# Support vector machines for compound activity and potency prediction

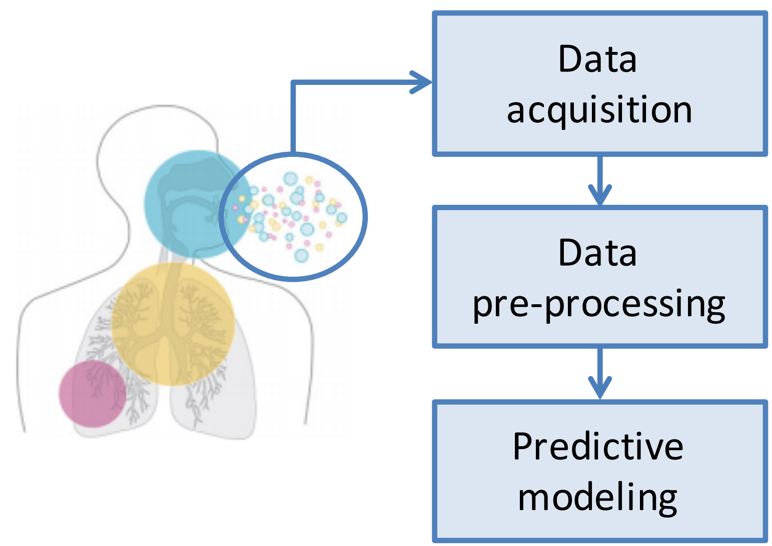
Raquel Rodríguez Pérez
B-IT Life Science Informatics
Rheinische Friedrich-Wilhelms-University Bonn

*BigChem School, October 2017*

# Background

- **Biomedical engineering** (B.Sc. and M.Sc.)


- Research at **the Institute for Bioengineering of Catalonia**
  - Data analysis for biomarker discovery in exhaled breath



# BIGCHEM Project: ESR1

- **Machine learning methodologies for mining large compound data sets**
  - Explore different methods to build models with large data sets
  - Develop machine learning strategies to predict compounds with desired multi-target activity profiles

| | |
|---|---|
| UNIVERSITÄT BONN | Sept. 2016 – Mar. 2018 |
| Boehringer Ingelheim | Mar. 2018 – Sept. 2019 |
| Chemotargets | Secondment (2018/2019) |

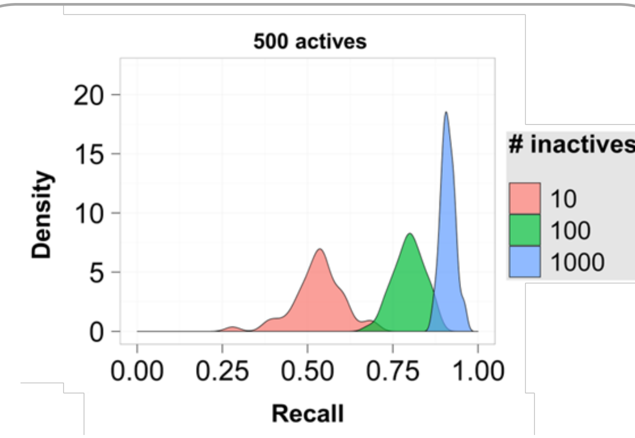
# Overview of the 1st year: Training

- BigChem schools and online courses
- Chemistry course (2 weeks; September 2016)
- German course A1.1 (Winter semester 2016)
- German course A2.1 (Summer semester 2017)
- Teaching in the Programming lab of Life Sciences Informatics Master (2 sessions; Summer semester 2017)
- Chemogenomics workshop mainly given by Dr. J.B. Brown (attendance and a talk; August 2017)

# Overview of the 1st year: Research

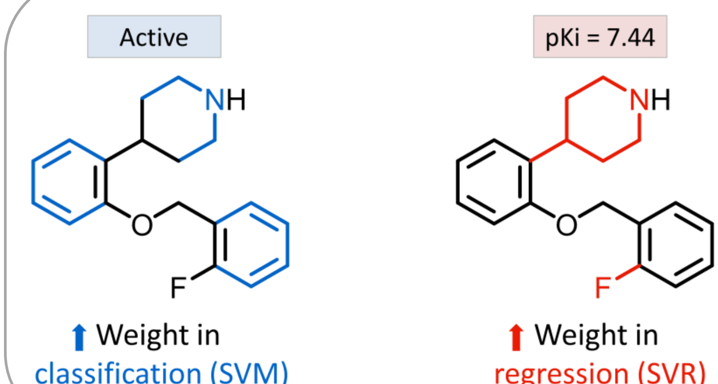
- Application of **support vector machines** classification (SVM) and regression (SVR)
  - Study 1: Influence of **training set composition and size** on SVM activity predictions

    Rodríguez-Pérez *et al. J. Chem. Inf. Model.* **2017**, *57,* 710-716.



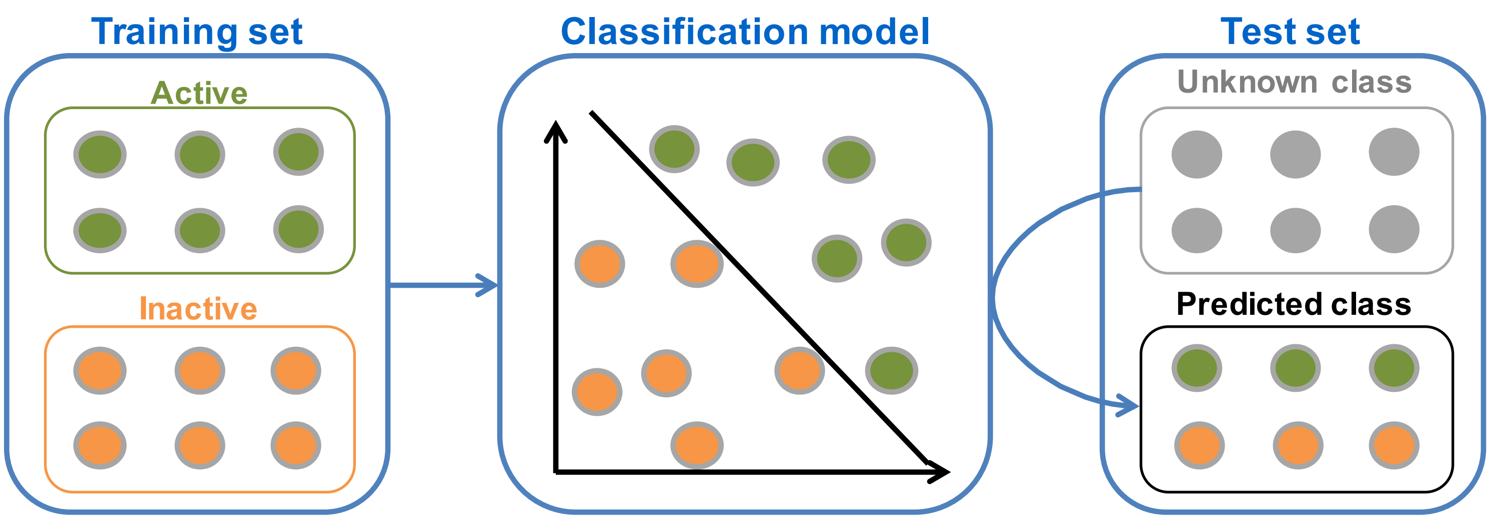

  - Study 2: Prioritized **structural features** for compound activity and potency predictions

    Rodríguez-Pérez *et al. ACS Omega.* **2017**, *2,* 6371-6379.



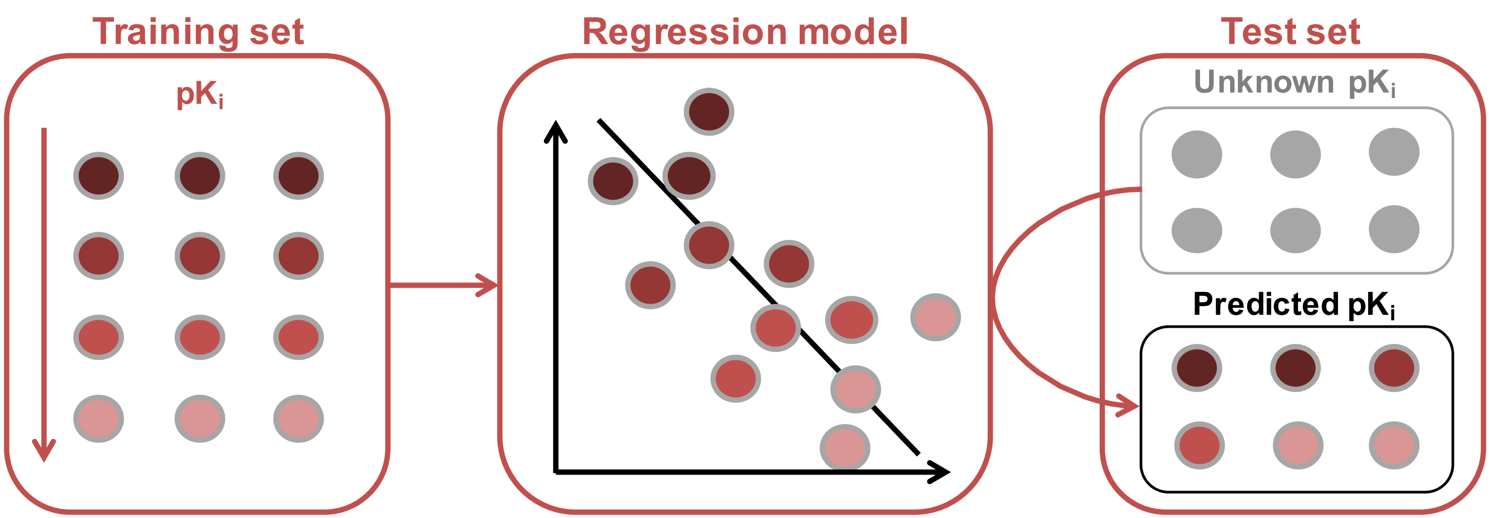
# Machine learning: SVM

- Derivation of computational models for the **prediction** of compound properties
  - Classification (SVM) → Binary activity (active/inactive)



# Machine learning: SVR

- Derivation of computational models for the **prediction** of compound properties
  - Classification (SVM) → Binary activity (active/inactive)
  - Regression (SVR) → Potency value ($pK_i$)

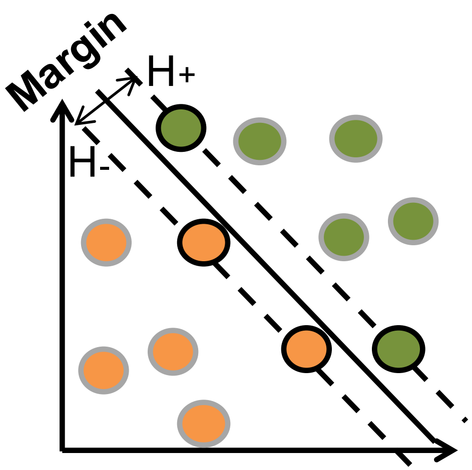


# Machine learning: SVM vs. SVR

## Classification model

Training data: feature vector $\boldsymbol{x} \in X$ and a categorical label $y \in \{-1,1\}$

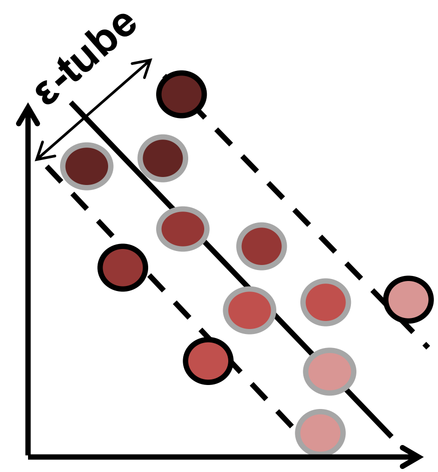
Derivation of: $H: < \boldsymbol{w}, \boldsymbol{x} > + b = 0$




## Regression model

Training data: feature vector $\boldsymbol{x} \in X$ and a numerical label $y \in \mathbf{R}$

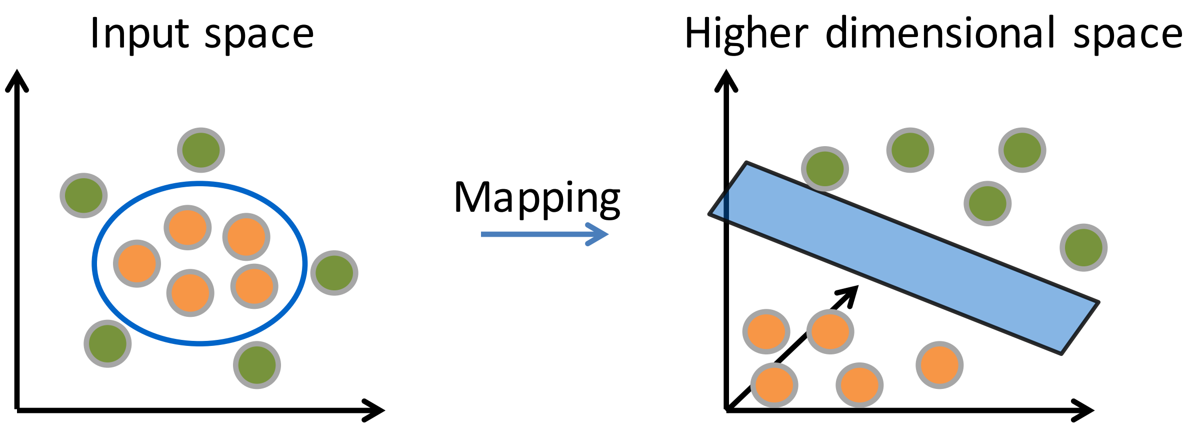
Derivation of: $f(\boldsymbol{x}) = < \boldsymbol{w}, \boldsymbol{x} > + b$




Support vectors
Training data

# Motivation: Model interpretation

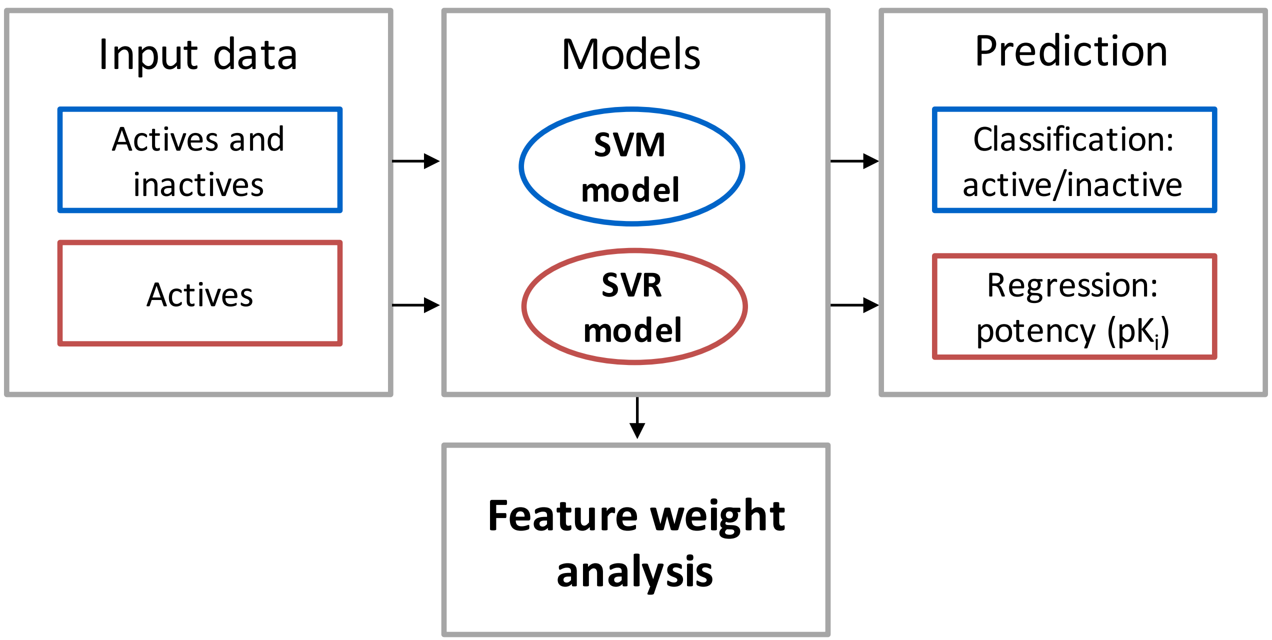
- Kernel trick: mapping into a high-dimensional space
- **Black box** character of SVM and SVR predictions




Identification of **features that determine classification** (SVM) **and regression** (SVR) model performance

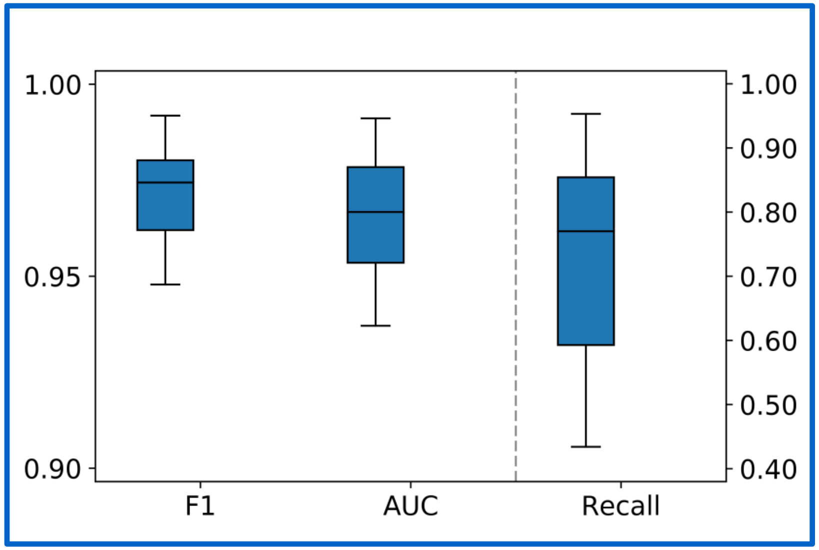
# Methods: Calculation protocol

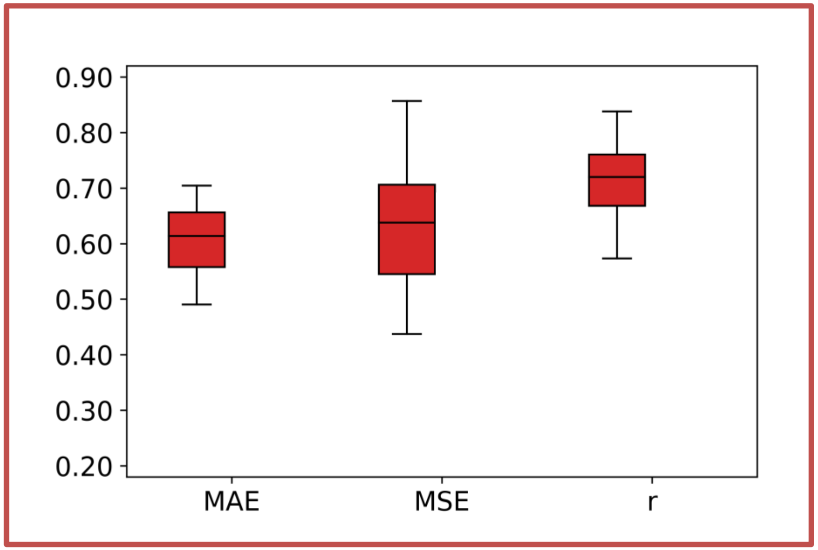
- Data:
  - 15 activity classes from ChEMBL 22
  - For classification inactives from ZINC
- Molecular fingerprints: MACCS and ECFP4



# Results: Global performance

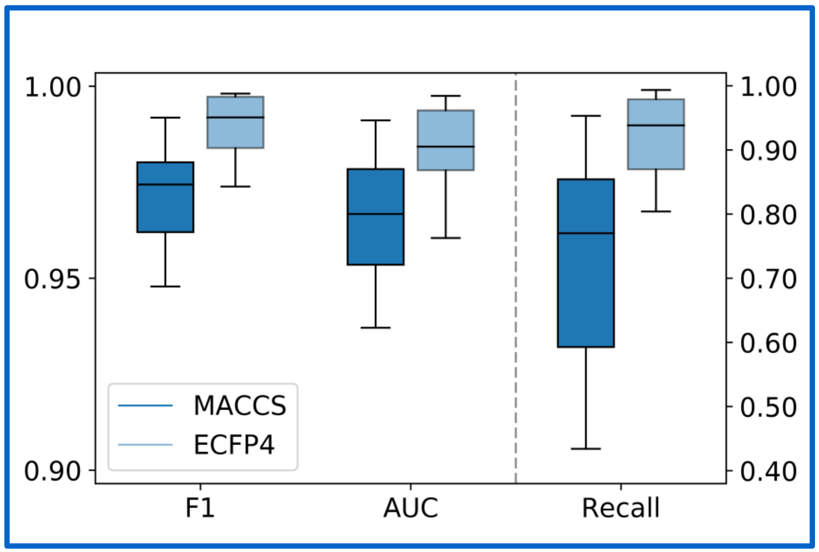
- Accurate classification of active and inactive compounds
- Errors of regression were less than one order of magnitude

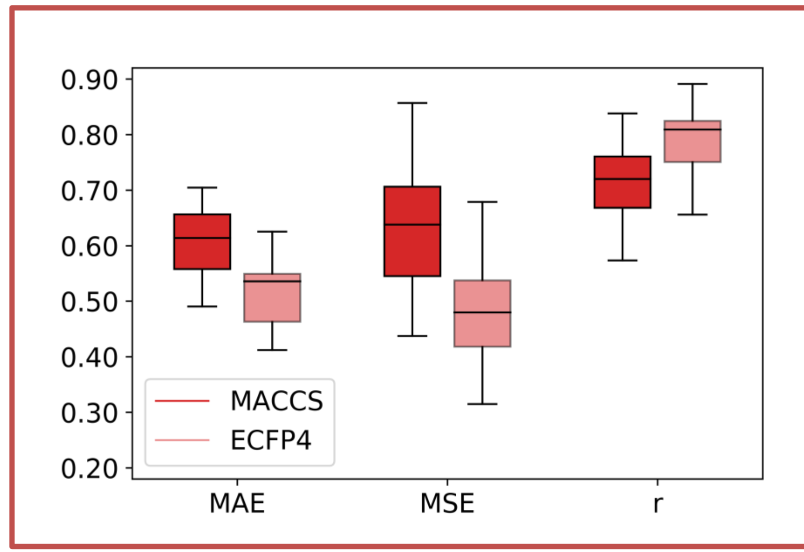


# Results: Global performance

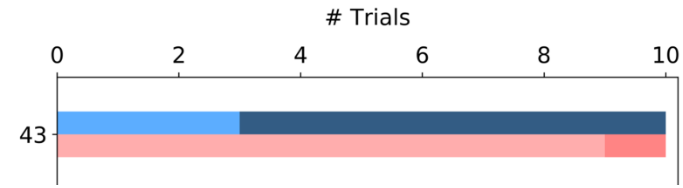
- Accurate classification of active and inactive compounds
- Errors of regression were less than one order of magnitude
- Higher performance of ECFP4 relative to MACCS

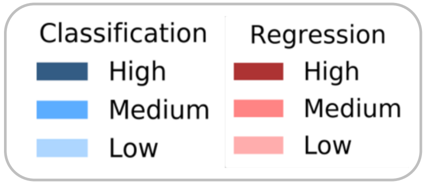


# Feature weight analysis: MACCS

- Some features had consistently high/low weights
- **The importance of many features differed between SVM and SVR**

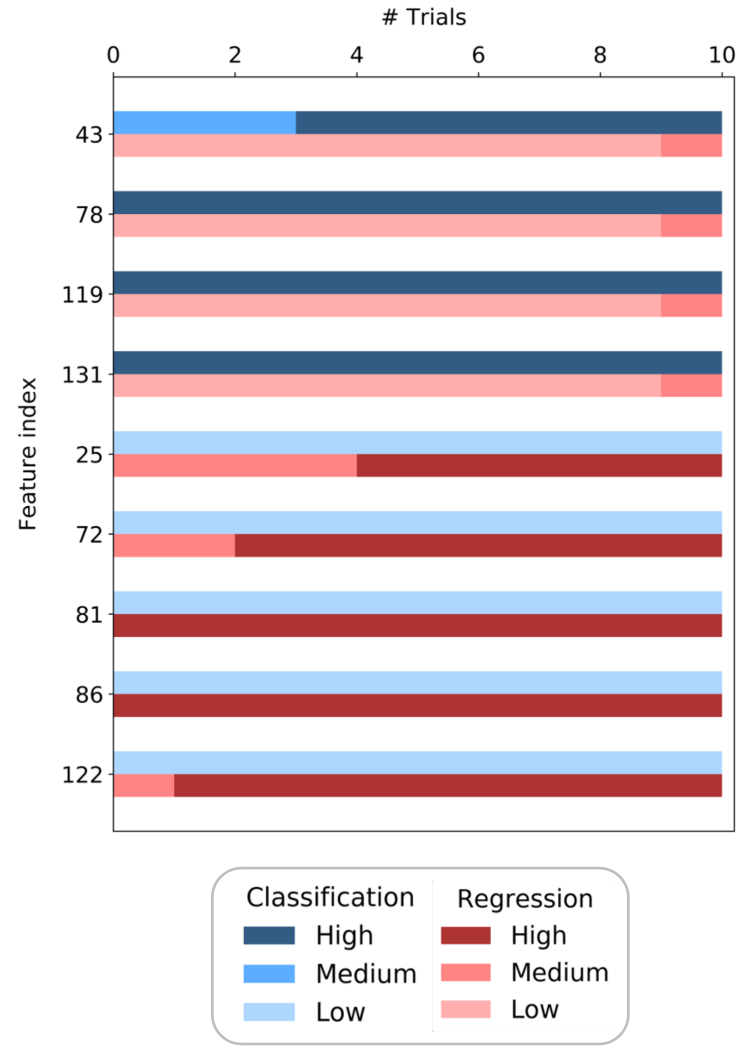



*Thrombin inhibitors*

# Feature weight analysis: MACCS

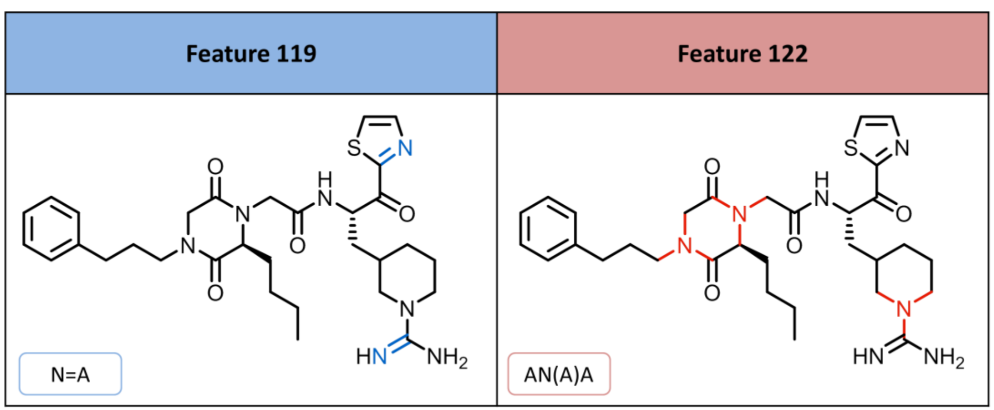
- Some features had consistently high/low weights
- **The importance of many features differed between SVM and SVR**



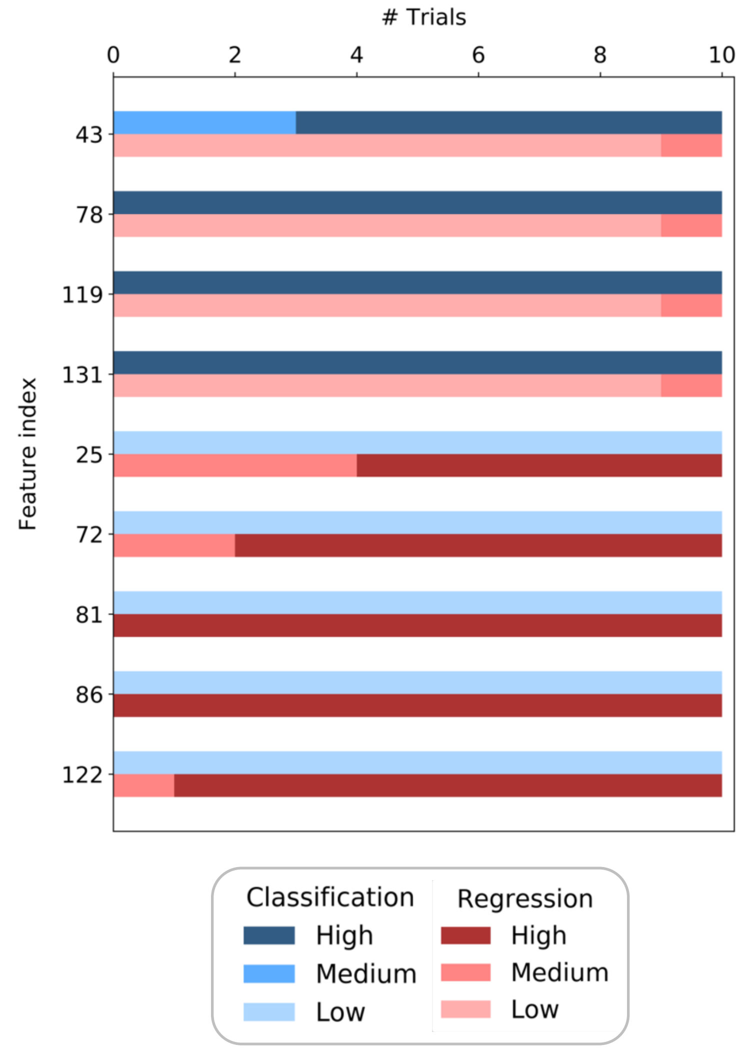

*Thrombin inhibitors*

# Feature weight analysis: MACCS

- Some features had consistently high/low weights
- **The importance of many features differed between SVM and SVR**

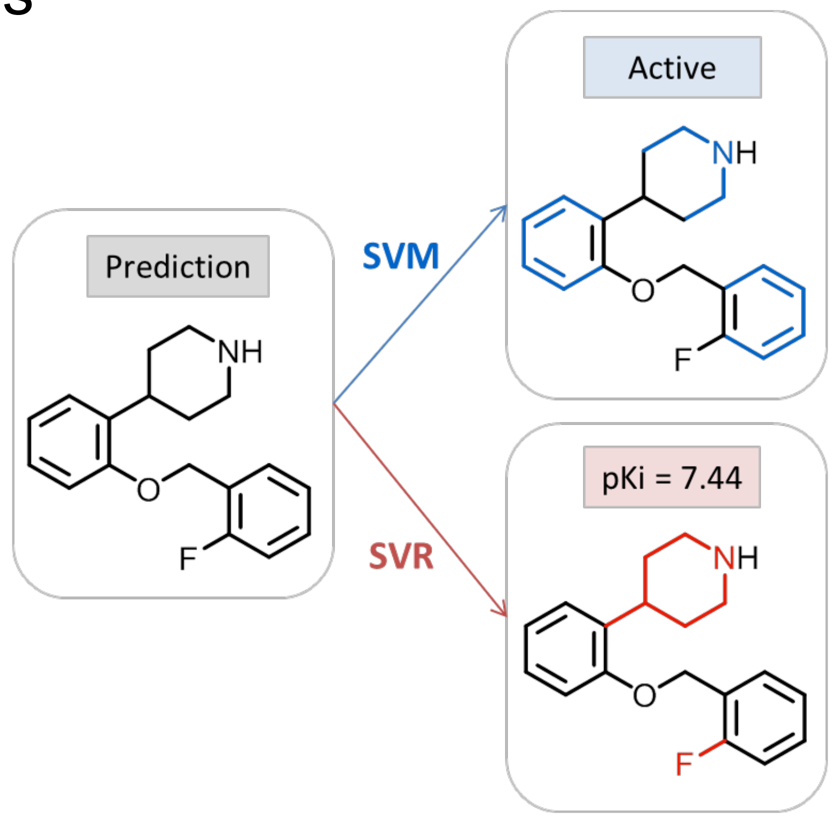



*Thrombin inhibitors*



# Mapping of ECFP4 features

- Highly weighted features were mapped to correctly predicted compounds
- **Different atom environments** (only partly overlapping) are important for activity and potency prediction




⬆ Weight in classification (SVM)

⬆ Weight in regression (SVR)

*Norepinephrine transporter inhibitor*

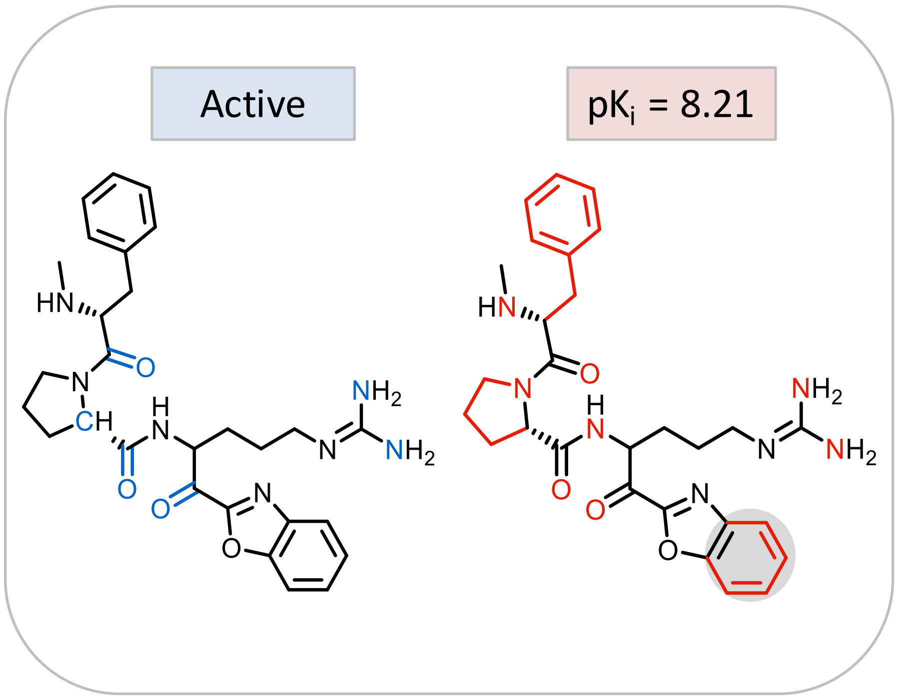
# Mapping of ECFP4 features

- Mapping of highly weighted features may reveal structure-activity relationships (SARs)

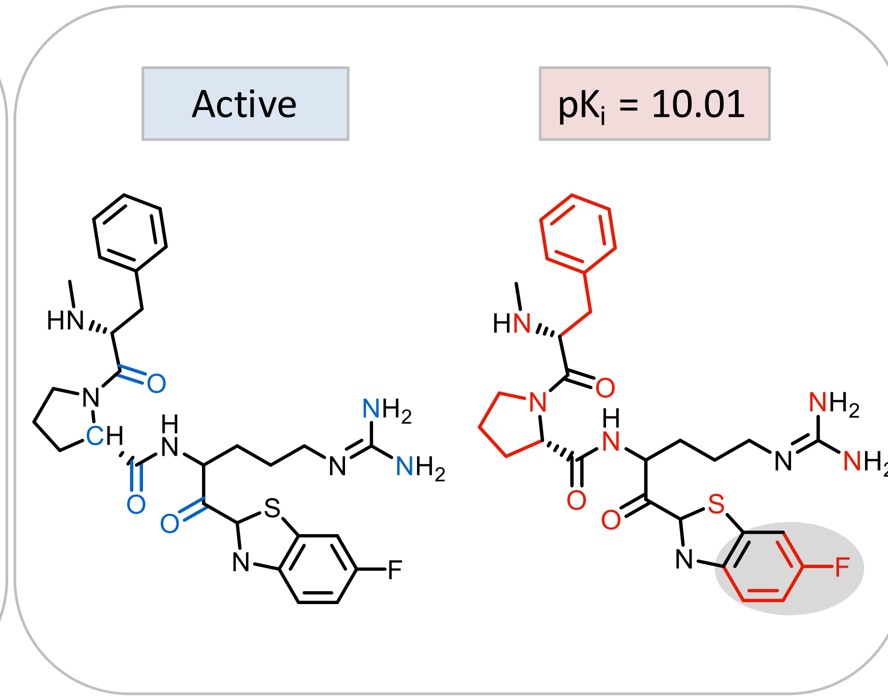
↑ Weight in classification (SVM)

↑ Weight in regression (SVR)

**Thrombin inhibitor A**




**Thrombin inhibitor B**



# Summary

- SVR is an extension of SVM algorithm
- Some **features contribute very differently** to classification and regression
- **Mapping of highly weighted features** helped to:
  - **Model interpretation**
  - Identification of **SAR-informative regions** of compounds

# Acknowledgment

Dr. Martin Vogt

Prof. Dr. Jürgen Bajorath


*E-mail: perez@bit.uni-bonn.de*