Machine Learning Strategies for Assay Transitions

Staffan Arvidsson McShane[✢,*], Ernst Ahlberg[✢,✾], Tobias Noeske[✻], and Ola Spjuth[✢]

✢Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University
✾Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca
✻Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, AstraZeneca

* staffan.arvidsson@farmbio.uu.se

Conclusions

Data acquired from legacy assays can be used in order to improve the efficency of predictive models. Evaluation through conformal prediction (CP) allows the user to monitor both efficiency and validity (calibration) of the produced models. Modeling strategies were the calibration set of the CP was exchangeable with test data (Table 1) always produced valid models. Contrary, the naïve approach of pooling new and legacy data produced valid models only in a few senarious and exclusively for the task of classification.

We propose that CP should be used, and to continously monitor both efficiency and calibration as new data is generated, before choosing the modeling strategy to use. Furthermore, the CCP_{AT} strategy was the overall most successful modeling strategy (exemplified in Figure 1). The full work was published in [1].

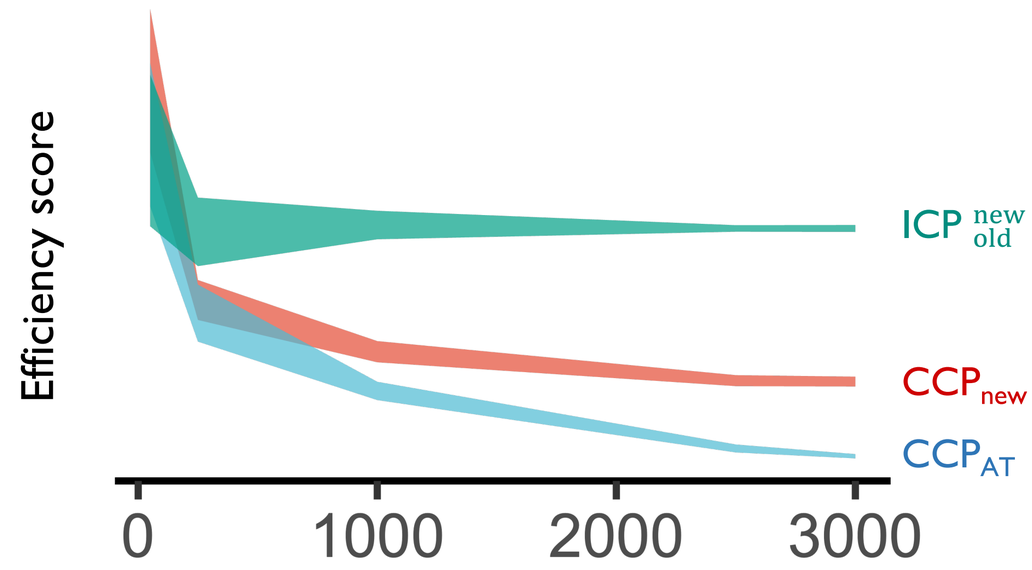

Figure 1. Efficiency plot for hERG regression data set using all available old data, colored areas correspond to the 95% CI from 10 replicate runs. The efficiency score (prediction interval width) should be minimized.

Table 1. Modeling strategies. Exchangeable denotes exchangeability vs the test set

Strategy	Calibration set	Exchangeable
CCP_{new}	new	x
CCP_{pool}	old U new	
ICP^{new}_{old}	new	x
CCP_{AT}	new	x

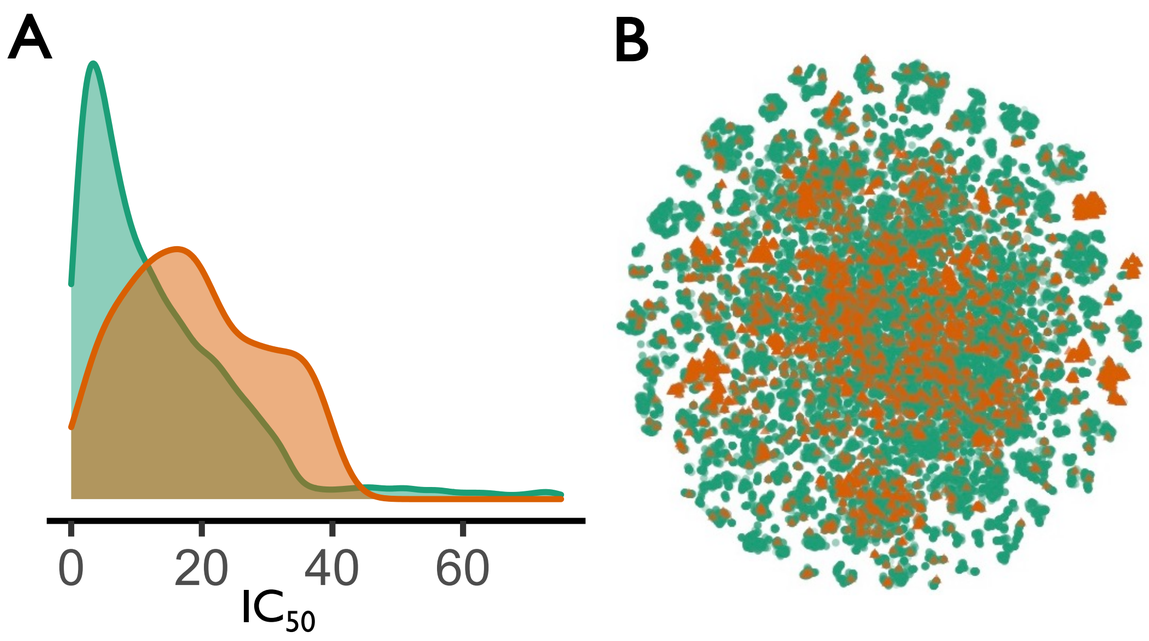

Figure 2. **A.** Distribution of label-space of the Na_V data set for the old vs new data sets. **B.** 2 dimensional t-SNE plot of descriptor space for hERG data set, using the same color scheme as in panel A.

Introduction

In this work we studied strategies of how to combine data from a *legacy* assay system (old) with data acquired from a new assay system (new). Combining data from different sources is problematic in ML as it violates the IID assumption, despite the multi-million $ that could have been invested in old assays. However, the performance of predictive models depend on the amount of available training data, making it desirable to include old data in order to build optimal models.

Data

Data was supplied from AstraZeneca for the hERG and Na_V endpoints, encompassing approx. 64k and 5k records from old assays and 5k and 200 from the new assay, respectively (see [1] for actual numbers). Old and new assays diverged both in descriptor space and label space (Figure 2). Four different data sets were studied.

Methods

Conformal Prediction (CP)[2] models with custom sampling strategies (Table 1) were used in order to construct models that keep exchangeability between the calibration set of the CP models and test data. Furthermore, the naïve approach of pooling all available data (CCP_{pool}) was evaluated. All CP models used SVMs as underlying scorer algorithm.

Study design

To simulate different user senarios, several combinations of data set size from the old and new assays were evaluated. Evaluation was exclusively performed on records from the new data sets (based on calibration and efficiency) using a 10-fold CV. Each combination was evaluated using 10 replicate runs, using different shuffling of records.

Results

Strategies were filtered based on calibration, excluding those with insufficient calibration. Remaining strategies were compared based on efficiency (e.g. Figure 1) were the best strategy varied depending on data set and combination of size of old and new data. The overall best strategy was the CCP_{AT}, which leverage old assay data in training of underlying models, but calibrate the CP model exclusively using new data. Contrary, the approach of simply pooling old and new data produced poorly calibrated models in most cases.

Refrences

[1] Arvidsson McShane, S., et al. (2021). Machine Learning Strategies When Transitioning between Biological Assays. *Journal of Chemical Information and Modeling*.
[2] Vovk, V., et al. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

I am a PhD student in the Spjuth lab, mainly working with machine learning algorithms that can produce measures of confidence in their produced predictions. The main application area has been QSAR modeling.

Uppsala University
Staffan Arvidsson McShane, M.Sc.
Department of Pharmaceutical Biosciences and Science for Life Laboratory
Uppsala Biomedical Centre (BCM), 751 24
Uppsala, Sweden

E-mail: staffan.arvidsson@farmbio.uu.se
Web: pharmb.io