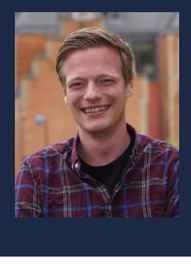
Generalized Calibration Across Liquid Chromatography Setups for Generic Prediction of Small-Molecule Retention Times

***Robbin Bouwmeester*[1,2,3], *Lennart Martens*[1,2], *Sven Degroeve*[1,2]**

1 VIB-UGent Center for Medical Biotechnology, Ghent, Belgium
2 Department of Biochemistry, Ghent University, Ghent, Belgium
3 Janssen Pharmaceutica N.V., Beerse, Belgium



Robbin.Bouwmeester@ugent.be

Retention time prediction in LC-MS^2

Accurate LC retention time (Rt) prediction of analytes is useful for better identification rates in untargeted MS and limiting experimental measurements in targeted MS. Rt prediction has been applied in untargeted MS to differentiate between isobaric lipids [1], limit the spectrum match search space [2] and score spectrum matches [3]. However, these predictors are not universally applied in the data analysis due to differences in experimental setups. Different experimental setups (columns, solvents, gradients, stationary phase, etc.) give rise to a multitude of prediction models that only predict accurate retention times for a specific experimental setup (Fig. 1).

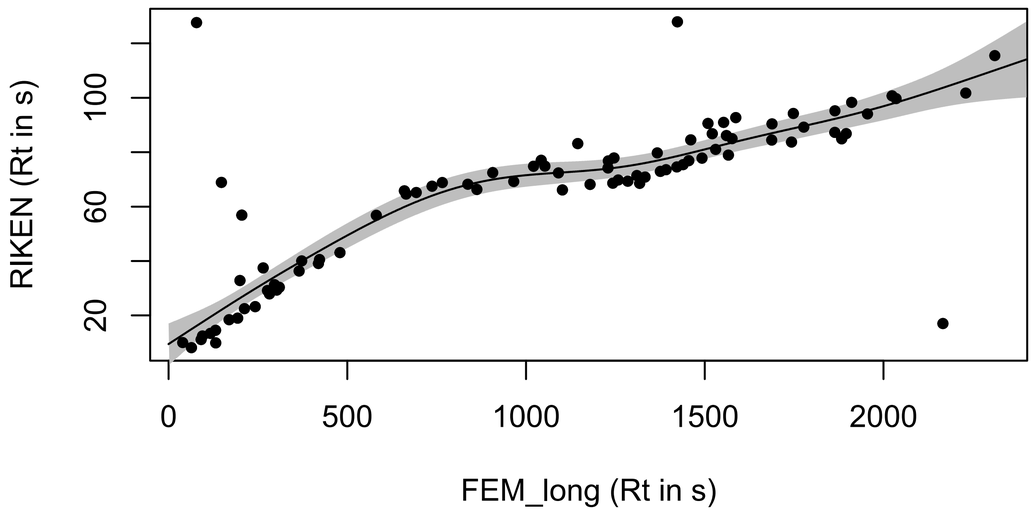


Figure 1: Retention times of the same analytes in two different setups. Figure adapted from [4].

Calibration between setups adds the necessary flexibility

The key idea presented here is to use calibration curves to overcome the problem of different experimental setups (Fig. 2) [5]. First, regression models are fitted to specific experimental setups and these models are used to make predictions for analytes from the setup of interest (Layer 1). Predictions from different setups and experimental measurements from the setup of interest are used to fit a calibration curve (Layer 2). Finally, the calibrated predictions are combined using a weighted average (Layer 3).

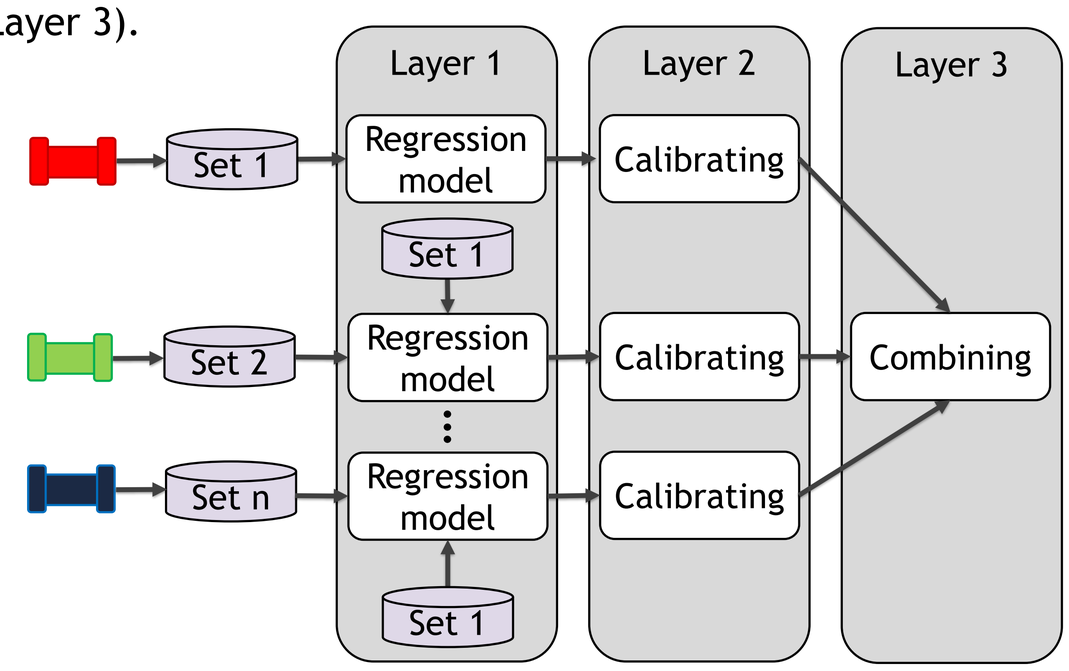


Figure 2: Concept for a retention time predictor that uses calibration. The layers illustrate the essential steps taken in the predictor.

Training the retention time predictor using calibrations

Data from 36 different experimental setups were used to fit the Rt predictor. A total of 151 features were extracted for 8305 metabolites and used to train five distinct machine learning models per experimental setup. The predictions from the setup specific models are calibrated using a generalized additive model (Fig. 3). Finally, a LASSO algorithm is used to train a model that combines the calibrated predictions from different setups and machine learning models.

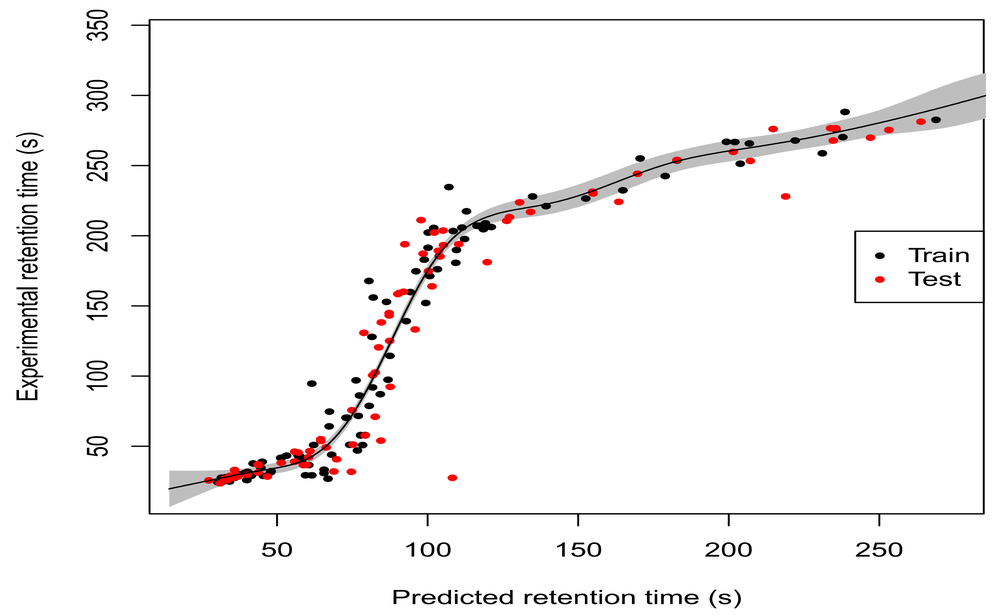


Figure 3: Calibration curve fitted using experimental data from the system of interest and predictions from models trained on different setups

Higher performance with calibration

Performance of the calibration method is evaluated using an internal comparison between the layers (Fig. 4a) and with a recently published predictor [1] (Fig. 4b). The internal evaluation shows that Layer 3 is able to outperform the other layers with a 3% decrease in error relative to the total elution time. The difference between layers becomes smaller when the number of initial training instances increases. The evaluation between the recently published predictor shows an improvement in the mean absolute error of more than 25% for the calibration method.

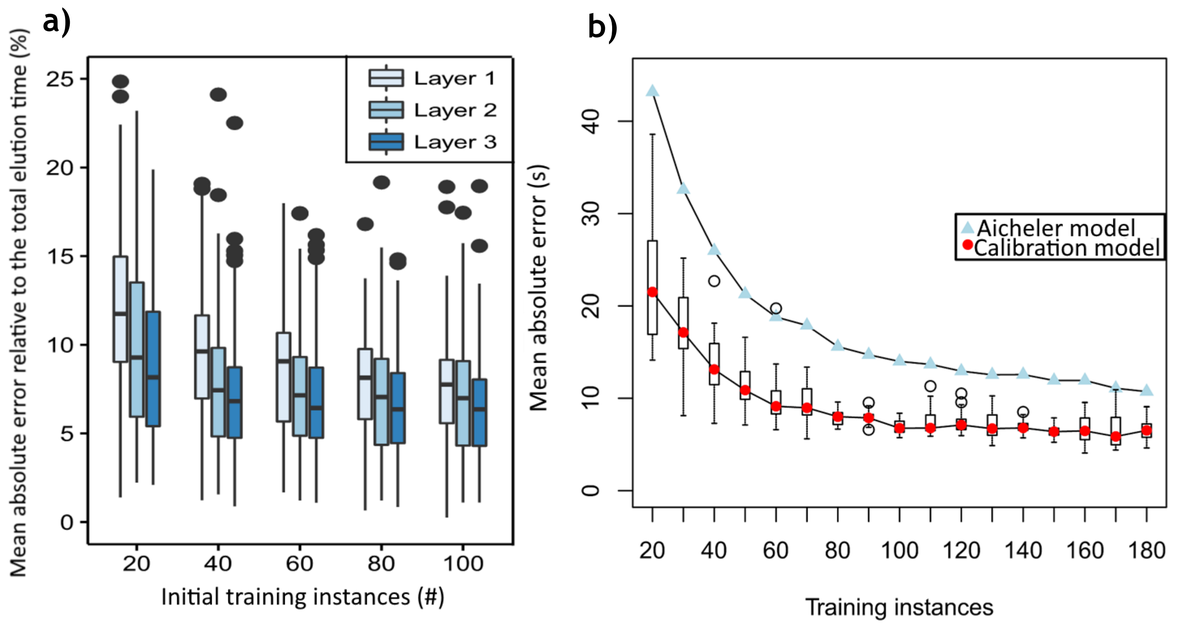


Figure 4: Internal (A) and external (B) performance evaluation using cross-validation for the retention time predictor with calibration

Currently, retention time (Rt) predictors are specific to experimental setups. Researchers are unable to use the Rt predictions in their experiments partly due to this inflexibility. Calibration between the setup specific regression models adds the necessary flexibility and gives more accurate predictions. The model for retention time prediction using calibration allows for *in silico* experimental design and better identification of analytes by comparing experimental to predicted retention times.

References
[1] Aicheler, Fabian, et al. "Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches." *Analytical chemistry* 87.15 (2015): 7698-7704.
[2] Spicer, Vic, et al. "Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversed-phase HPLC." *Analytical chemistry* 79.22 (2007): 8762-8768.
[3] Klammer, Aaron A., et al. "Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions." *Analytical Chemistry* 79.16 (2007): 6111-6118.
[4] Stanstrup, Jan, Steffen Neumann, and Urska Vrhovsek. "PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems." *Analytical chemistry* 87.18 (2015): 9421-9428.
[5] Bouwmeester, Robbin, Lennart Martens, and Sven Degroeve. "Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction." *Analytical chemistry* 91.5 (2019): 3694-3703.

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY
GHENT UNIVERSITY
CompOmics
janssen PHARMACEUTICAL COMPANIES OF Johnson & Johnson
SCIENCE MEETS LIFE