Graceful handling of large imbalanced datasets using Conformal Prediction

Ulf Norinder and Fredrik Svensson

ulfn@dsv.su.se , f.svensson@ucl.ac.uk

Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

Question of (un)certainty

Proceedings: 3rd Skövde Workshop on Information Fusion Topics, pp 59-62, 2009.

Utilizing Information on Uncertainty for *In Silico* Modeling using Random Forests

Henrik Boström
Dept. of Computer and Systems Sciences
Stockholm University
and
Informatics Research Centre
University of Skövde

Ulf Norinder
AstraZeneca R&D Södertälje

Introducing Uncertainty in Predictive Modeling—Friend or Foe?

Ulf Norinder*,†,‡,⊥ and Henrik Boström§

J. Chem. Inf. Model. 2012, 52, 2815−2822

Representing descriptors derived from multiple conformations as uncertain features for machine learning

Ulf Norinder · Henrik Boström

J Mol Model (2013) 19:2679–2685

Question of (un)certainty

"What we ideally would like to know is in fact that a particular prediction is derived from an area of property space from which reliable predictions are to be expected"

Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination

Ulf Norinder,[*,†] Lars Carlsson,[‡] Scott Boyer,[‡,⊥] and Martin Eklund[‡,§]

J. Chem. Inf. Model. 2014, 54, 1596–1603

Question of (un)certainty

Proceedings: 3rd Skövde Workshop on Information Fusion Topics, pp 59-62, 2009.

Utilizing Information on Uncertainty for *In Silico* Modeling using Random Forests

Henrik Boström
Dept. of Computer and Systems Sciences
Stockholm University
and
Informatics Research Centre
University of Skövde

Ulf Norinder
AstraZeneca R&D Södertälje

Proceedings: Stream KDD '10 First International Workshop on Novel Data Stream Pattern Mining Techniques, Washington, D.C. ,2010

Conformal Prediction for Distribution-Independent Anomaly Detection in Streaming Vessel Data

Rikard Laxhammar
University of Skövde
Skövde, Sweden
&
Security and Defense Solutions, Saab AB
Järfälla, Sweden
rikard.laxhammar@his.se

Göran Falkman
University of Skövde
Skövde, Sweden
goran.falkman@his.se

Introducing Uncertainty in Predictive Modeling—Friend or Foe?

Ulf Norinder*,†,‡,⊥ and Henrik Boström§

J. Chem. Inf. Model. 2012, 52, 2815−2822

Representing descriptors derived from multiple conformations as uncertain features for machine learning

Ulf Norinder · Henrik Boström

J Mol Model (2013) 19:2679–2685

Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

Question of (un)certainty

Proceedings: 3rd Skövde Workshop on Information Fusion Topics, pp 59-62, 2009.

Utilizing Information on Uncertainty for *In Silico* Modeling using Random Forests

Henrik Boström
Dept. of Computer and Systems Sciences
Stockholm University
and
Informatics Research Centre
University of Skövde

Ulf Norinder
AstraZeneca R&D Södertälje

Proceedings: Stream KDD '10 First International Workshop on Novel Data Stream Pattern Mining Techniques, Washington, D.C. ,2010

Conformal Prediction for Distribution-Independent Anomaly Detection in Streaming Vessel Data

Rikard Laxhammar
University of Skövde
Skövde, Sweden
&
Security and Defense Solutions, Saab AB
Järfälla, Sweden
rikard.laxhammar@his.se

Göran Falkman
University of Skövde
Skövde, Sweden
goran.falkman@his.se

Introducing Uncertainty in Predictive Modeling—Friend or Foe?

Ulf Norinder*,†,‡,⊥ and Henrik Boström§

J. Chem. Inf. Model. 2012, 52, 2815−2822

Representing descriptors derived from multiple conformations as uncertain features for machine learning

Ulf Norinder · Henrik Boström

J Mol Model (2013) 19:2679–2685

Question of (un)certainty

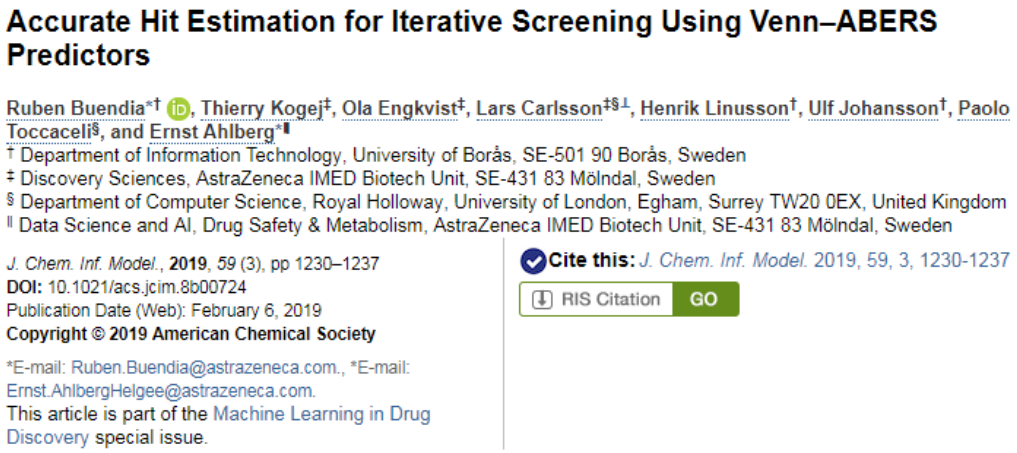
Conformal Prediction
What is it good for …?

Confidence Predictors

- Conformal Predictors

Confidence Predictors

- Conformal Predictors
- Venn (-Abers) Predictors



Accurate Hit Estimation for Iterative Screening Using Venn–ABERS Predictors

Ruben Buendia*†, Thierry Kogej‡, Ola Engkvist‡, Lars Carlsson‡§⊥, Henrik Linusson†, Ulf Johansson†, Paolo Toccaceli§, and Ernst Ahlberg*∥

† Department of Information Technology, University of Borås, SE-501 90 Borås, Sweden
‡ Discovery Sciences, AstraZeneca IMED Biotech Unit, SE-431 83 Mölndal, Sweden
§ Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom
∥ Data Science and AI, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, SE-431 83 Mölndal, Sweden

J. Chem. Inf. Model., **2019**, *59* (3), pp 1230–1237
DOI: 10.1021/acs.jcim.8b00724
Publication Date (Web): February 6, 2019

*E-mail: Ruben.Buendia@astrazeneca.com., *E-mail: Ernst.AhlbergHelgee@astrazeneca.com.
This article is part of the Machine Learning in Drug Discovery special issue.

Cite this: *J. Chem. Inf. Model.* 2019, 59, 3, 1230-1237
RIS Citation GO

Calibrated probabilities: p0 and p1

representing the minimum and maximum probability limits

Conformal Prediction

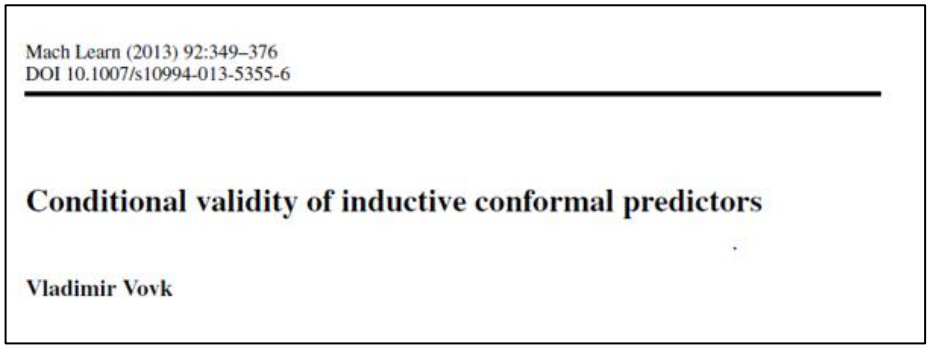
Why Conformal Prediction?

- Win situation
- Statistical guarantees (on validity)

Conformal Prediction

If {Exchangeability} then {conformal predictors are **always** valid}

Mathematical proof



Mach Learn (2013) 92:349–376
DOI 10.1007/s10994-013-5355-6

Conditional validity of inductive conformal predictors

Vladimir Vovk

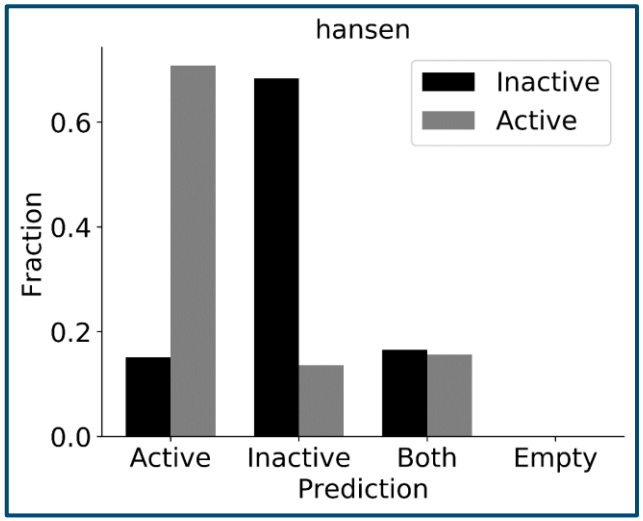
Vovk V, Gammerman A, Shafer G (2005) **Algorithmic learning in a random world**, Springer, New York

If 20 % prediction errors on **validity** acceptable --->

CP will give, at most, 20 % errors!!

Conformal Prediction

validity



Classes

Active

Inactive

Both {active, inactive)

Empty {null}

Binary classification

In conformal prediction:

If a classification contains the correct class it is correct

both = always correct, ***empty*** = always erroneous

Validity = % of correct classifications (for each class)

Efficiency = % of single label classifications (right or wrong)

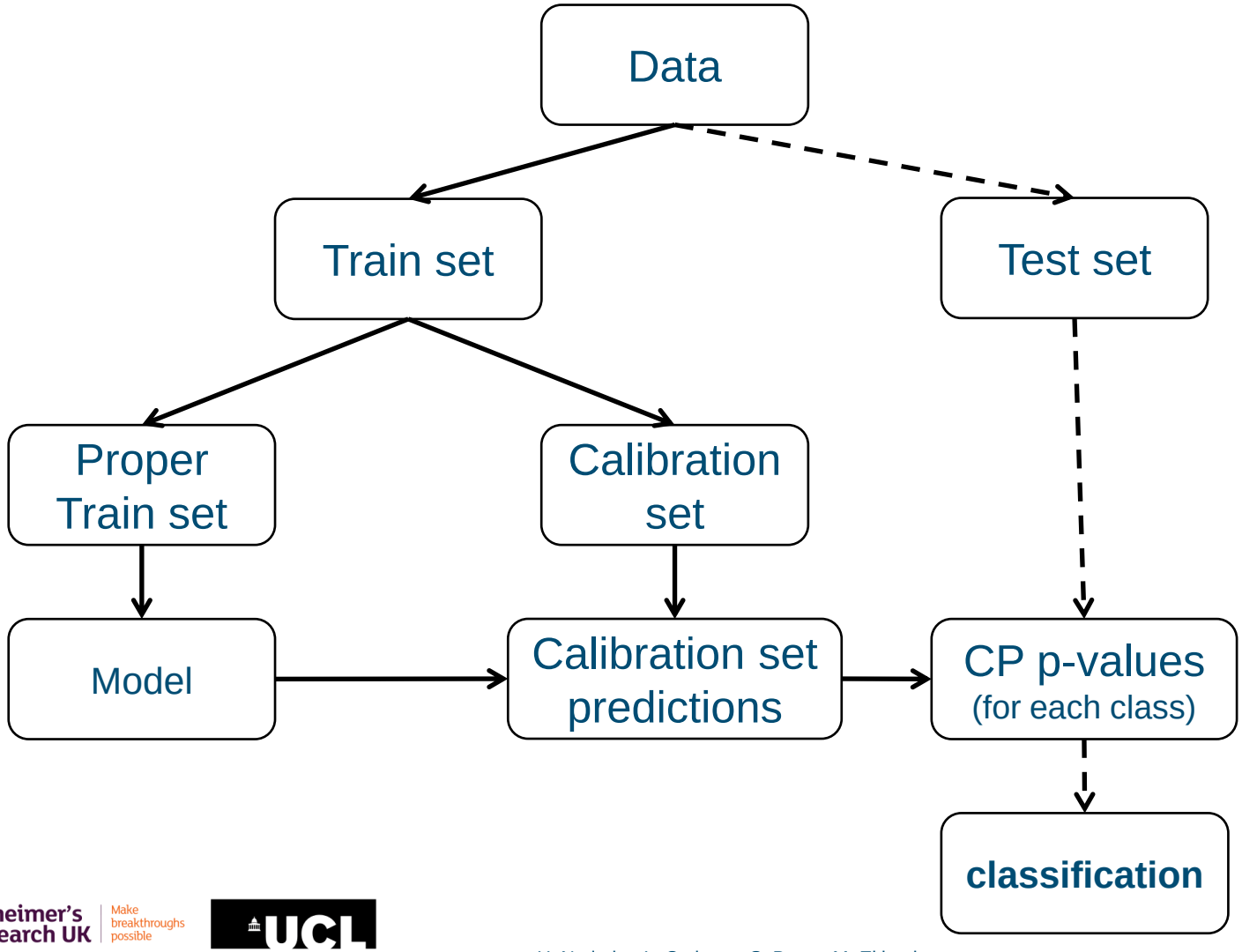
Conformal Prediction

Why Conformal Prediction?

- Win situation
- Statistical guarantees (on validity)
- CP is instance-based
- The risk is known up-front for the decision taken
- Applicability domain closely linked to model development
  - CP strictly defines the level of similarity (conformity) needed
  - No ambiguity anymore
- Gracefully handles (severely) imbalanced datasets
  - Ratios of 1:100 – 1:1000
  - No need for over- or undersampling
- CP is a framework (almost any ML algorithm will work)

Conformal Prediction

How does this work?



Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

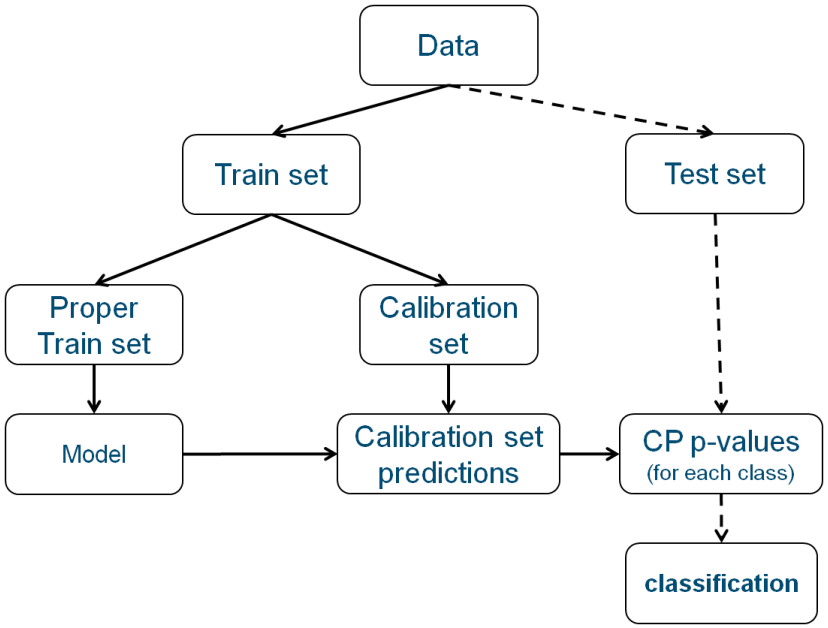
U. Norinder, L. Carlsson, S. Boyer, M. Eklund,
Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination, *J. Chem. Inf. Model.*, 2014, 54 ,596–1603

Stockholm University

CP is a framework (almost any ML algorithm will work)

- ML algorithm must provide a ranking
- Use current models, descriptors, algorithms
- Add calibration set -

  New examples in time



Conformal Prediction

- (non-) similarity function → (non-) conformity function in CP
- Compares new compounds to old (calibration) compounds
  - Defined by the user
  - **Probability from the RF trees**
  - Distance to decision plane in SVM
  - (Random numbers)

Conformal Prediction

- (non-) similarity function → (non-) conformity function in CP
- Compares new compounds to old (calibration) compounds
  - Defined by the user
  - **Probability from the RF trees**
  - Distance to decision plane in SVM

Ranking problem

CP p-value

$$|\{i = 1,\ldots,n : \beta_i \leq \beta_{new}\}| / (n+1) \geq \varepsilon$$

β_i = probability for the calibration compound i

β_{new} = probability for the new test compound

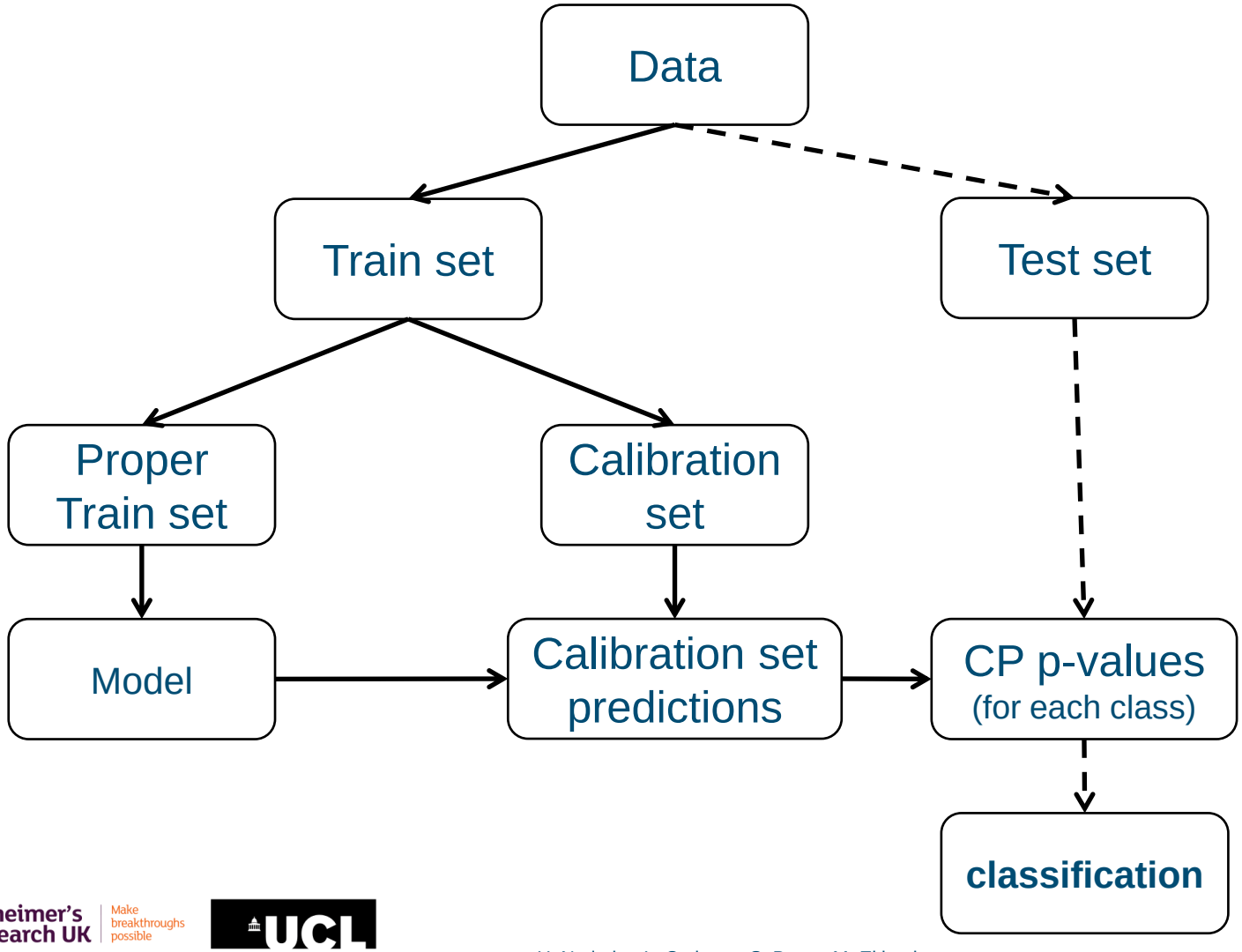
n = number of calibration set compounds

ε = significance level (% acceptable errors)

The number of calibration set compounds with probabilities ≤ probability for the new compound divided by (n+1) must be ≥ ε to be assigned a class label

Conformal Prediction

How does this work?



Alzheimer's Research UK | Make breakthroughs possible

UCL

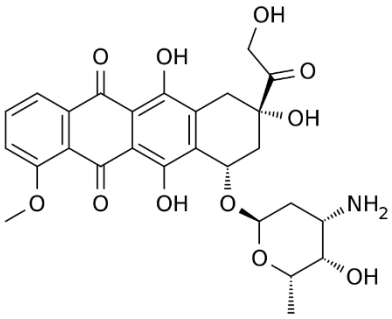
DRUG DISCOVERY INSTITUTE

U. Norinder, L. Carlsson, S. Boyer, M. Eklund,
Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination, *J. Chem. Inf. Model.*, 2014, 54 ,596–1603

Stockholm University

Conformal Prediction

Example: Predicting Toxicity



New compound to predict
(is toxic)

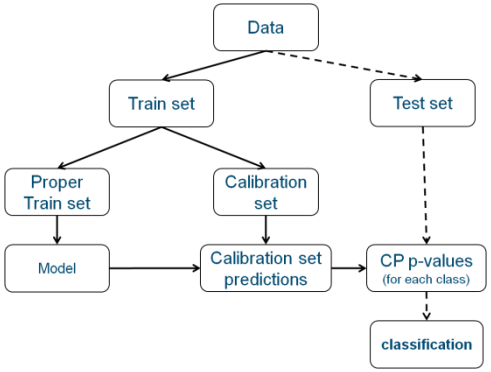
Imbalanced dataset
(toxic minority class)

A binary RF classifier (100 trees) gives the output:

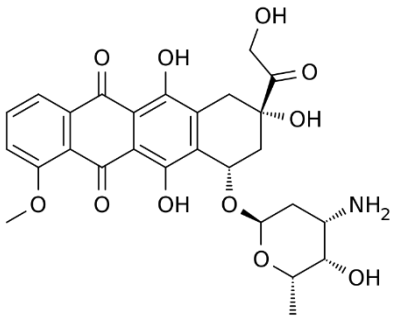
32 trees: toxic
68 trees: non-toxic

Conformal Prediction

Example: Predicting Toxicity



Calibration set, 6 toxic, 7 non-toxic compounds
N trees predicting correct class



New compound to predict
(is toxic)

32 trees: toxic
68 trees: non-toxic

Calibration set

Toxic	Non-Toxic
45	91
42	88
36	85
30	82
28	79
27	79
	77

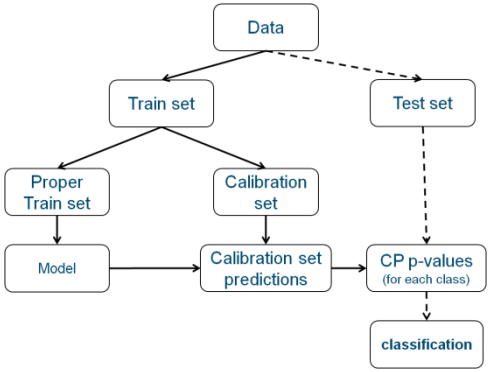
Mondrian Conformal Prediction

Alzheimer's Research UK | Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

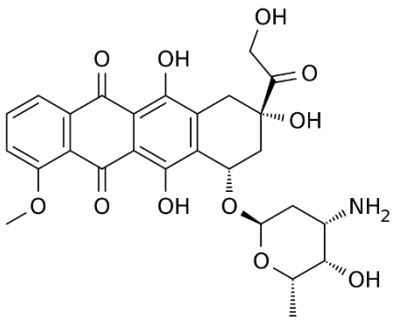
Stockholm University

Conformal Prediction

Example: Predicting Toxicity



Calibration set, 6 toxic, 7 non-toxic compounds
N trees predicting correct class



New compound to predict
(is toxic)

32 trees: toxic
68 trees: non-toxic

Calibration set

Toxic	Non-Toxic
45	91
42	88
36	85
30	82
28	79
27	79
	77

Mondrian Conformal Prediction

Alzheimer's Research UK | Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

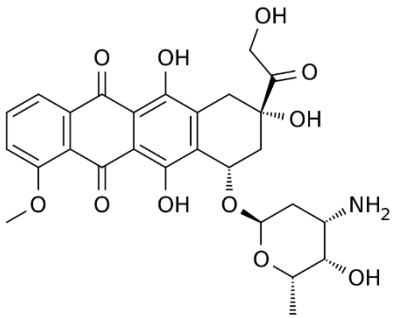
Stockholm University

Conformal Prediction

Example: Predicting Toxicity

Based on the similarity to the known examples in the calibration set:

Position toxic: 3/7
Position non-toxic: 0/8



New compound to predict
(is toxic)

32 trees: toxic
68 trees: non-toxic

Calibration set

Toxic	Non-Toxic
45	91
42	88
36	85
30	82
28	79
27	79
	77

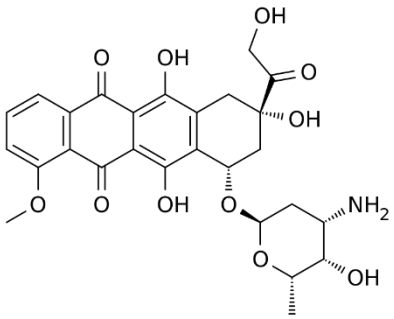
Mondrian Conformal Prediction

Alzheimer's Research UK | Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

Stockholm University

Conformal Prediction

Example: Predicting Toxicity



New compound to predict
(is toxic)

32 trees: toxic
68 trees: non-toxic

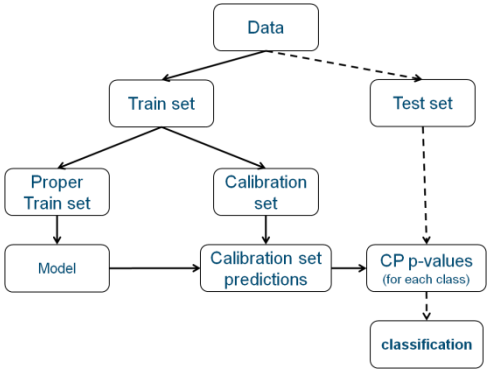
Using 80% confidence level
(0.2 significance level):

3/8 = 0.43 > 0.2 therefore the compound is assigned to the toxic class

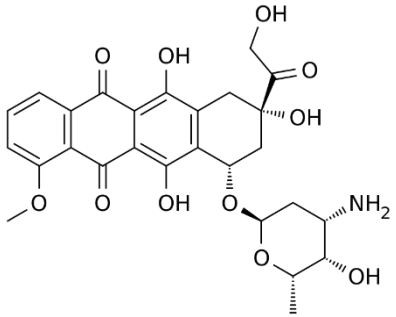
0/8 = 0.0 < 0.2 therefore the compound is not assigned to the non-toxic class

Conformal Prediction

Example: Predicting Toxicity



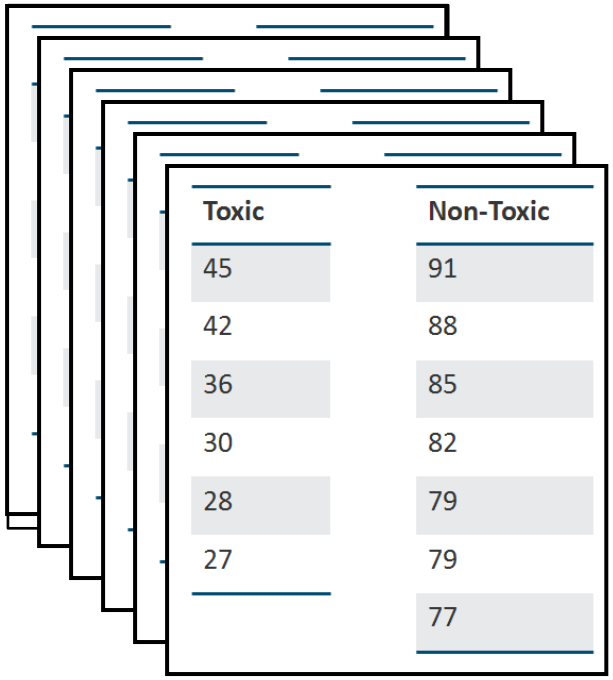
Calibration set, 6 toxic, 7 non-toxic compounds
N trees predicting correct class



New compound to predict (is toxic)

32 trees: toxic
68 trees: non-toxic

Several pairs of proper train and calibration sets



Toxic	Non-Toxic
45	91
42	88
36	85
30	82
28	79
27	79
	77

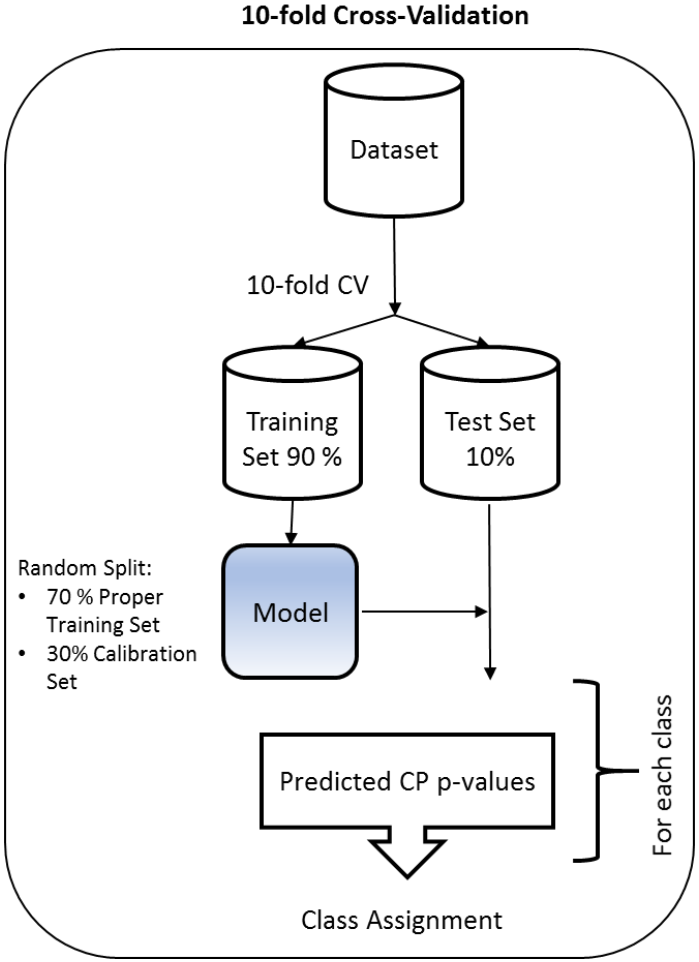
Several p-values (for each class): Use median p-value

Aggregated Mondrian Conformal Prediction
Mondrian Cross-Conformal Prediction

Alzheimer's Research UK | Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

Stockholm University

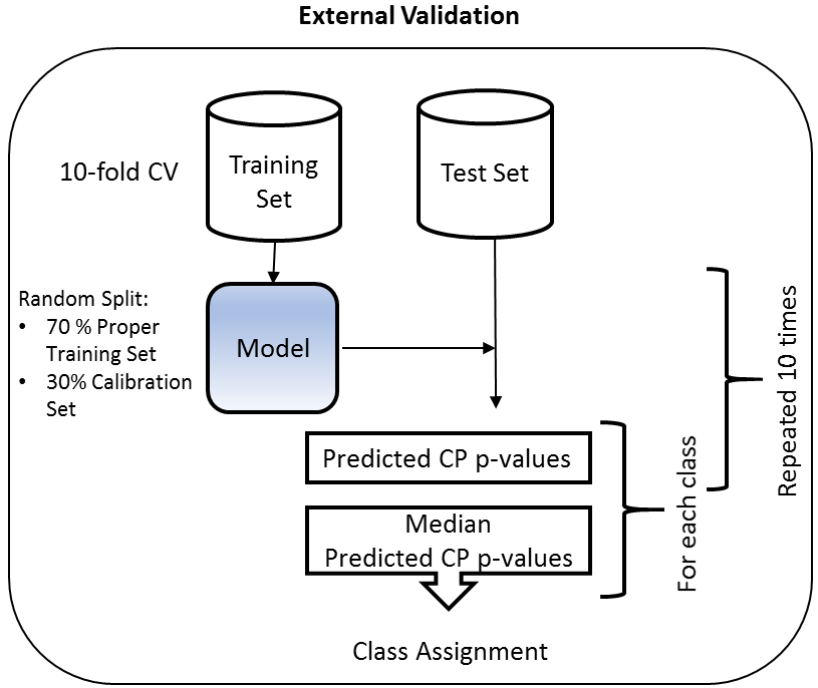
Mondrian Cross-Conformal Prediction



Mondrian = for computing p-values → study each class separately

Class 1 $|\{i = 1,\ldots,n : \beta_i \leq \beta_{new}\}| /(n+1) \geq \varepsilon$

Class 2 $|\{i = 1,\ldots,n : \beta_i \leq \beta_{new}\}| /(n+1) \geq \varepsilon$



Alzheimer's Research UK | Make breakthroughs possible | UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

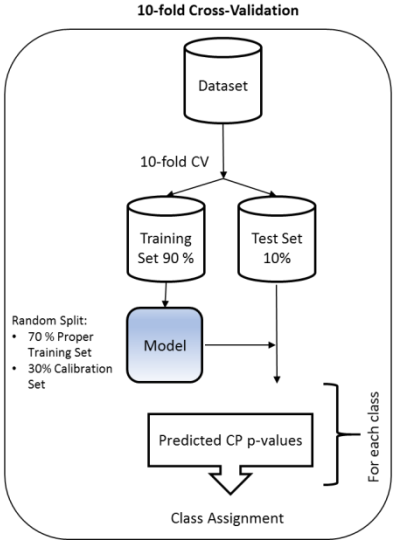
Binary Mondrian Conformal Prediction p-values

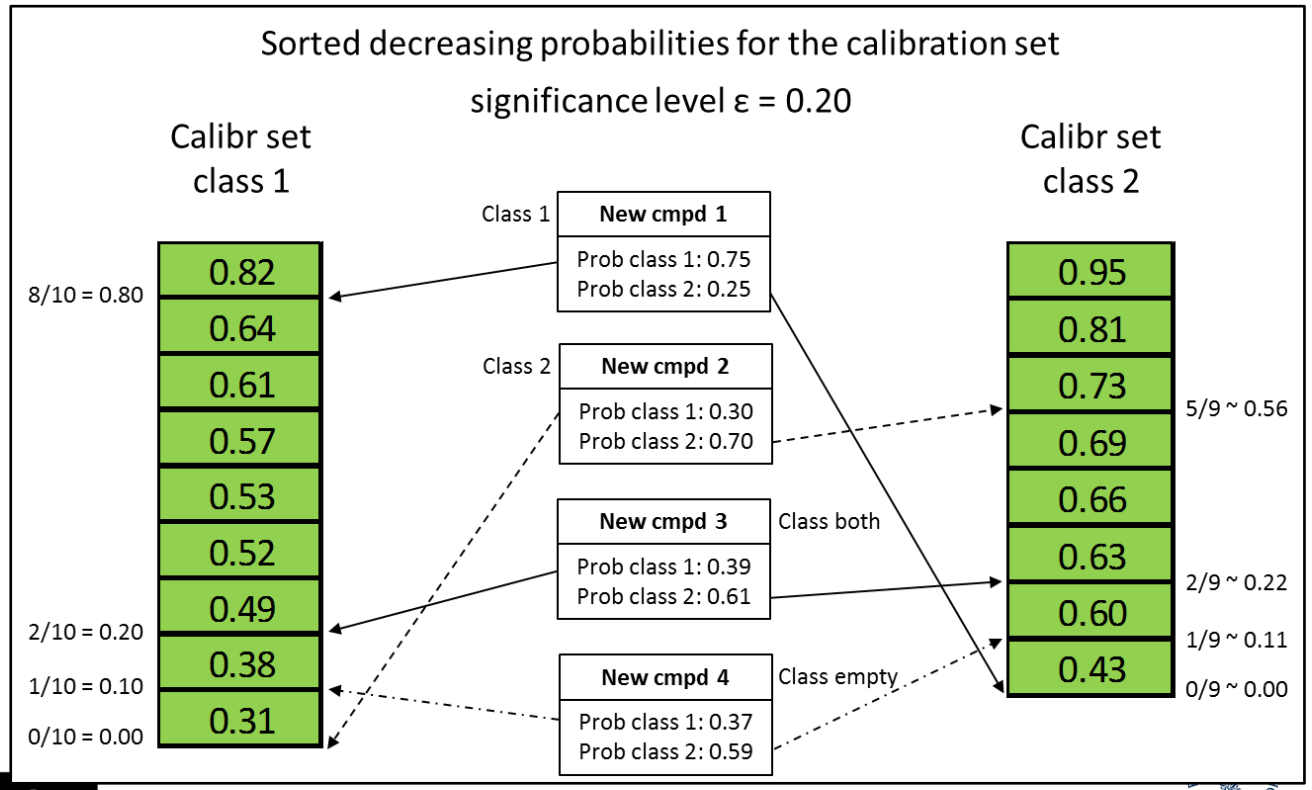
Random Split:
- 70 % Proper Training Set
- 30% Calibration Set

Model

- Build model using the proper training set
- Use calibration set for calculating p-values

For each class $|\{i = 1,\ldots,n : \beta_i \leq \beta_{new}\}| / (n+1) \geq \varepsilon$





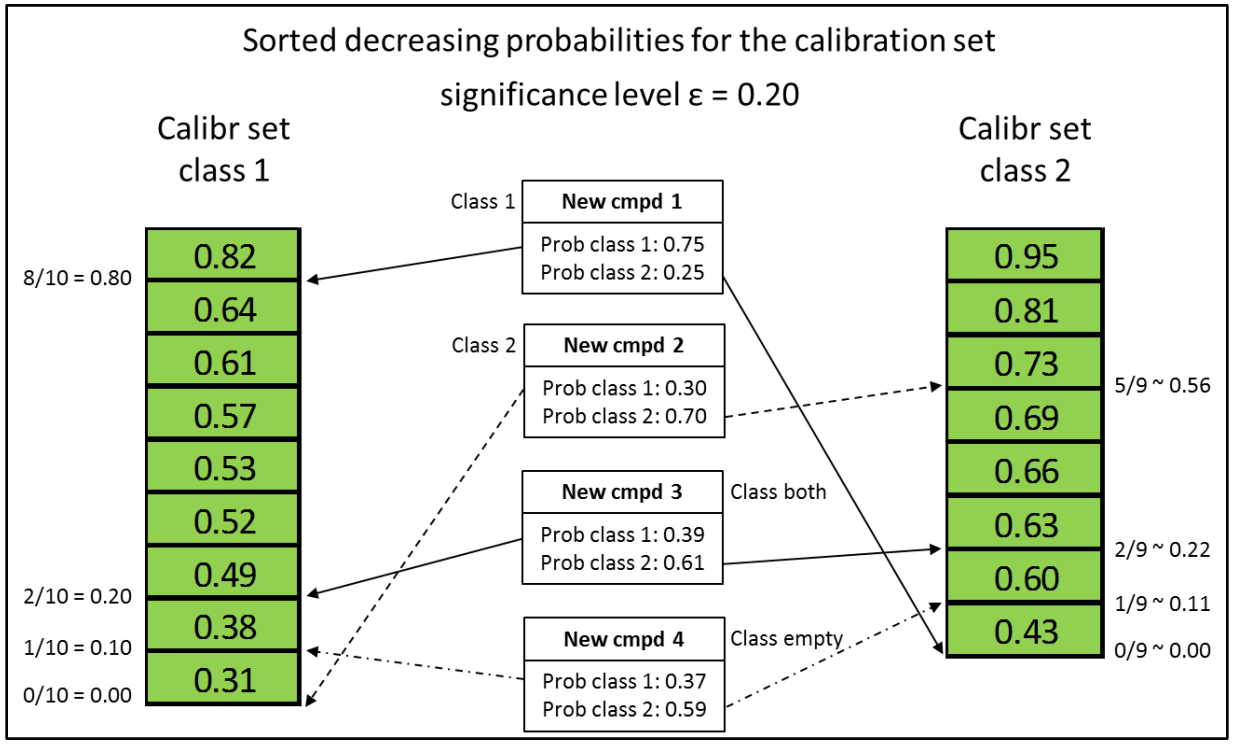
Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

Binary Mondrian Conformal Prediction p-values



In conformal prediction:

If a classification contains the correct class it is correct

both = always correct, ***empty*** = always erroneous

Validity = % of correct classifications (for each class)

Efficiency = % of single label classifications (right or wrong)

PubChem Cytotox Assays

- Results from 16 high throughput cell viability (tox) screens from PubChem
- On average 0.8% toxic compounds

AID	Tested compounds	Toxic compounds	%active	ratio non-tox/tox
624418	386 360	524	0.14	736.3
504648	367 995	600	0.16	612.3
602141	359 040	1302	0.36	274.8
620	86 701	364	0.42	237.2
847	41 152	194	0.47	211.1
903	52 783	338	0.64	155.2
2275	29 938	193	0.64	154.1
588856	404 016	3018	0.75	132.9
1825	290 605	2259	0.78	127.6
2717	299 957	3181	1.06	93.3
648	86 121	924	1.07	92.2
719	84 841	937	1.10	89.5
1486	217 851	2408	1.11	89.5
463	56 465	706	1.25	79.0
430	62 627	1121	1.79	54.9
598	85 162	5139	6.03	15.6

PubChem Cytotox Assays

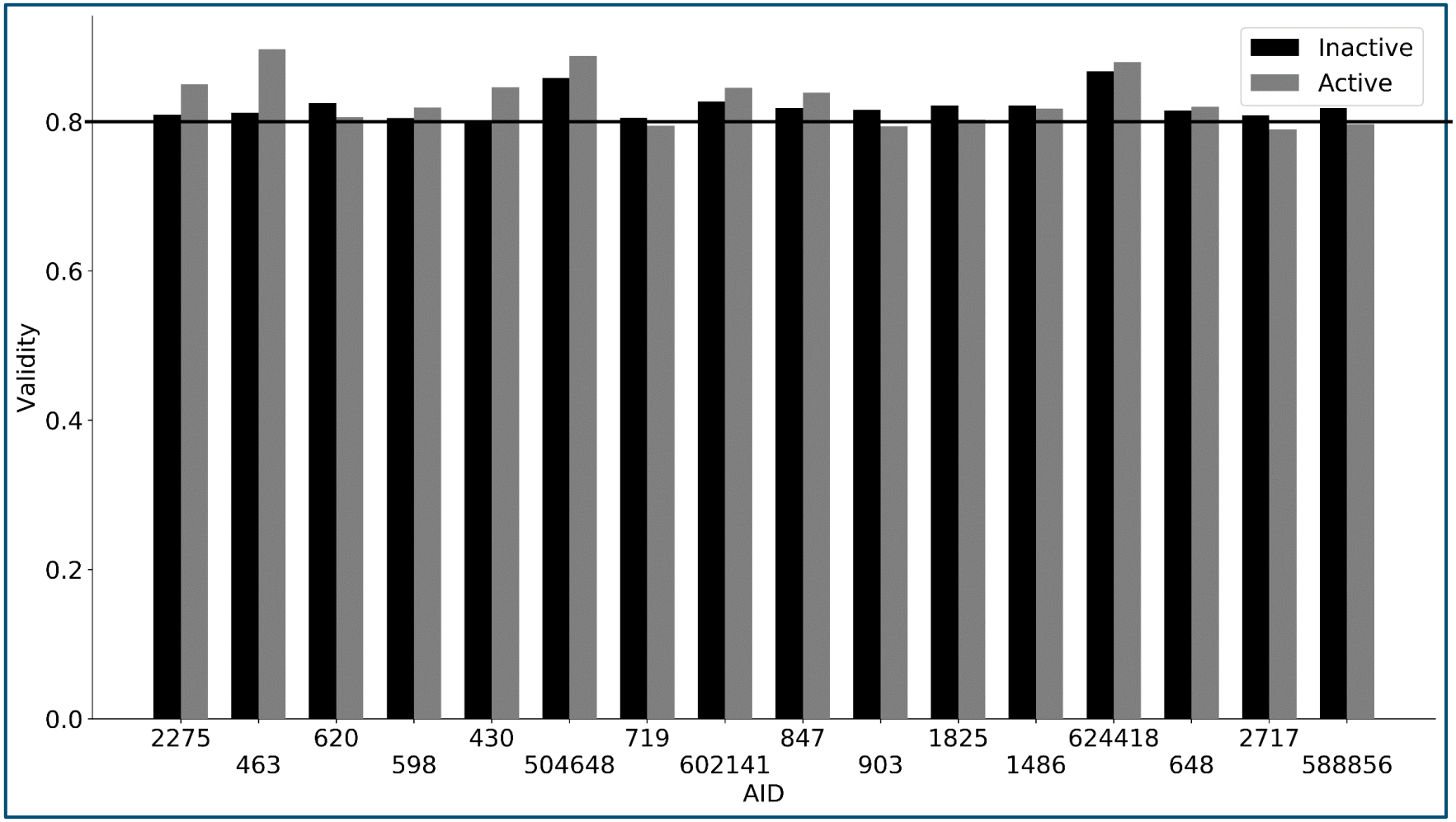
- Results from 16 high throughput cell viability (tox) screens from PubChem
- On average 0.8% toxic compounds
- RDKit descriptors
- RF, 500 trees, ensemble of 100 models
- 80 % training set, 20 % external test set

Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

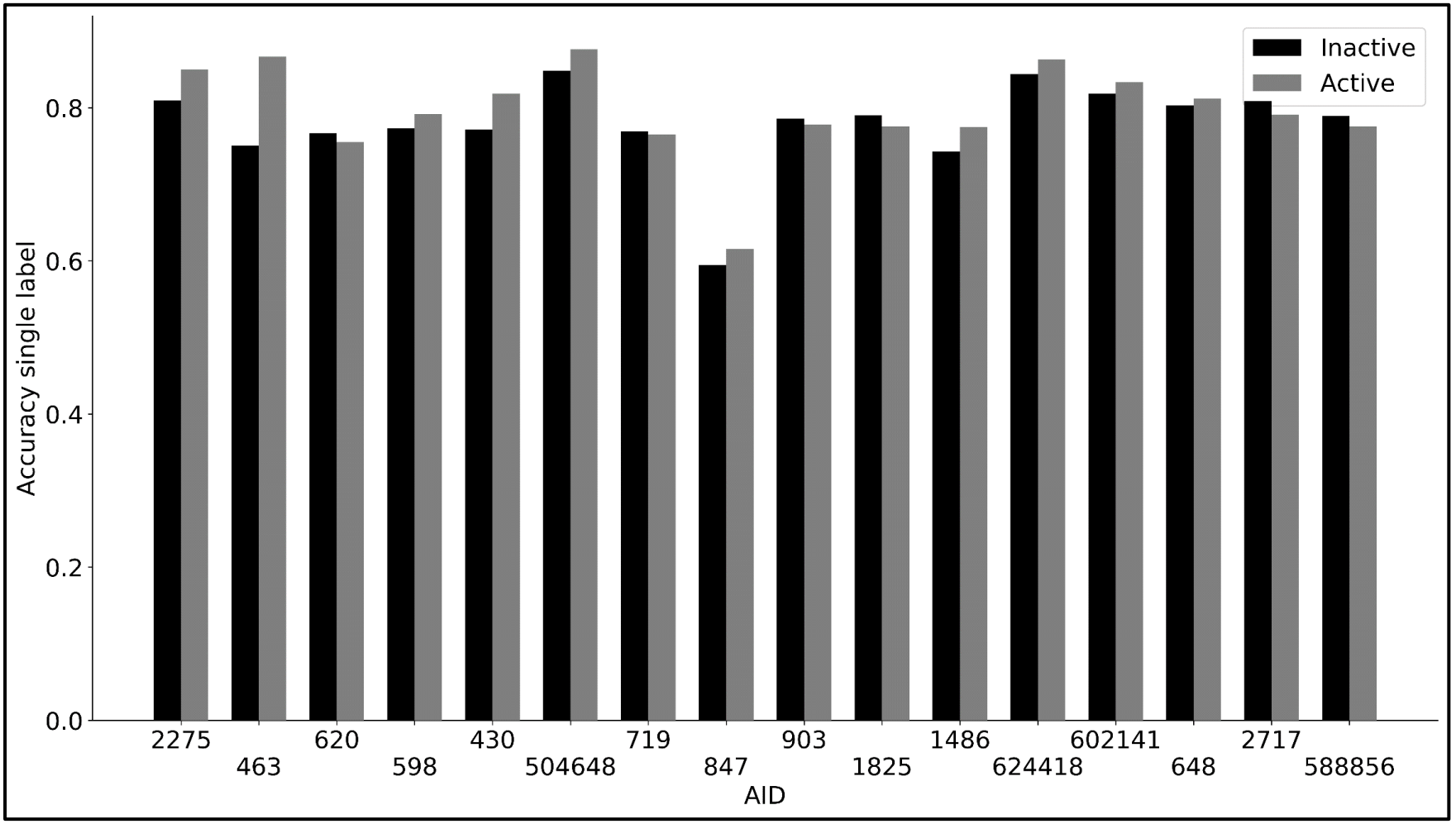
Stockholm University



Validity of the predictions (test sets) at the 80% confidence level. Models are valid for both classes.

Alzheimer's Research UK
Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

Stockholm University



Accuracy of the single label predictions (test sets) at the 80% confidence level.

The accuracy is similar for both the active and the inactive class.

Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

PubChem & Hansen Datasets

- Four dataset of different sizes and class imbalances
- 10 % randomly selected training sets
- Signature descriptors of heights 0–2 for chemical structure characterization
- Support vector machines (SVM) C-SVC, RBF kernel, parameters C =50, gamma = 0.002
- Ensemble of 100 SVM models

Binary classification of imbalanced datasets using conformal prediction

Ulf Norinder*, Scott Boyer

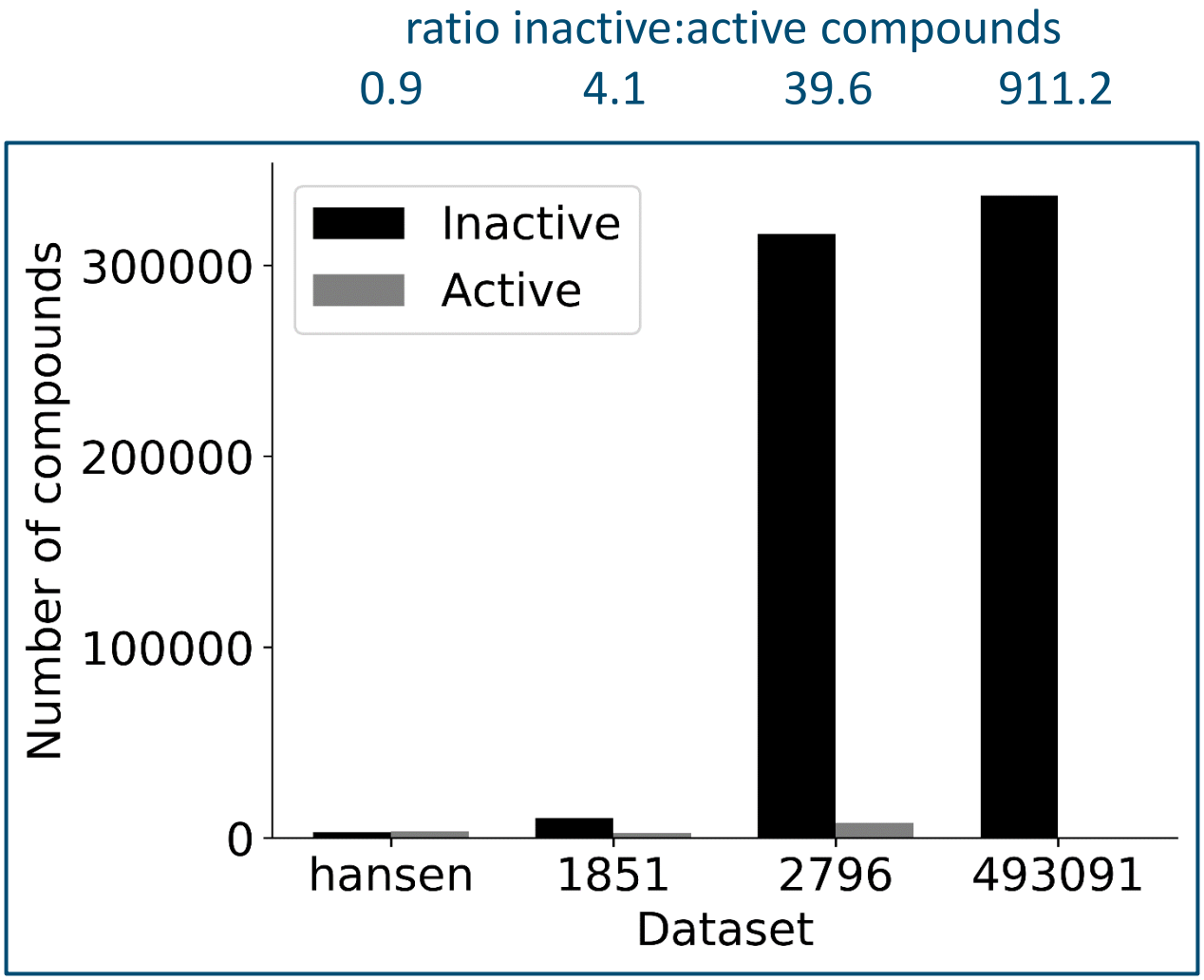
Journal of Molecular Graphics and Modelling 72 (2017) 256–265

Alzheimer's Research UK | Make breakthroughs possible

UCL

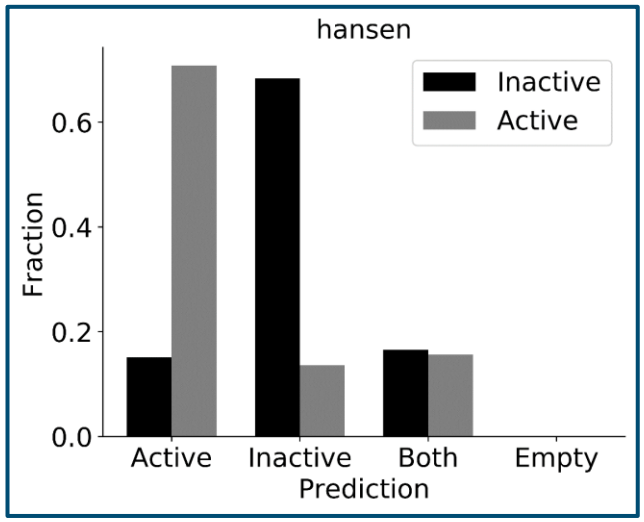
DRUG DISCOVERY INSTITUTE

Stockholm University

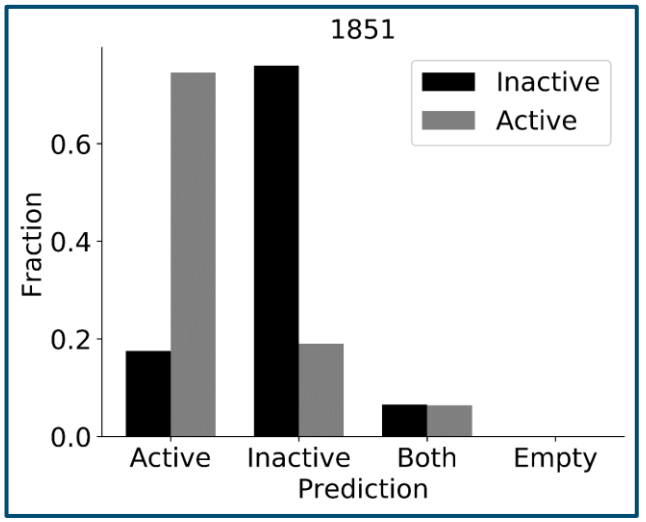


Size and imbalance differs considerably between the datasets.

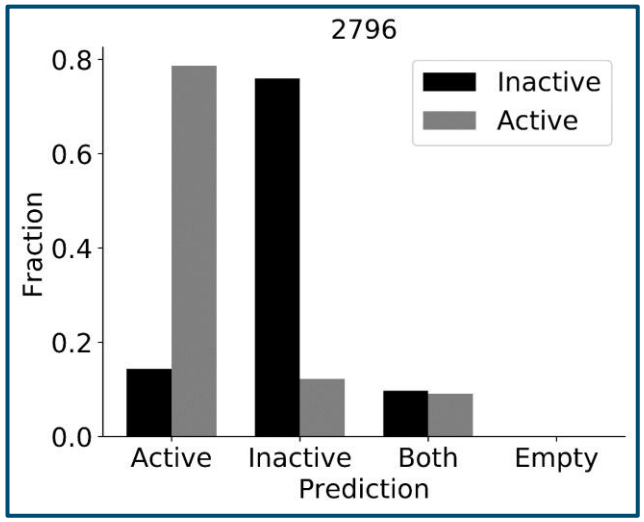
Fraction predicted active and inactive compounds.



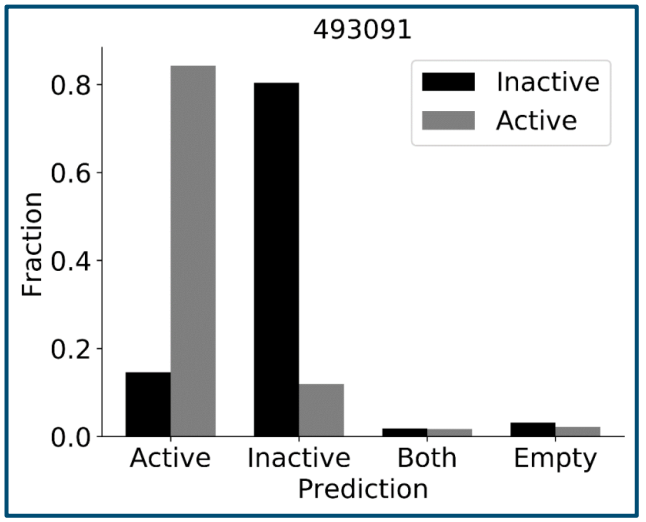
0.9



4.1



39.6



911.2

Alzheimer's Research UK | Make breakthroughs possible

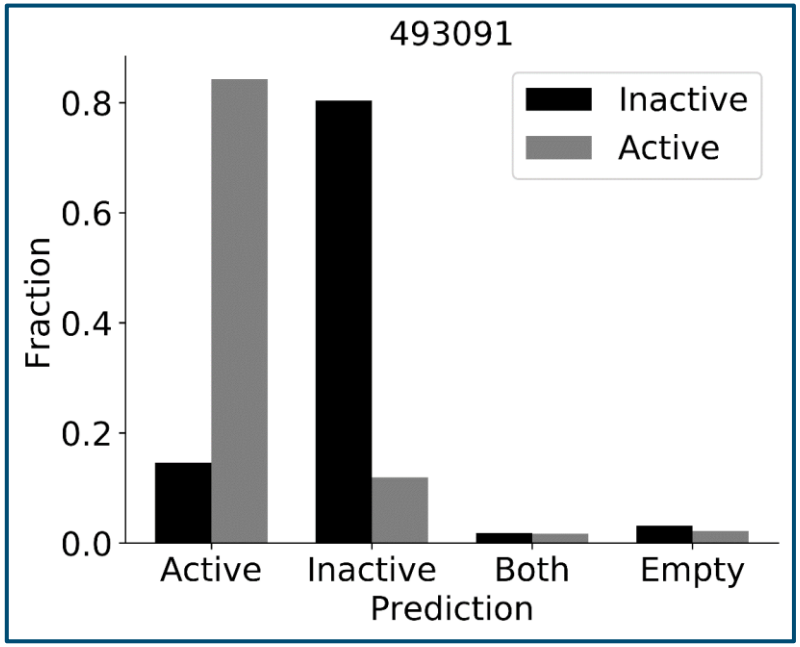
UCL

DRUG DISCOVERY INSTITUTE

Results are similar across the datasets despite the varying imbalance.

Stockholm University

#compounds in ***both*** class & ***empty*** class

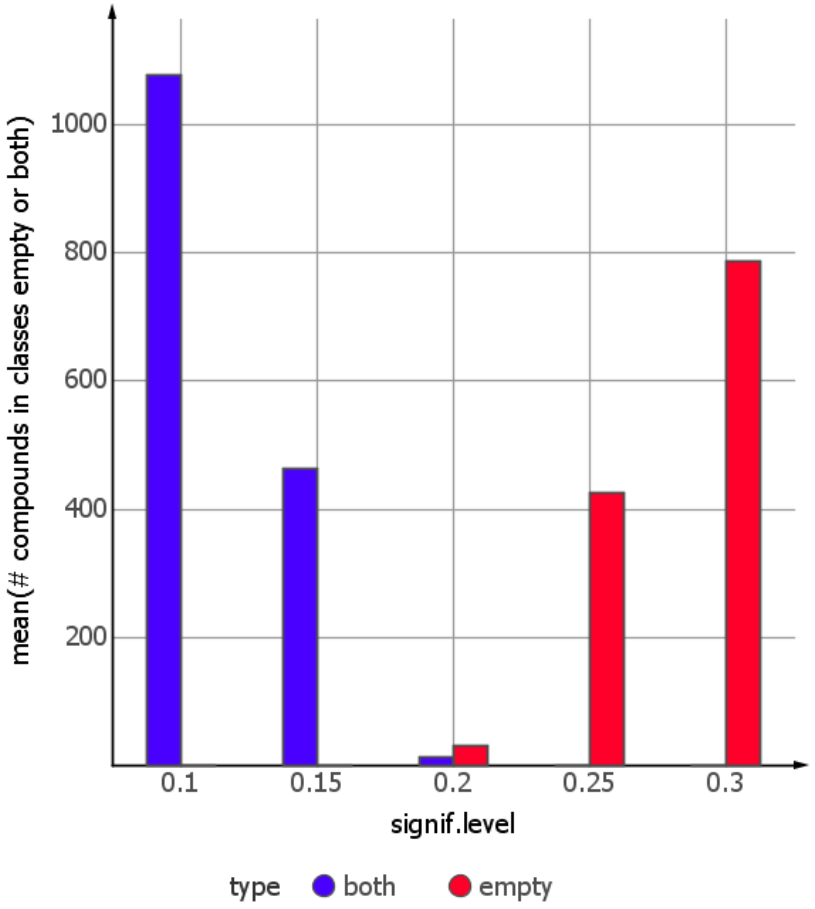


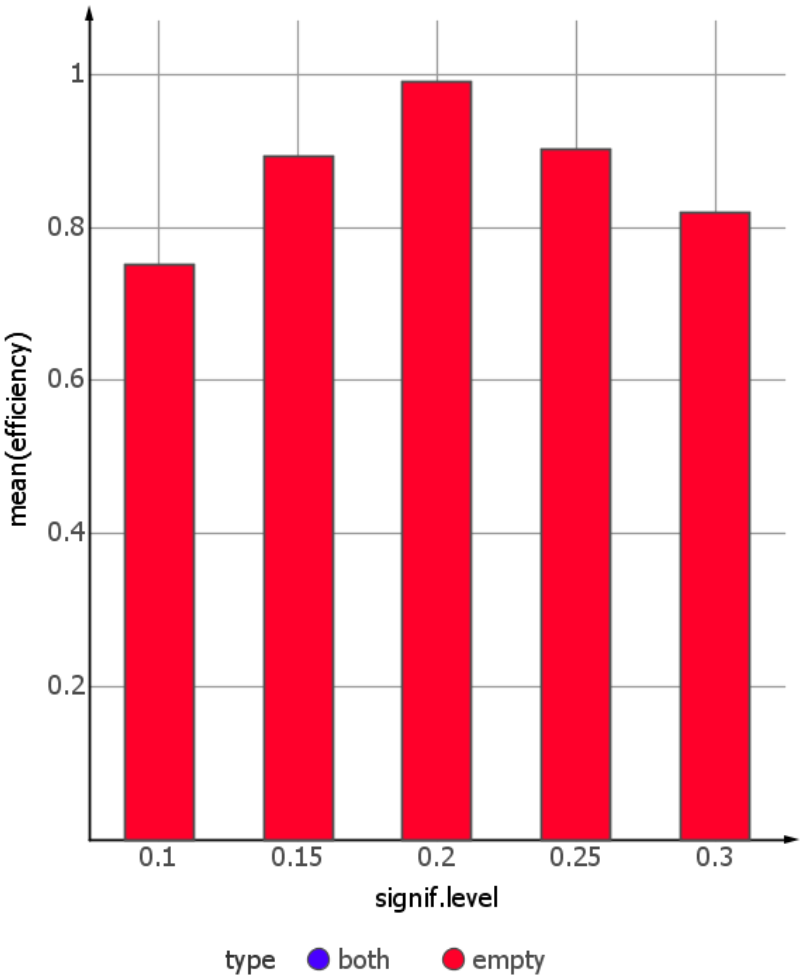
@acceptable significance level:

Results from new data →

- Many predictions in ***empty*** class → outside AD of current model → measure and update model
- Many predictions in ***both*** class → inside AD of current model → lack of information → add new information (descriptors), develop better model (classifier, algorithm)

@acceptable significance level (decided by the user)





If a classification contains the correct class it is correct
both = always correct, ***empty*** = always erroneous

Efficiency = % of single label classifications (right or wrong)

Alzheimer's Research UK | Make breakthroughs possible

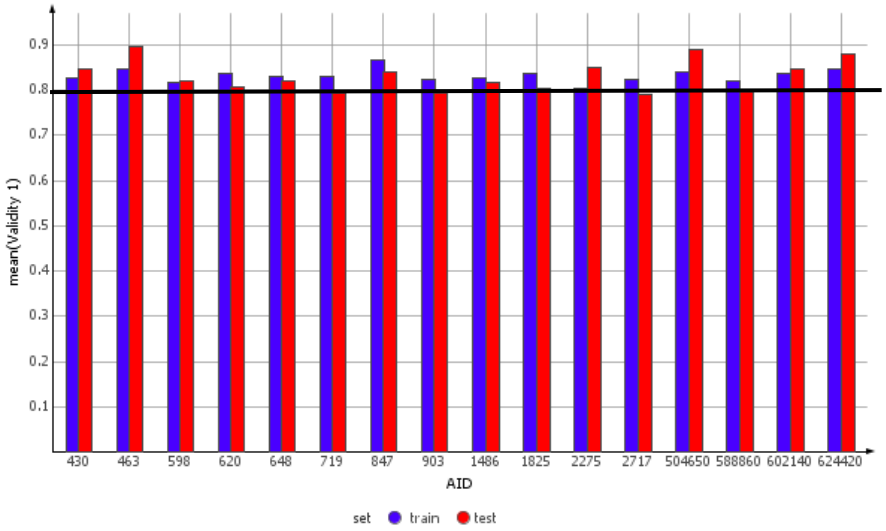
UCL

DRUG DISCOVERY INSTITUTE

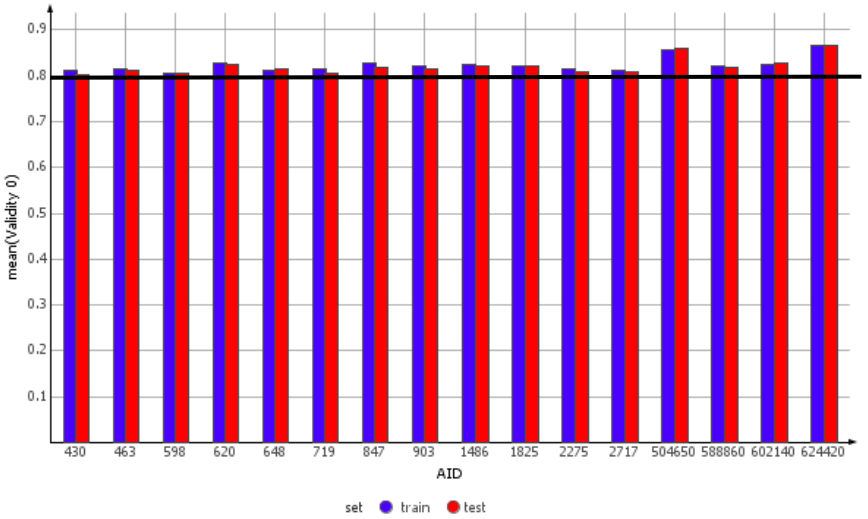
Stockholm University

Not over-optimistic models

Validity minority class



Validity majority class



Signif. level 0.2

Alzheimer's Research UK | Make breakthroughs possible

UCL

DRUG DISCOVERY INSTITUTE

Stockholm University

Acknowledgements

Dr. Andreas Bender, Cambridge Univ

Dr. Lars Carlsson, Royal Holloway, Univ London

Dr. Martin Eklund, Karolinska Institutet

Dr. Ola Spjuth, Uppsala Univ

Dr. Scott Boyer, former Swetox

Dr. Natalia Aniceto, Cambridge Univ

The Francis Crick Institute