Probabilistic Predictions of Metabolism using Venn-Predictors

Staffan Arvidsson Mc Shane*, Lars Carlsson**, Paolo Toccaceli***, Ola Spjuth*

*Department of Pharmaceutical Biosciences, Uppsala University
**Quantitative Biology, Discovery Sciences, Innovative Medicines & Early Development, AstraZeneca
***Department of Computer Science, Royal Holloway

Contact:
Staffan Arvidsson Mc Shane
staffan.arvidsson@farmbio.uu.se

ABSTRACT

Prediction of drug metabolism is an important topic in the drug discovery process, and we here present a study using probabilistic predictions applying Cross Venn-ABERS Predictors (CVAPs) on data for site-of-metabolism. We used a dataset of 73599 biotransformations, applied SMIRKS to define biotransformations of interest and constructed five datasets where chemical structures were represented using signatures descriptors. The results show that CVAP produces well-calibrated predictions for all datasets with good predictive capability, making CVAP an interesting method for further exploration in drug discovery applications.

DATASET GENERATION

The datasets were generated from the MDL Metabolite database (2005) and processed according to:

- MCS matching for all reactions
- SMIRKS filtration for 60 reaction types
  - Yellow → only match in substrate (class label 0)
  - Red → match in full reaction (class label 1)
- Filtration to keep one record per substrate, in favor for full reaction, for each of the 60 datasets

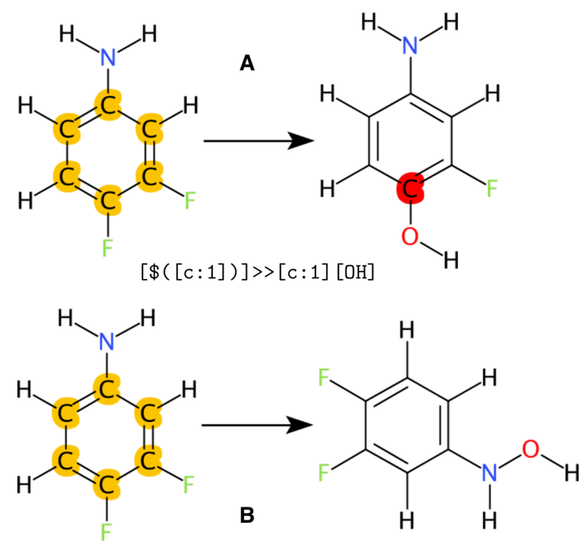
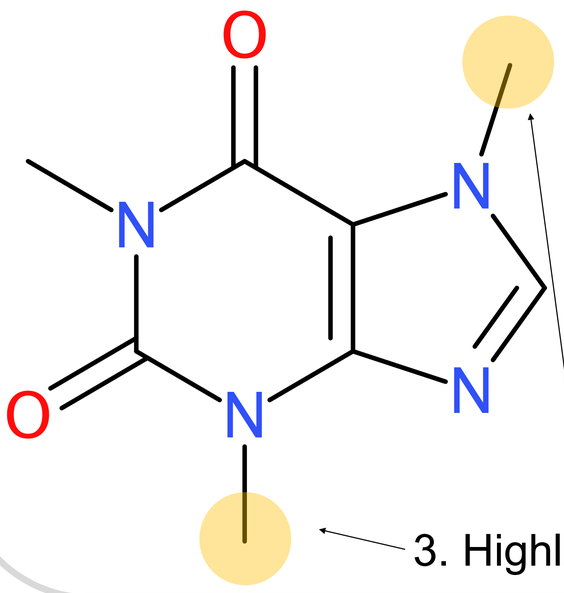


Table 1: Datasets

Dataset	Reactions	Class distribution
Alkyl hydroxylation	17793	68 : 32
Aromatic hydroxylation	14691	85 : 15
Carboxylation	12580	64 : 36
Oxidation of tertiary amine	11040	97 : 3
Aromatization	9518	25 : 75

RESEACH AIM

1. Input a drug candidate



2. List all potential reactions after probability of reaction to occur:

88.5% Reaction type 1
75.2% Reaction type 2
:
0.0% Reaction type N

3. Highlight Reaction Centers

Cross Venn-ABERS algorithm

Normal classification setting, using observations $z = (y, x)$, with labels $y \in \{0, 1\}$ for objects x.

Training procedure

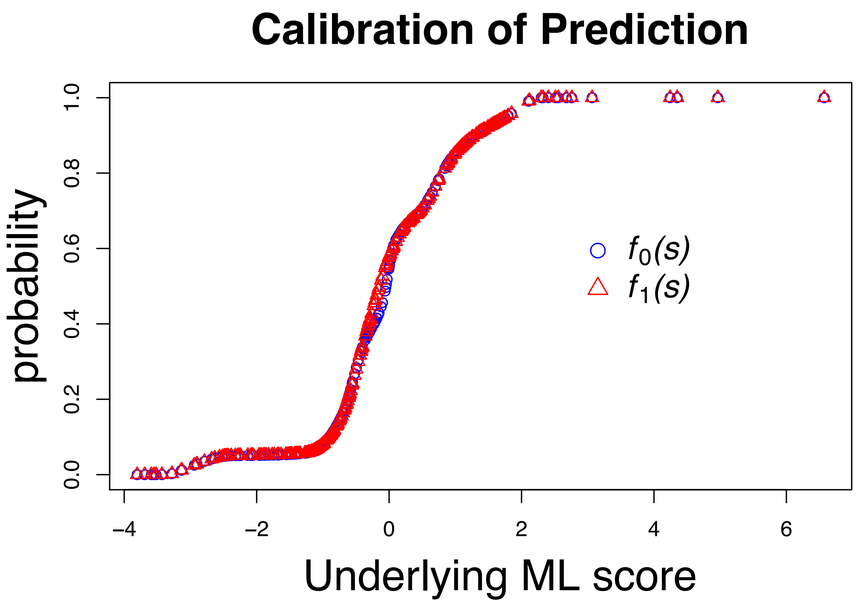
Split dataset in k non-overlapping folds, for iteration $i = 1,\ldots,k$ do:

1. Set fold i as *calibration set*, remaining observations is the *proper training set*.
2. Train underlying scoring model on *proper training set*.
3. Predict all observations in *calibration set* using the scoring model (step 2).
4. Save scoring model together with calibration points consisting of tuples of (*score*, *y*).
5. Redo step 1-4 for remaining k-1 folds, each leaving one fold for calibration.

Predicting procedure

1. For a new test object x_{new}, for each of the k Venn-Predictors, do:
2. Predict $score_{new}$ using the scoring model, fit two isotonic regression functions $f_0(s)$ and $f_1(s)$ using (Figure below);
   - Append ($score_{new}$, 0) to calibration points → fit $f_0(s)$
   - Append ($score_{new}$, 1) to calibration points → fit $f_1(s)$
3. Set $p_0 = f_0(score_{new})$ and $p_1 = f_1(score_{new})$, creating a probability *interval* $[p_0, p_1]$.
4. Aggregate the k intervals using formula (optimal under log-loss):

$$p = \frac{GM(\overrightarrow{p_0})}{GM(1 - \overrightarrow{p_0}) + GM(\overrightarrow{p_1})}$$



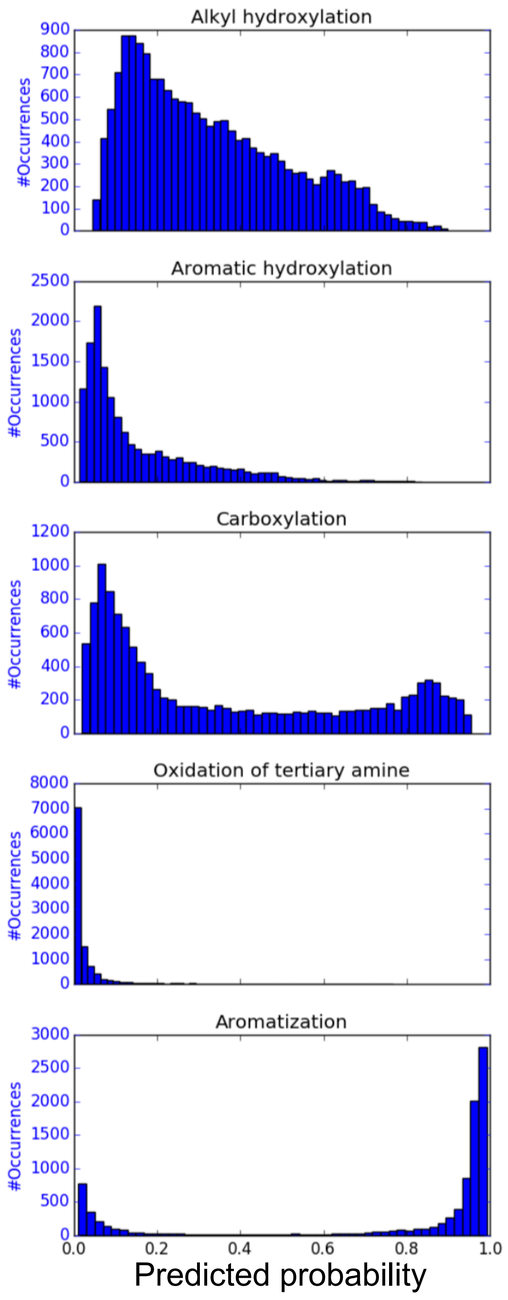
RESULTS

Pre-study containing 5 reaction types showed promising results with OK to good AUC scores (Table 2). Full paper available at http://proceedings.mlr.press/v60/

Further study comprising more reactions under way, including publication of predictive models as microservices part of the OpenRiskNet.

Table 2: Performance metrics for QSAR

Dataset	Log Loss	AUC
Alkyl hydroxylation	0.538	0.753
Aromatic hydroxylation	0.348	0.793
Carboxylation	0.41	0.881
Oxidation of tertiary amine	0.093	0.904
Aromatization	0.173	0.964



OpenRiskNet

A 3 year project to build an open e-Infrastructure to support Data Sharing, Knowledge Integration and in silico Analysis in Predictive Toxicology and Risk Assessment. Currently focusing on annotated OpenAPI specification for microservices deployed on OpenShift.

Read more at https://openrisknet.org/

OpenRiskNet (Grant Agreement 731075) is a project funded by the European Commission within Horizon2020 Programme