Prediction of Metabolic Transformations using Cross Venn-ABERS Predictors

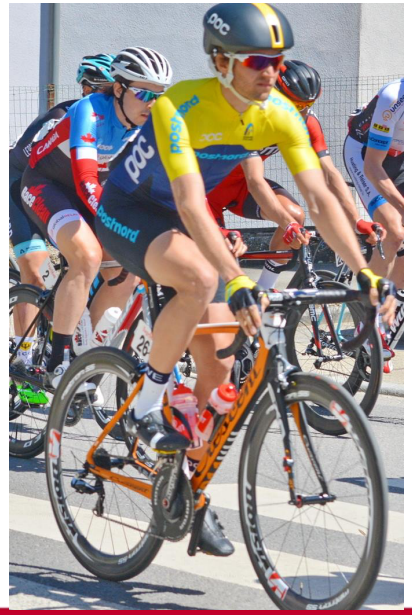
Staffan Arvidsson[†], *Ola Spjuth*[†], *Lars Carlsson*[‡], *Paolo Toccaceli*[*]

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

Contact:
Staffan Arvidsson
staffan.arvidsson@farmbio.uu.se
Pharmaceutical Biosciences
Uppsala University

About me

- PhD student - supervisor Ola Spjuth
- Uppsala University
- Department of Pharmaceutical Biosciences
- Free time: Cyclist



UPPSALA UNIVERSITET

Presentation outline

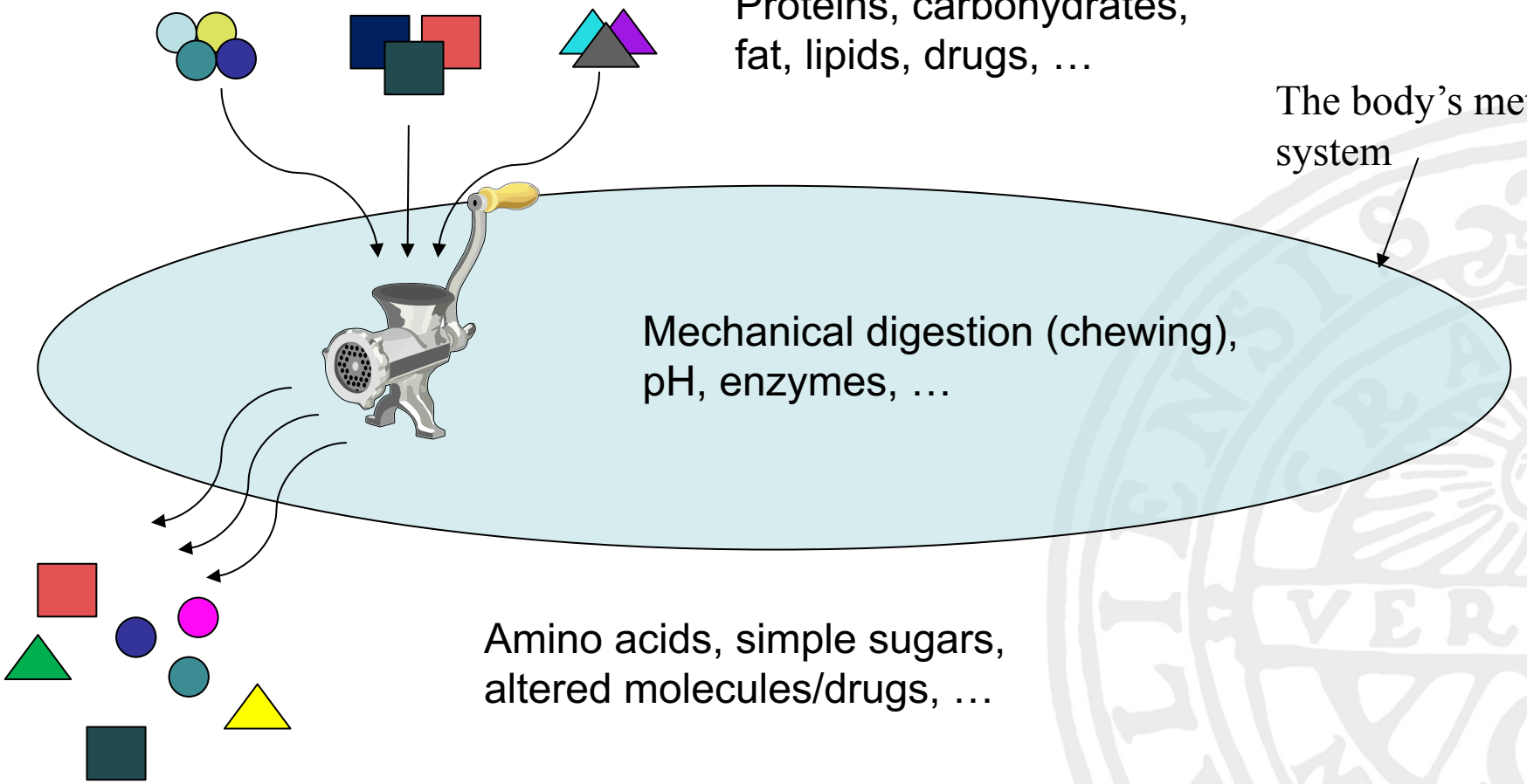
1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

UPPSALA UNIVERSITET

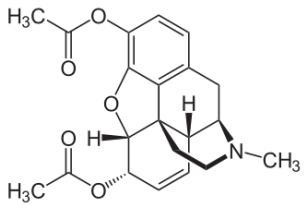
Presentation outline

1. Defining the problem setting
2. Defining SMIRKS as reaction types
3. Preprocessing
4. ML context
5. Results
6. Conclusions

Problem setting – Metabolism

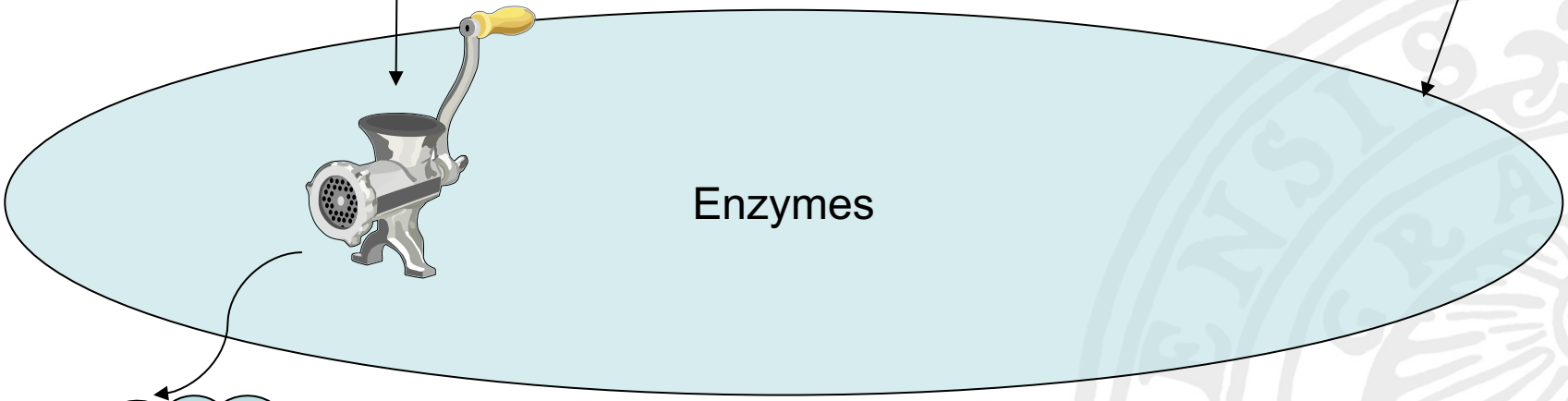


Problem setting – Drug Metabolism

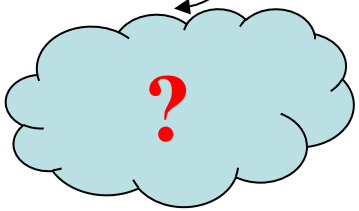


Drug candidate

The body's metabolic system



Enzymes



Same molecule? Altered?
How? Toxic? …

Predicting drug metabolism in drug discovery process

- Aid drug discovery process
  - Duration and intensity of pharmacological action
  - Metabolites might be biologically active
  - Provide early warnings → decrease costs
- Use known biotransformations
  - MDL Metabolite Database[1] 73599 recorded biotransformations
  - Use Machine Learning to predict metabolism for new drug candidates

[1]Elsevier MDL. Mdl metabolite database, 2005.
URL http://accelrys.com/products/ collaborative-science/databases/bioactivity-databases/biovia-metabolite.html.

UPPSALA UNIVERSITET

Presentation outline

1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

Describing reactions - SMIRKS

Definition

SMIRKS is a language used for describing chemical reaction transformations in a generic way[1]

Syntactic

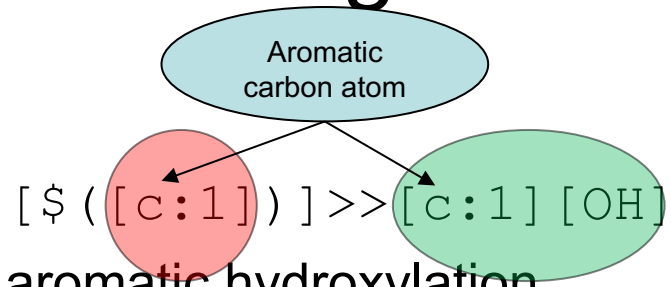
`[Reactant]>>[Product]`

Where `Reactant` and `Product` are represented as SMILES or reaction graphs

[1] Chemical Information Systems Inc. DAYLIGHT. Smirks - a reaction transform language, 2008. URL http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html.

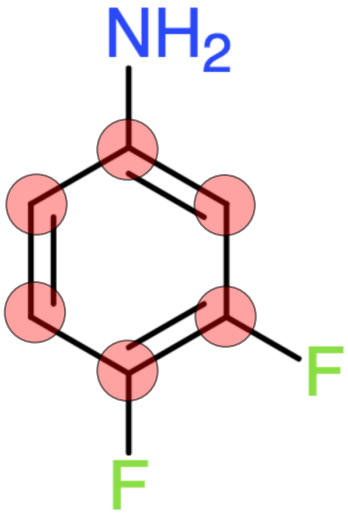
Describing reactions - SMIRKS

Example

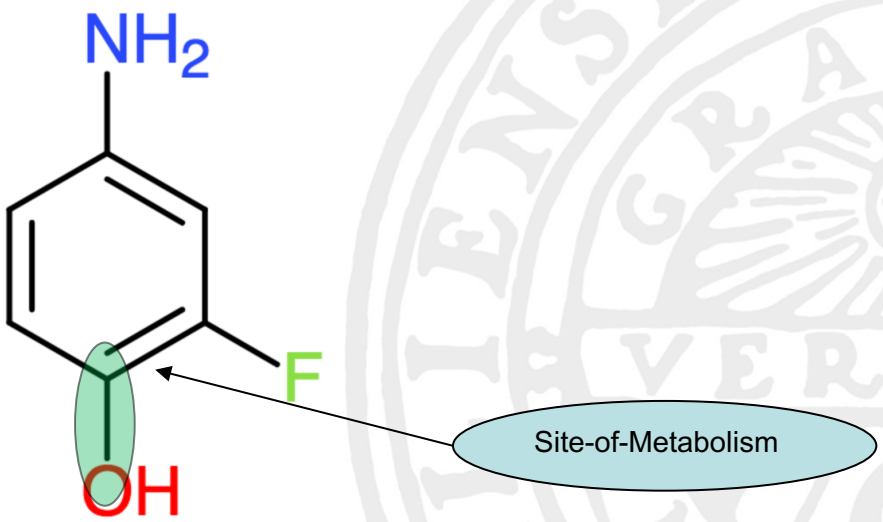


aromatic hydroxylation

Reactant



Product



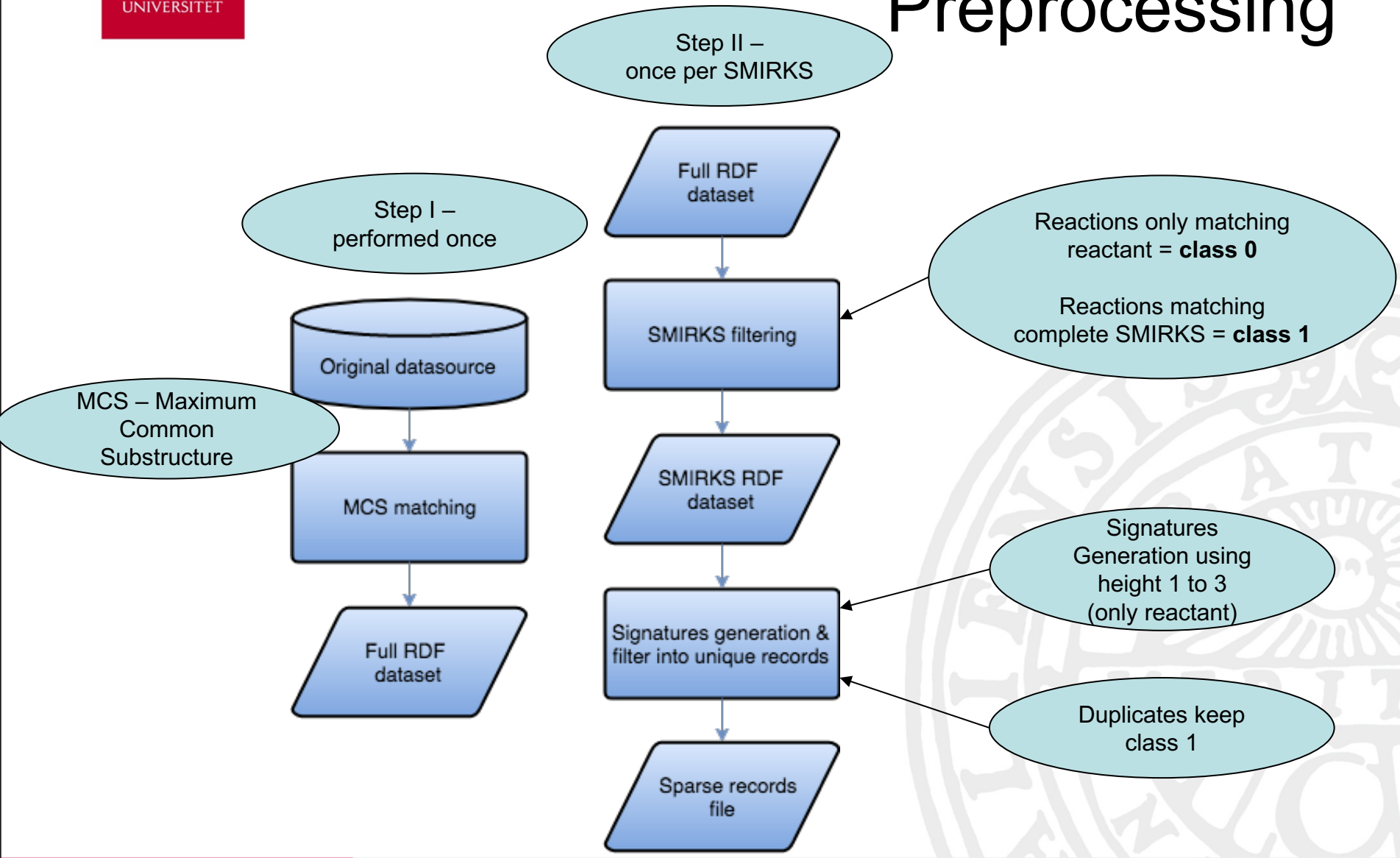
UPPSALA UNIVERSITET

Presentation outline

1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

UPPSALA UNIVERSITET

Preprocessing



UPPSALA UNIVERSITET

Derived datasets

Dataset name	SMIRKS	Biotrans-formations	Class 0
Alkyl hydroxylation	[$([C:1])]>>[C:1][OH]	17793	68%
Aromatic hydroxylation	[$([c:1])]>>[c:1][OH]	14691	85%
Carboxylation	[$([CH3:1])]>>[C:1](=O)O	12580	64%
Oxidation of tertiary amine	[$([N;X3:1])]>>[N+:1][O-]	11040	97%
Aromatization	[$([*;R;!a:1])]>>[a:1]	9518	25%

Skewed class distribution

UPPSALA UNIVERSITET

Presentation outline

1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

Machine learning context

- *Observations* $z = (x, y)$
  - x: Features derived using Signatures Descriptors[1] of height 1 to 3, total number of features was 112710
  - $y \in \{0, 1\}$, 0="no SMIRKS match", 1="SMIRKS match"
- Cross Venn-ABERS predictors
  - Using 10 Inductive Venn-ABERS predictors
  - Datasets taken fold-wise i.e. 10% in calibration set and 90% in proper training set

[1]Jean-Loup Faulon, et al. The signature molecular descriptor using extended valence sequences in qsar and qspr studies. Journal of chemical information and computer sciences, 43(3):707–720, 2003.

Machine learning context cont.

- Evaluation using outer *k*-fold cross validation, *k*=10
- Scoring algorithm: SVM in Python Scikit-learn[1]
  - RBF kernel
  - $C = 50$, $\gamma = 0.002$ previously found optimal values[2]
- Study both (p_0, p_1) interval and single prediction derived using: $p = \mathrm{GM}(p_1)/(\mathrm{GM}(1 - p_0) + \mathrm{GM}(p_1))$

[1]F. Pedregosa, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
[2]J. Alvarsson, et al. Benchmarking study of parameter variation when using signature fingerprints together with support vector machines. Journal of chemical information and modeling, 54(11):3211–3217, 2014.

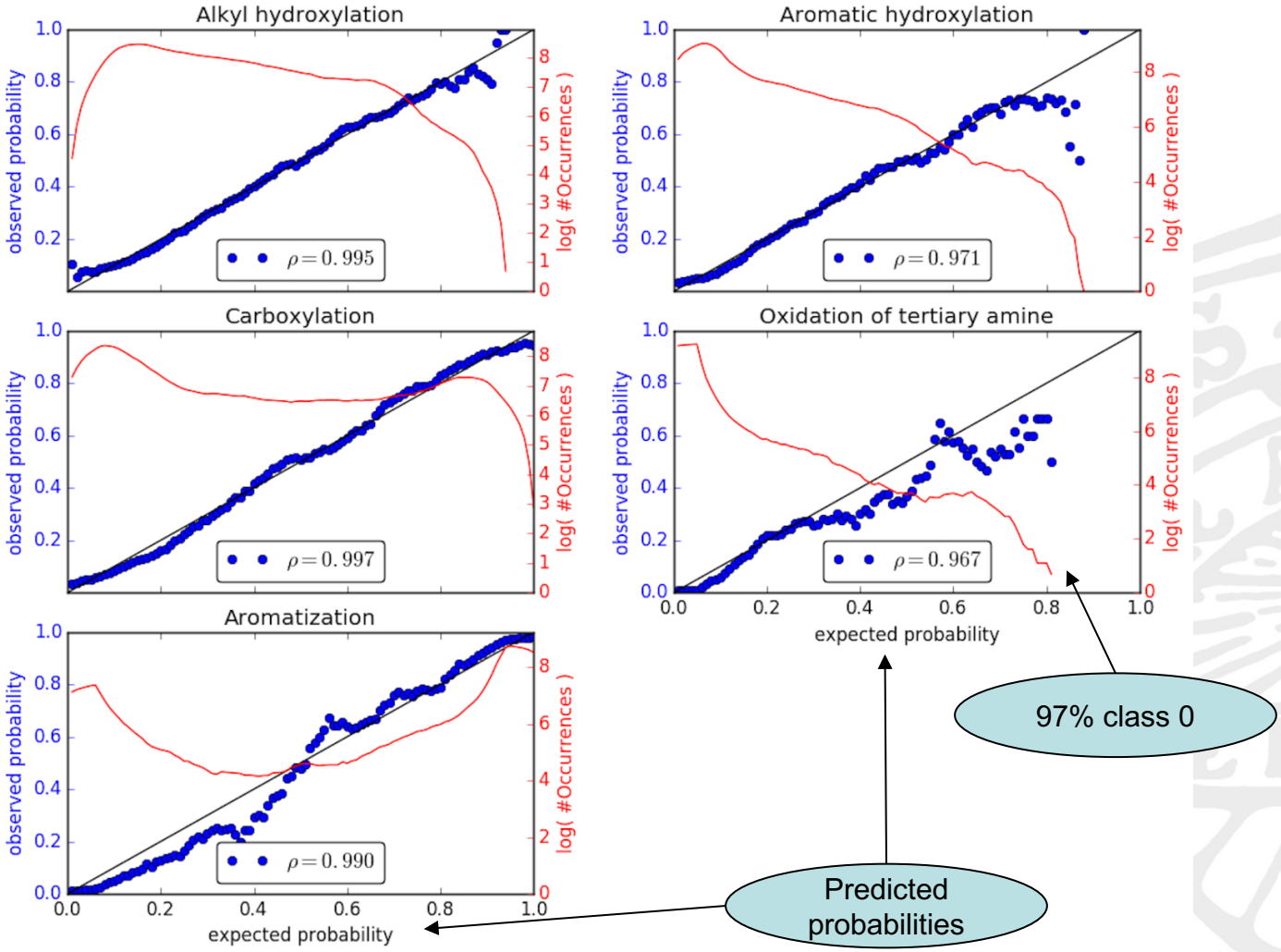
UPPSALA UNIVERSITET

Presentation outline

1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

UPPSALA UNIVERSITET

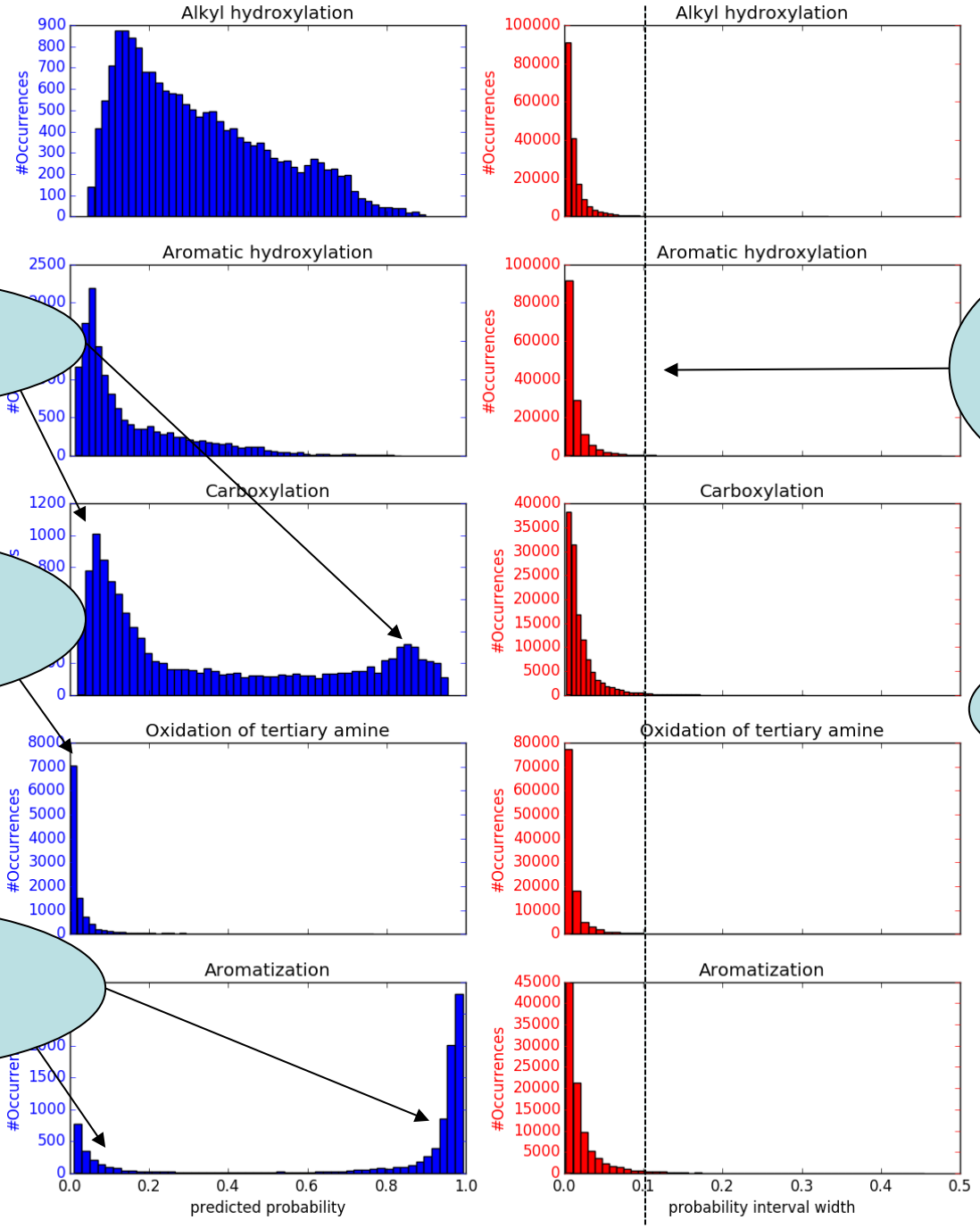
Calibration results



Observed frequencies of true label *y*=1 within bucket of size 0.1

97% class 0

Predicted probabilities



Peaks around desired prediction outcome

Almost only predicting class 0 – again due to skewed class distribution

High frequency of "informative predictions" – either class 0 or 1

Almost all objects probability intervals are less than 0.1 – median ranging between 0.006 to 0.013

Small intervals → High certainty in the predictions

UPPSALA UNIVERSITET

Performance metrics

Dataset name	Log Loss	AUC
Alkyl hydroxylation	0.538	0.753
Aromatic hydroxylation	0.348	0.793
Carboxylation	0.410	0.881
Oxidation of tertiary amine	0.093	0.904
Aromatization	0.173	0.964

UPPSALA UNIVERSITET

Presentation outline

1. Defining the problem setting
2. Describing reaction types using SMIRKS
3. Preprocessing
4. ML context
5. Results
6. Conclusions

UPPSALA UNIVERSITET

Conclusions

- Solve issue with skewed class distributions
  - Incorporate stratified sampling
  - Re-sampling of minority class
  - Mondrian classification in Conformal Prediction might be a better approach
- We achieve very good predictors in most of the datasets
  - Narrow prediction intervals for all datasets even though not always informative predictions

UPPSALA UNIVERSITET

Questions?