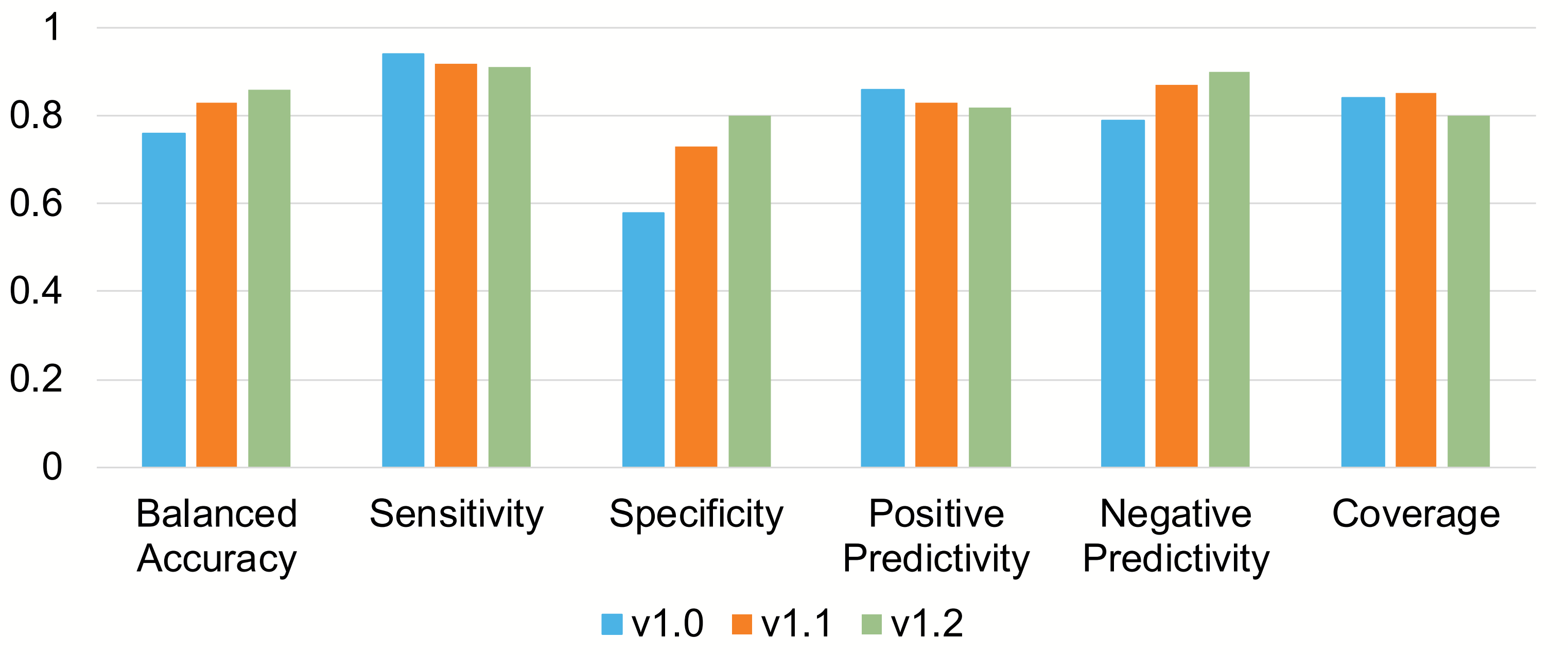


Figure 5. Performance of models for the adenosine A2a receptor trained on different datasets assessed by 4:1 cross-validation.

Conclusions & Future Work

The collaboration between Lhasa and its members is helping define the properties of federated qualitative QSAR models, which have been trained on multiple proprietary datasets. Workflows have been generated to curate data from the public domain that can contribute to the training of federated models, that will allow users to cover wider areas of chemical space during profiling for secondary pharmacology.

The next steps for the consortium is to distill knowledge from each proprietary dataset to lead to the production of federated QSAR models. These models can then be validated prospectively as new data is generated. The models generated in this study can be deployed in Effiris, a container of models for secondary pharmacology endpoints.

References

[1] Bowes *et al.* Nat. Rev. Drug Discov., 2012, 11, 909-922; [2] Lynch III *et al.* J. Pharmacol. Toxicol. Methods, 2017, 87, 108-126; [3] Gaulton *et al.* Nucleic Acids Res., 2017, 45, D945-D954; [4] Sun *et al.* J. Cheminform., 2017, 9:17; [5] Hanser *et al.* J Cheminform, 2019, 11:9.

shared **knowledge** • shared **progress**