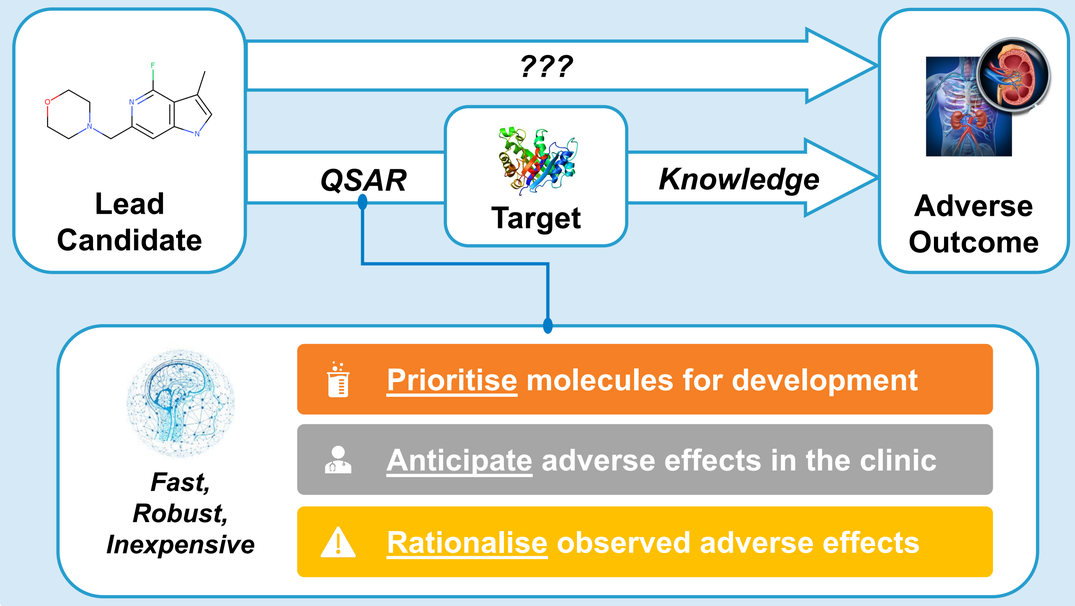
Utilising ChEMBL Data for Building Secondary Pharmacology Models

Robert Davies, Adrian Fowkes, Steven Kane, Jeffery Plante, Andrew Thresher, Lilia Fisk, Richard Williams

Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS

1 QSAR models for secondary pharmacology

Adverse drug reactions (ADRs) are often the consequence of a direct interaction between a drug and a specific protein. Many proteins have been identified as the loci of molecular initiating events (MIEs) leading to ADRs and these relationships have been published in the literature[1]. This knowledge can be leveraged in screening strategies to identify toxicity liabilities present on drug candidates using quantitative structure-activity relationship (QSAR) models.

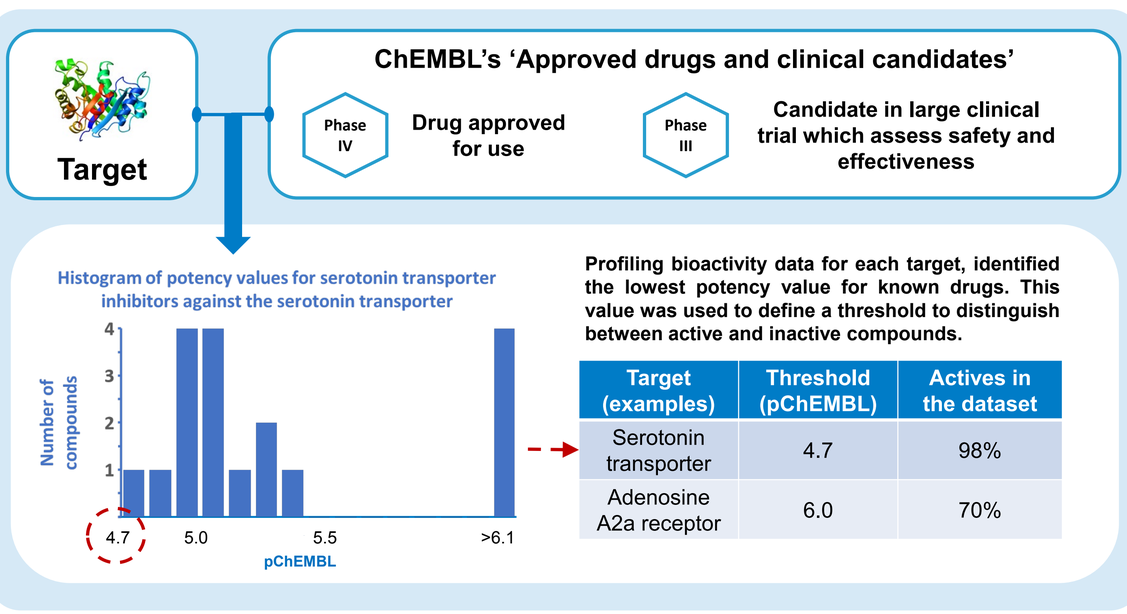


Study aim

Develop classification QSAR models for targets identified by Bowes *et al.*[1], which are suitable for secondary pharmacology workflows

2 What do 'active' compounds look like?

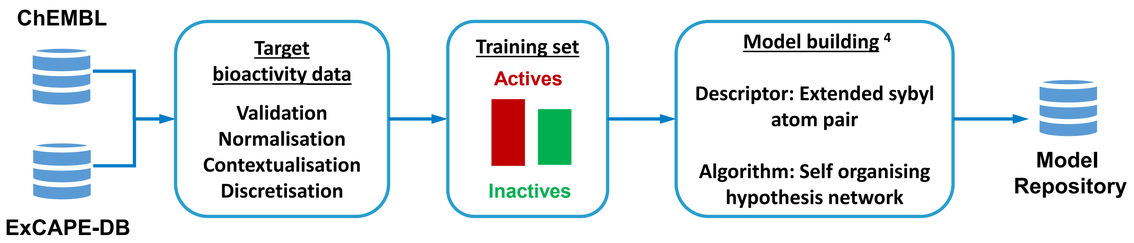
The ChEMBL database[2] contains quantitative bioactivity data. Therefore, thresholds are required for training classification QSAR models. The establishment of thresholds is a subjective process and different approaches can be applied. As the MIEs described by Bowes *et al.*[1] are also pharmacological targets, the potency values of drugs designed to interact with the target were profiled using *in vitro* data present in the database. The lowest potency value was identified from these compound sets and used to establish a threshold for discretising molecules between 'active' and 'inactive' classifications.



Target (examples)	Threshold (pChEMBL)	Actives in the dataset
Serotonin transporter	4.7	98%
Adenosine A2a receptor	6.0	70%

3 Data & model pipeline

Bioactivity data is often stored at the assay level. Therefore, data curation is required to support the generation of target-specific models, as opposed to assay-specific models which can have limited applicability. Bioactivity data were extracted from ChEMBL and ExCAPE-DB[3] and curated to support the generation of training sets and QSAR models. Additional data from ExCAPE-DB was derived from high-throughput screens and typically contains inactive compounds and therefore is able to support the generation of balanced datasets.



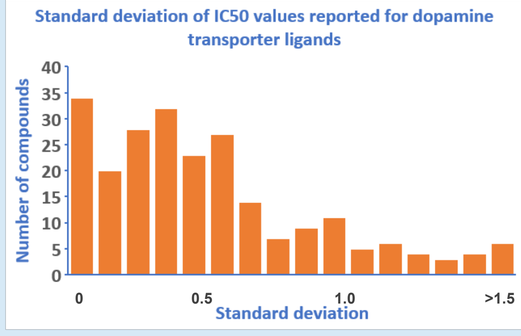
Number of targets In Bowes *et al.*[1]	Number of Bowes *et al.* targets in ChEMBL associated with >500 compounds	Number of Bowes *et al.* targets with additional data in ExCAPE-DB
44	38	19

References

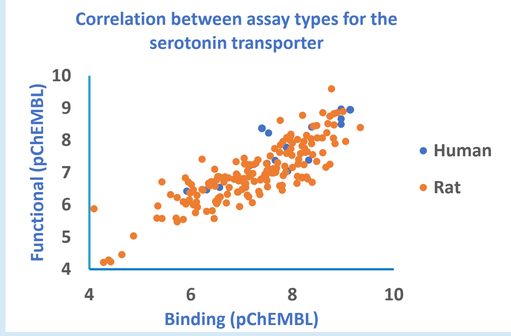
[1]Bowes *et al.* Nat Rev Drug Discov, 2012, 11, 909-922 [2] Gaulton *et al.* Nucleic Acids Res, 2017, 45, D945-D954 [3] Sun *et al.* J Cheminform, 2017, 9:17 [4] Hanser *et al.* J Cheminform, 2019, 11:9 [5] Golbraikh *et al.* J Chem Inf Model, 2014, 54, 1-4 [6] Hanser *et al.* SAR & QSAR in Environmental Research, 2016, 27, 865-881.

4 Data analysis & model evaluation

The wealth and depth of data in ChEMBL enables model builders to assess the suitability of target bioactivity for inclusion in a QSAR model's training set. The following plots highlight types of analyses that can be performed on both the underlying data in ChEMBL and models trained on ChEMBL data:

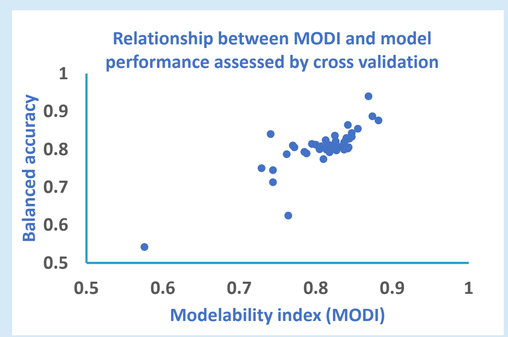


'Do similar assays produce similar results?' Compounds which have been tested multiple times in similar assays can be used to examine the reproducibility of the bioactivity results.

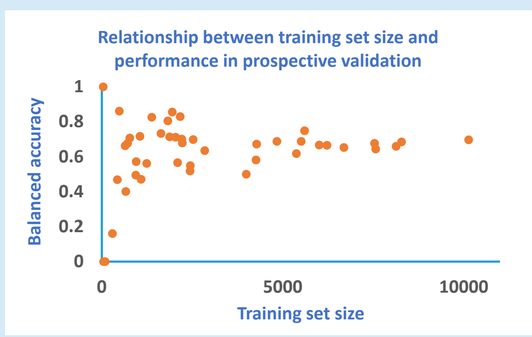


'Are different assay types for the target concordant?' Compounds which have been tested in different assay types can be used to determined the concordance between assay types.

Analysing the data in this manner enables semi-automated workflows or model developers to determine which data is suitable for inclusion in the model's training set.



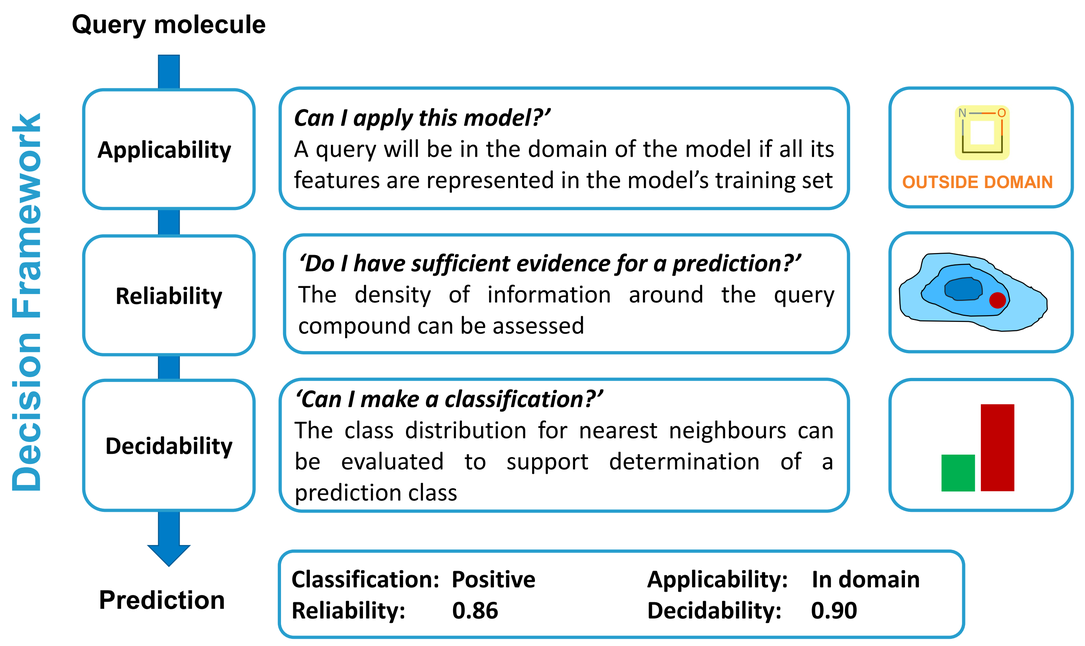
'Is my dataset modelable?' Generating models from datasets can be resource intensive. Descriptors such as the modelability index (MODI) can give an indication of the potential of generating QSAR models. Golbraikh *et al.*[5] state that datasets with a MODI > 0.65 are modelable, indicating that datasets generated from ChEMBL are suitable for generating QSAR models for secondary pharmacology targets.



'Does my model correctly classify compounds published more recently?' Creating test sets comprised of compounds published more recently, allows validation exercises to be more reflective of real-world scenarios. Typically, secondary pharmacology workflows are designed to predict for new chemical series, which are often poorly represented in model training sets.

5 Integration into decision workflows

The models generated in this study adhere to a transparent and interpretable applicability domain framework developed at Lhasa Limited[6]. The framework enables users to rationalise the output of the model and therefore subsequent decision making can be made with increased confidence.



The applicability domain is established by the model builder and defines whether a query compound is in the domain of the model or not. The reliability and decidability metrics are on a continuous scale and enable the user to tune how the model operates in their secondary pharmacology workflows.

6 Conclusions and future work

The ChEMBL database enables relevant bioactivity data to be rapidly analysed and support the training for QSAR models suitable for secondary pharmacology workflows. Future work will focus on:

- **Using federative distillation** to incorporate proprietary training data
- **External validation** of the models against proprietary data
- **Deploying** models in Effiris – a model container developed by Lhasa Limited
- **Developing additional QSAR models** for targets of high priority
- **Generating semi-quantitative** predictions to further support risk assessments

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