Combining various open data sources for P-gp/BCRP selectivity profiling

Barbara Zdrazil, Floriane Montanari, and Gerhard F. Ecker

University of Vienna, Dept. of Pharmaceutical Chemistry, Pharmacoinformatics Research Group, Althanstraße 14, 1090, Vienna, Austria

Email to: barbara.zdrazil@univie.ac.at

Background:

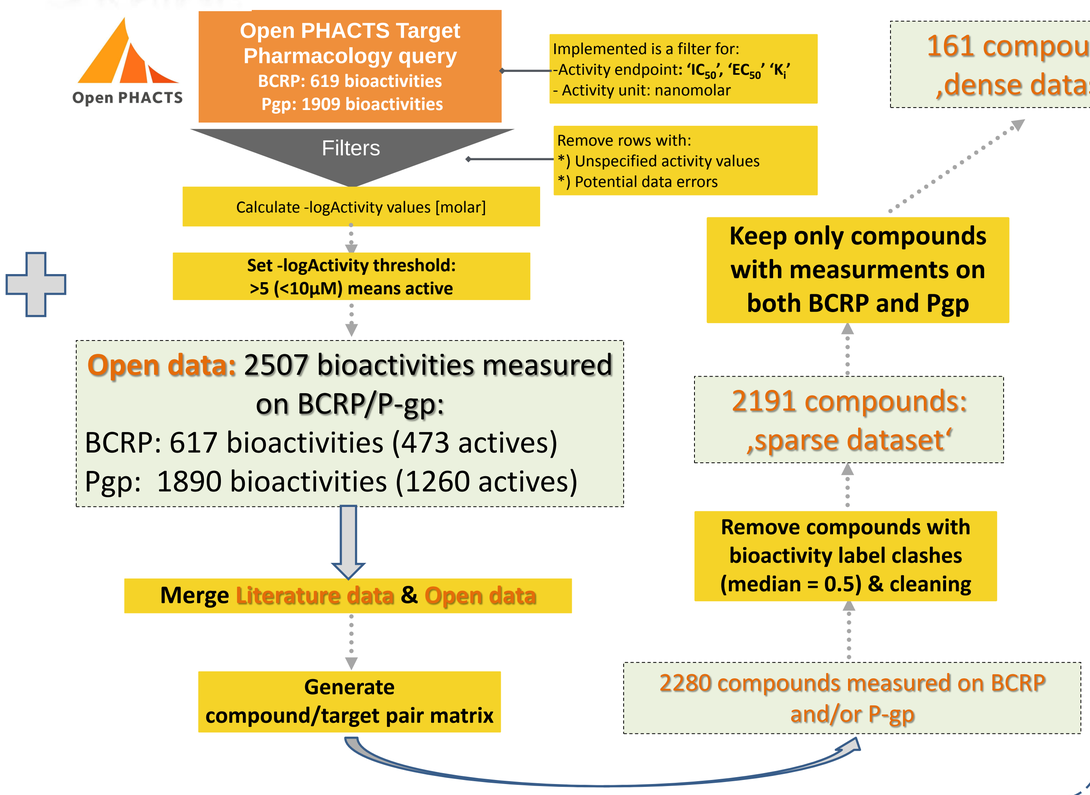
The human ATP binding cassette (ABC) transporters **Breast Cancer Resistance Protein (BCRP)** and **Multidrug Resistance Protein 1 (P-gp)** are **co-expressed in many tissues and barriers**, especially at the blood-brain barrier and at the hepatocyte canalicular membrane. Understanding their interplay in **affecting the pharmacokinetics of drugs** is of prime interest. ***In silico* tools to predict inhibition and substrate profiles towards BCRP and P-gp might serve as early filters** in the drug discovery and development process. However, to build such models, pharmacological data must be collected for both targets, which is a tedious task, often involving manual and poorly reproducible steps.

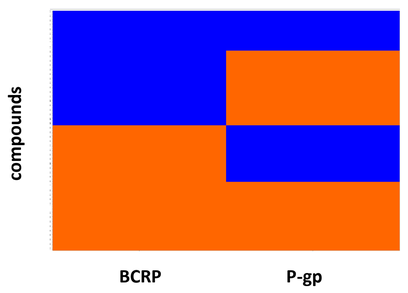
Literature data: 978 unique compounds measured for their BCRP inhibitory activity extracted from 47 sources [1]

Open data sources are explored by using Open PHACTS API calls [3].

Fully flexible KNIME workflow: can be adapted to any protein target(s).

KNIME -workflow for data collection [2]





Heatmap respresentation of bioactivities measured against BCRP/Pgp in the dense dataset: orange bars…..active; blue bars…..inactive

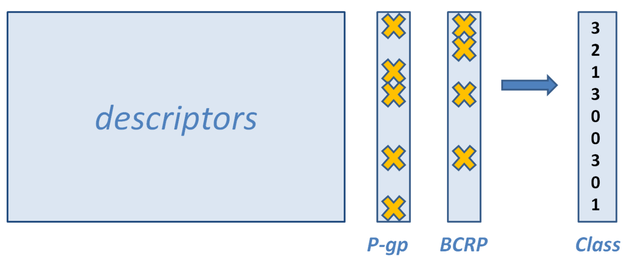
Approach:

1. Determination of **compound overlap.**
2. Use of machine learning approaches in order to establish **multi-label classification models** for P-gp/BCRP.
3. Different ways of addressing multi-label problems are explored and compared:

   Label powerset, binary relevance and classifier chains.

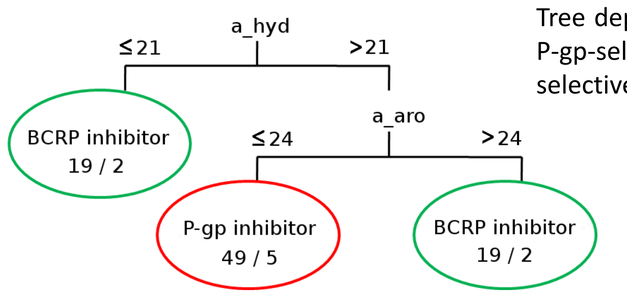
Multi-class classification for P-gp/BCRP profiling

1. The label powerset transformation (dense dataset)



Class	Compound number	Description
0	27	Non-inhibitors
1	48	P-gp selective inhibitors
2	39	BCRP-selective inhibitors
3	47	Dual inhibitors

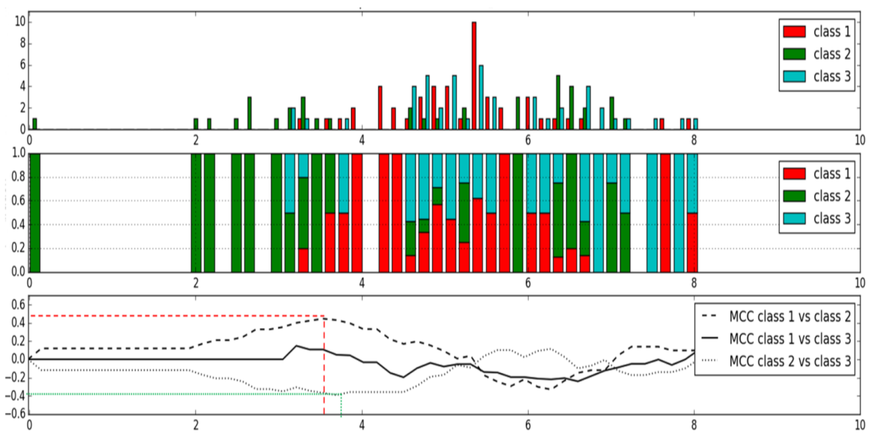
1B. Two-class classification: exploring selectivity



Tree depiction of the **JRip model** to separate P-gp-selective inhibitors (red leaf) from BCRP-selective inhibitors (green leaves).

Only **two descriptors** were sufficient to **separate selective BCRP inhibitors from selective P-gp inhibitors.**

1A. Three-class classification: distinguishing between different types of transport inhibitors



Bagging of J48 / 71 interpretable MOE descriptors

→ **Most important descriptors:**

- *SlogP*
- *the number of hydrophobic atoms*
- *the number of aromatic atoms*
- *the number of donor and acceptor atoms*
- *the Wiener path (sum of the lengths of the shortest paths between all pairs of heavy atoms)*

Distribution of SlogP among the three kinds of inhibitors Top panel: barplot of the counts per binned value of SlogP. Middle panel: proportions of each class in each bin, by putting each bin count to 100%. Lower panel: Matthews Correlation Coefficient (MCC) that would be obtained by splitting the data at each SlogP value. MCC values that peak above or below 0 show ideal thresholds to separate the data between classes. The colored dotted lines corresponds to the peaks of MCC and the corresponding SlogP values (between 3 and 4) for separating class 1 from 2 (red dotted lines) and class 2 from 3 (green dotted lines).

2. Binary relevance and Classifier chains: exploiting all the data (sparse dataset)

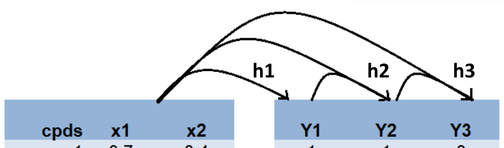
cpds	x1	x2
1	0.7	0.4
2	0.6	0.2
3	0.1	0.9
4	0.3	0.1

h1: X → Y1
h2: X → Y2
h3: X → Y3

Y1	Y2	Y3
1	1	0
1	1	0
0	0	1
0	0	0

Binary relevance: building independent models for each label and using them together for the final prediction; ☹: labels are treated as independent variables

Algorithms	Macro-accuracy	Macro-MCC	Macro-AUC
Binary relevance, Logistic Regression	0.812	0.594	0.793
Classifiers chain, Logisitic Regression	0.812	0.594	0.793
Binary relevance, RandomForest	0.835	0.641	0.808
Classifiers chain, RandomForest	**0.836**	**0.643**	**0.809**
Binary relevance, SVM	0.766	0.504	0.749
Classifiers chain, SVM	0.767	0.504	0.750



cpds	x1	x2
1	0.7	0.4
2	0.6	0.2
3	0.1	0.9
4	0.3	0.1

Y1	Y2	Y3
1	1	0
1	1	0
0	0	1
0	0	0

h1: X → Y1
h2: X+Y1 → Y2
h3: X+Y1+Y2 → Y3

Classifier chains: list of labels is shuffled and a model is trained using the first label and all the data for which there is an annotation for that label → predict this label (as a score between 0 and 1) for all compounds of the dataset (even those for which there was no information for that label) → prediction is appended to the features matrix and serves as additional descriptor for training the next model, on the second label and so on….

Conclusions:

- The workflow proved a useful tool to **merge data from diverse sources**. It could be used for building multi-label datasets of any set of pharmacological targets where there is data available in the open domain *or in-house*.
- **Label-powerset revealed important molecular features for selective or polyspecific inhibitory activity.**
- By using the sparse dataset with **missing annotations**, predictive models can be derived in cases where **no accurate dense dataset is available.**

References:

[1] Montanari F. et al., Mol. Inform. 2014, 33, 322.
[2] http://www.knime.org
[3] Williams AJ et al., Drug Discov Today, 2012, 17, 1188.

Acknowledgements:

*The research leading to this work has received support from the **Austrian Science Fund** (FWF), Grant **F03502** and from a '**Back to Research Grant**' funded by the Faculty of Life Sciences, University of Vienna. We acknowledge support from the **Innovative Medicines Initiative** Joint Undertaking under grant agreement no. [115191], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.*