EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BIO-4372
CHEMINFORMATICS
Winter Semester 2024-25

Dr. Philipp Thiel
Institute for Bioinformatics
and Medical Informatics

QSPR Modeling Competition

Submission deadline: January 24, 2025, 6 p.m.

In this project, you are expected to construct a predictive model based on a given QSPR data set where the compounds are annotated with effectivity end-point measurements for *Pimephales promelas*. Please evaluate suitable approaches in order to generate a predictive model to estimate compound toxicity. All teams will receive the same data set and the task is to finally come up with the best predictor you have constructed. The performance will be evaluated on an undisclosed external test data set and the team handing in the best predictor will be awarded. Yes, it is a contest!

The training data set is provided in the file `chin-qspr-dataset.sdf`. It contains 375 compounds in SD format (SDF) annotated with

$$pLC_{50} = log(\frac{1}{LC_{50}})$$

values, where LC_{50} is the lethal concetration 50% for *P. promelas*. The undisclosed test data set will have the same format. Thus, make sure that the predictor you will hand in reads exactly that format without any modifications. As an output, your predictor shall write a CSV file that contains two columns: (1) 'compound_id' and (2) 'pred_pLC50'. You can choose between Python or KNIME to work out the project. The summary statistics to assess prediction quality of the final model will be the root-mean-square error (RMSE) and, if required, R^2 as a second criterion.

Deliverables

Please write a report with maximally four pages that describes your QSPR project as precisely as required. As usual, please use the report template and this time the structure of a scientific publication. Together with your report, please hand-in your final predictor. Please make sure that your script or your workflow does not retrain the model. As usual, add sufficient documentation to run and understand your predictor.

Team

Please form teams of size two and let me know if there are problems forming teams! It is sufficient if only one team member hands in the deliverables.

Presentation

Each presentation has a slot of 7 minutes (5+2). Every team member is required to present a significant part. The presentations take place in course of the last lecture.

Grading

The project is graded with 40 points (two assignment sheets). Participating in and working out the project is required for exam admission (c.f. introductory slide deck).

▷ Please use Slack to discuss problems in the first place
▷ If you have confidential questions, don't hesitate to drop by or write an e-mail