Applicability domain of QSAR models: status quo and perspectives

Iurii Sushko

8 October 2012

Introduction

Types of models: specific and general

The are many types of models:

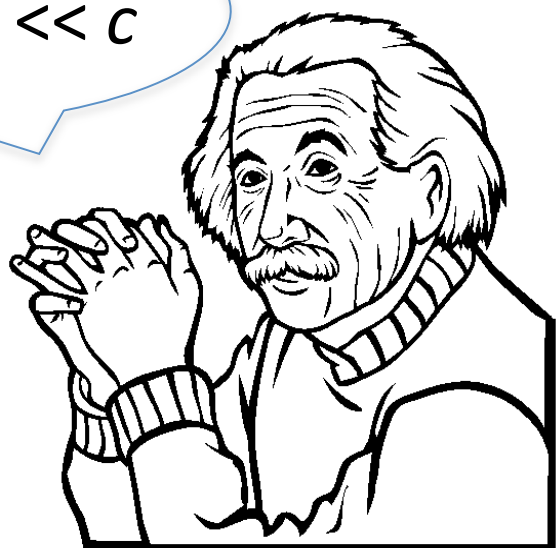
- QSAR models for risk assessment
- Chemoinformatics and bioinformatics models
- Financial markets forecast
- Weather forecast
- Chemistry and physics laws

No models are universal

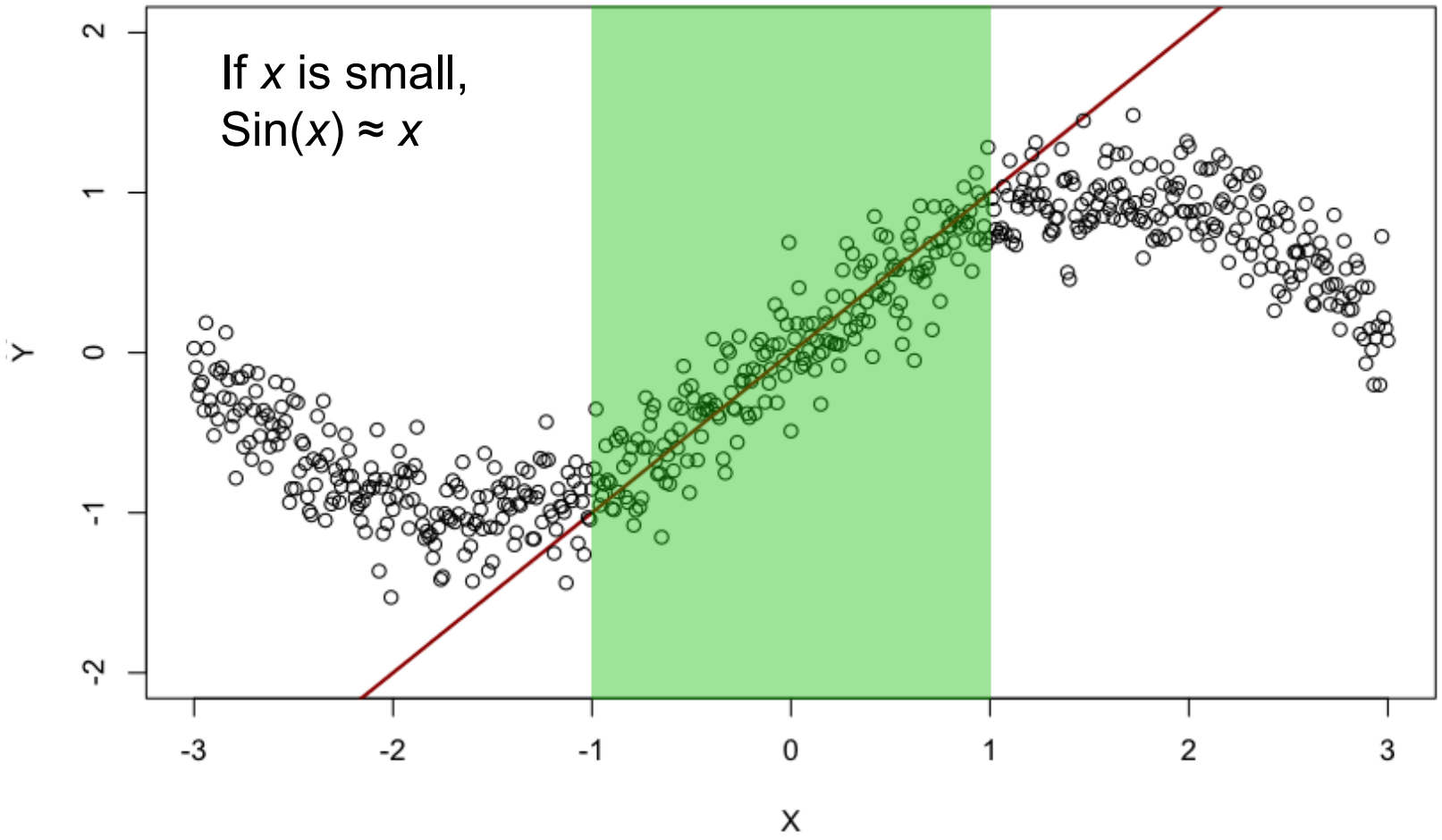
Example 1: The Newton's laws

$F = m \bullet a$

…only if $v << c$



Example 2: simple regression



QSAR truth is even more bitter

- Predicted dependencies are very complex
  (chemistry, biology)
- Dimensionality is high
  (hundreds or even thousands of descriptors)
- Data is limited
  (infinitely small coverage of infinitely large chemical space)

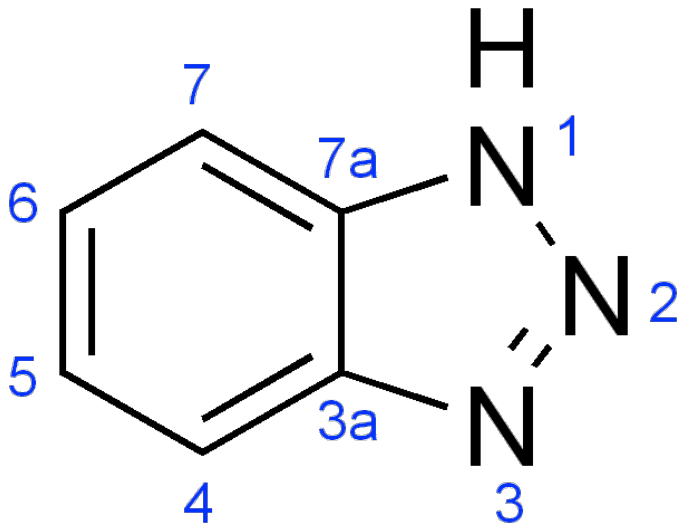
QSAR model are very limited
(also recognized by OECD)

How to define the applicability domain of QSAR models?
How to distinguish accurate and inaccurate predictions?

AD assessment:
from simple to complex

Structural applicability domain

Limit the applicability domain to particular chemical classes.
Example: benzotriazoles, perflourinated compounds, Platinum complexes, etc.



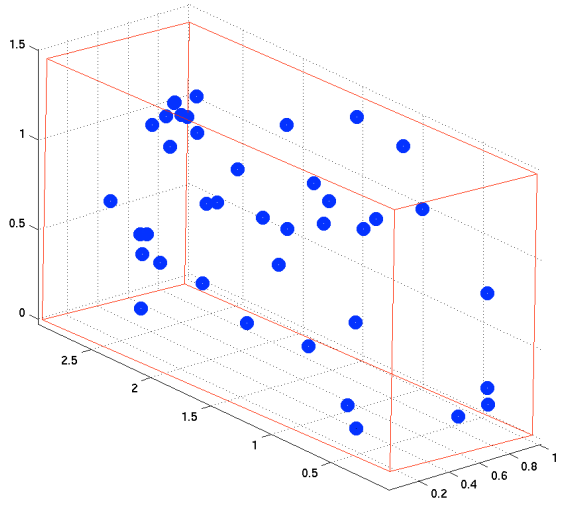
Pros:
- Simplicity
- Easy interpretability by chemists
- Easy technical implementation (set of SMARTS patterns)

Cons:
- The definition is too broad
- Not always possible to define

Descriptors bounding box

Disallow each molecular descriptor to exceed the range from the training set.

Extensions: PCA bounding box, convex hull.



Pros:
- Simplicity
- Easy implementation

Cons:
- The definition is too broad and naïve
- Intolerant to multiple clusters
- Intolerant to non-convex clusters
- Difficult chemical interpretation

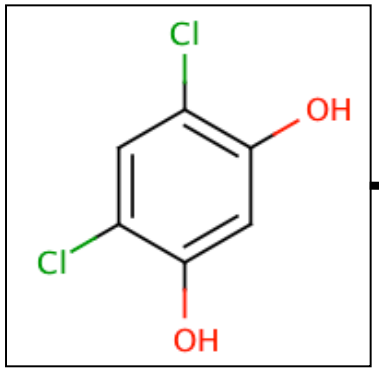
Distances to models (DMs)

DM is any numerical measure of the prediction reliability for a particular chemical compound by a given model

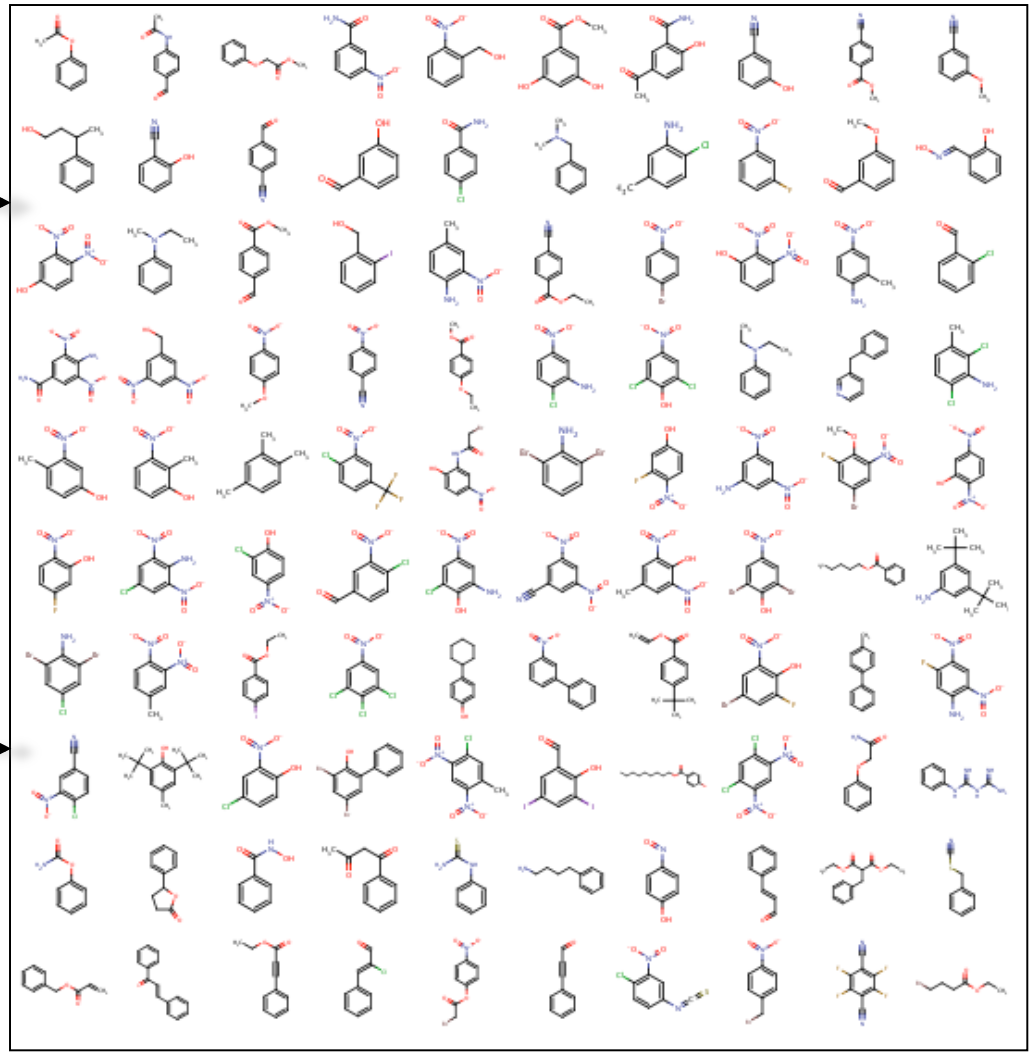
f (*model, chemical compound*) → *prediction reliability*

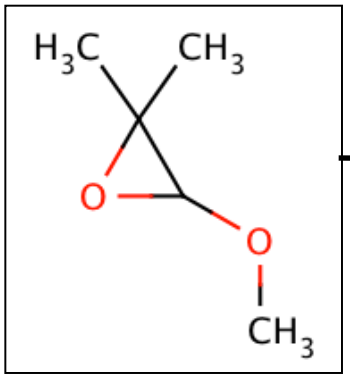
Applicability domain can be defined as
all molecules with **DM less than a particular threshold**

Dissimilarity to the training set



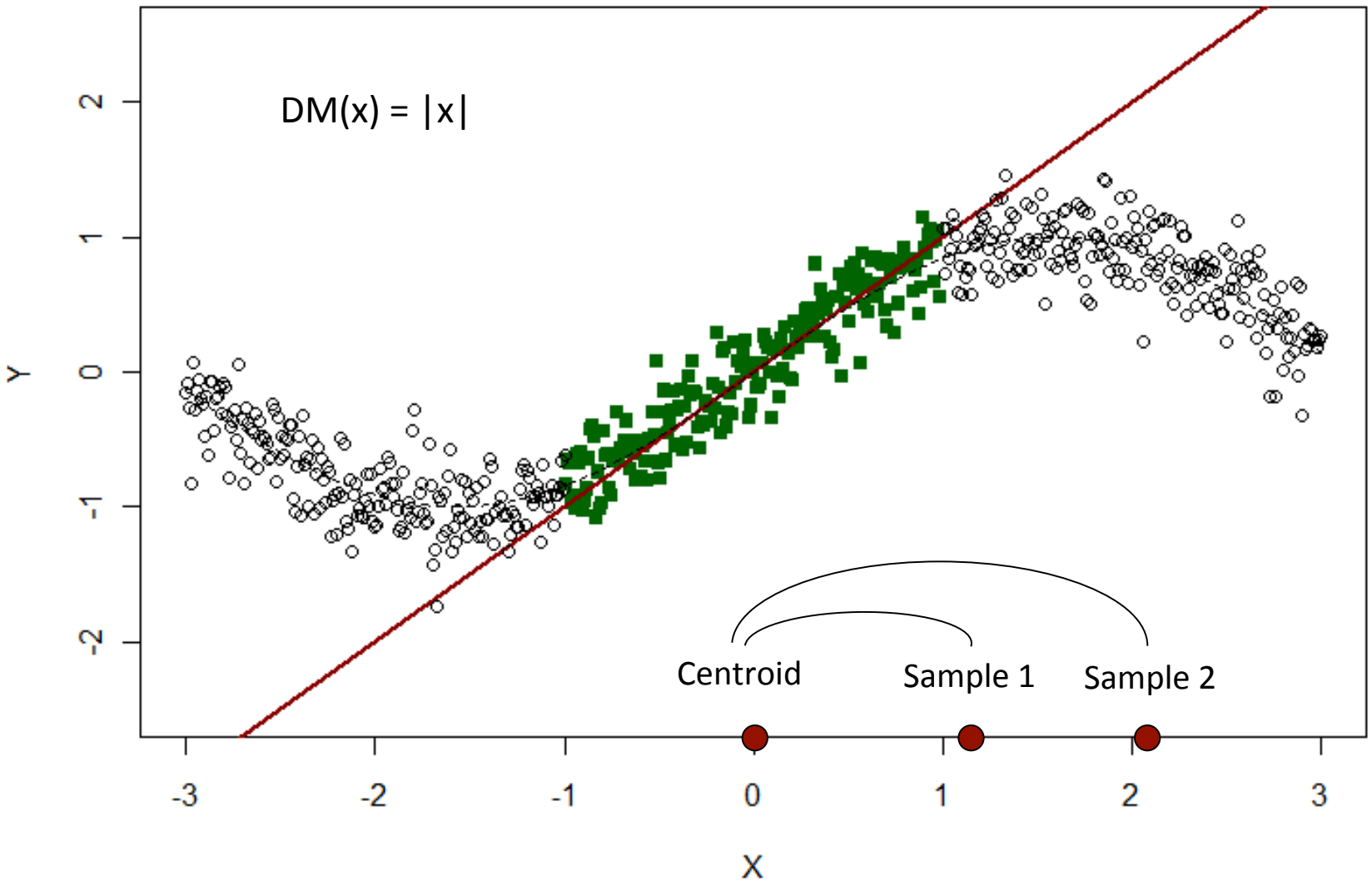
similar





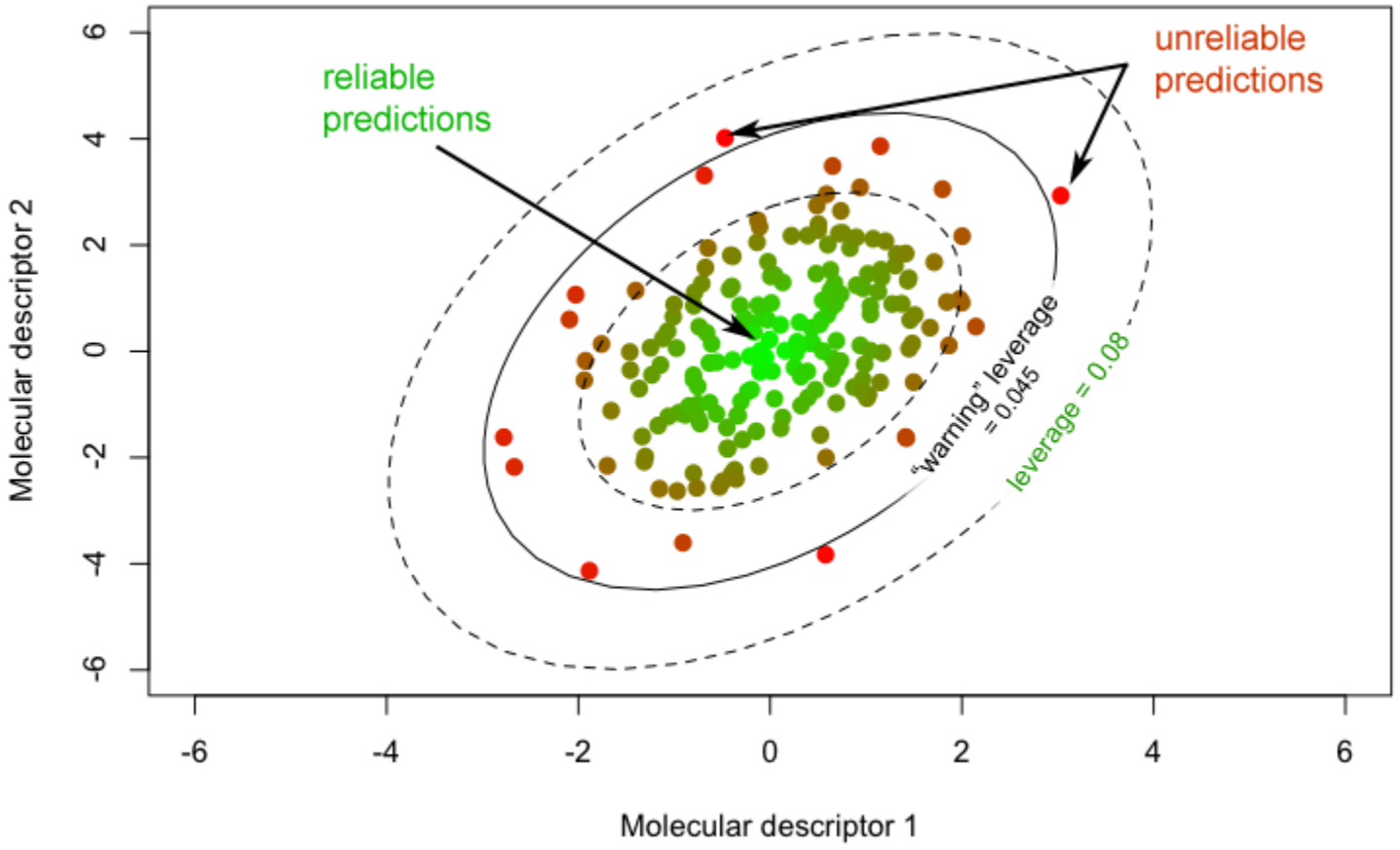
dissimilar

Dissimilarity: Distance to centroid

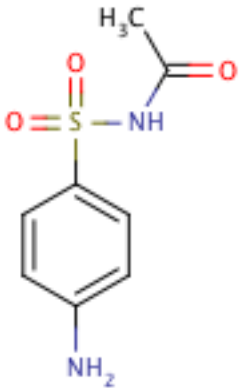


Dissimilarity: Leverage statistic

DM(J) = LEVERAGE(J) = $\overrightarrow{x(J)} \cdot (X^T \cdot X)^{-1} \cdot \overrightarrow{x(J)}^T$



Disagreement in prediction



logS = ?

Model 1 logS = -1.51

Model 2 logS = -1.70

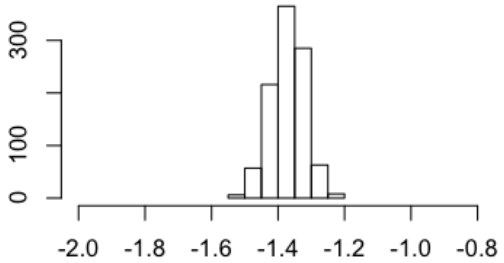
Model 3 logS = -1.43

Model 4 logS = -1.81

.....

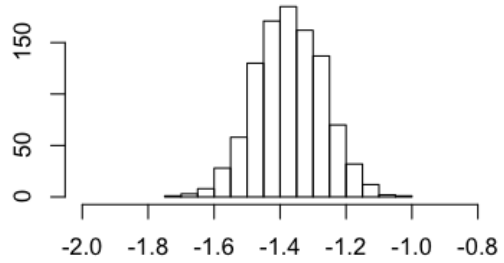
Model 99 logS = -2.01

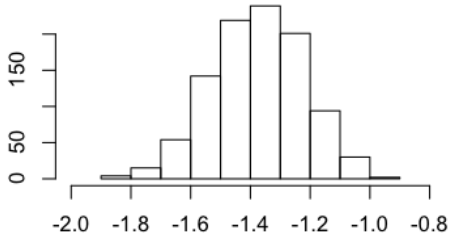
Model 100 logS = -1.54



Low uncertainty

STD = 0.05

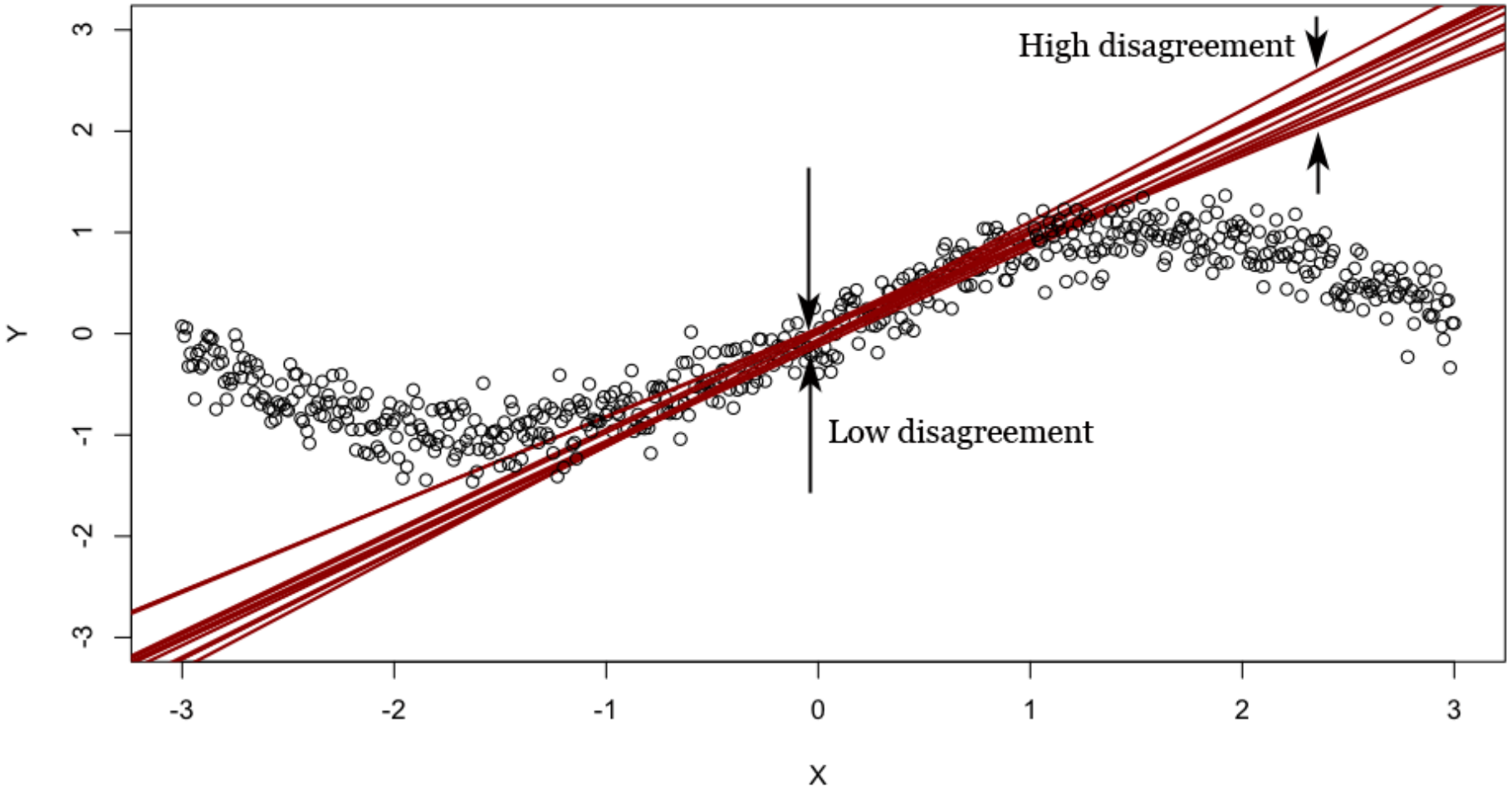




High uncertainty

STD = 0.15

DM 2: Disagreement



DMs overview

Similarity-based

- Distance to centroid
- Leverage
- Tanimoto

Agreement-based

- Standard deviation
  - Ensemble
  - Consensus
- Voting concordance

DM versus Accuracy

"Reliability" (DM) is *subjective*. Accuracy is *objective*.

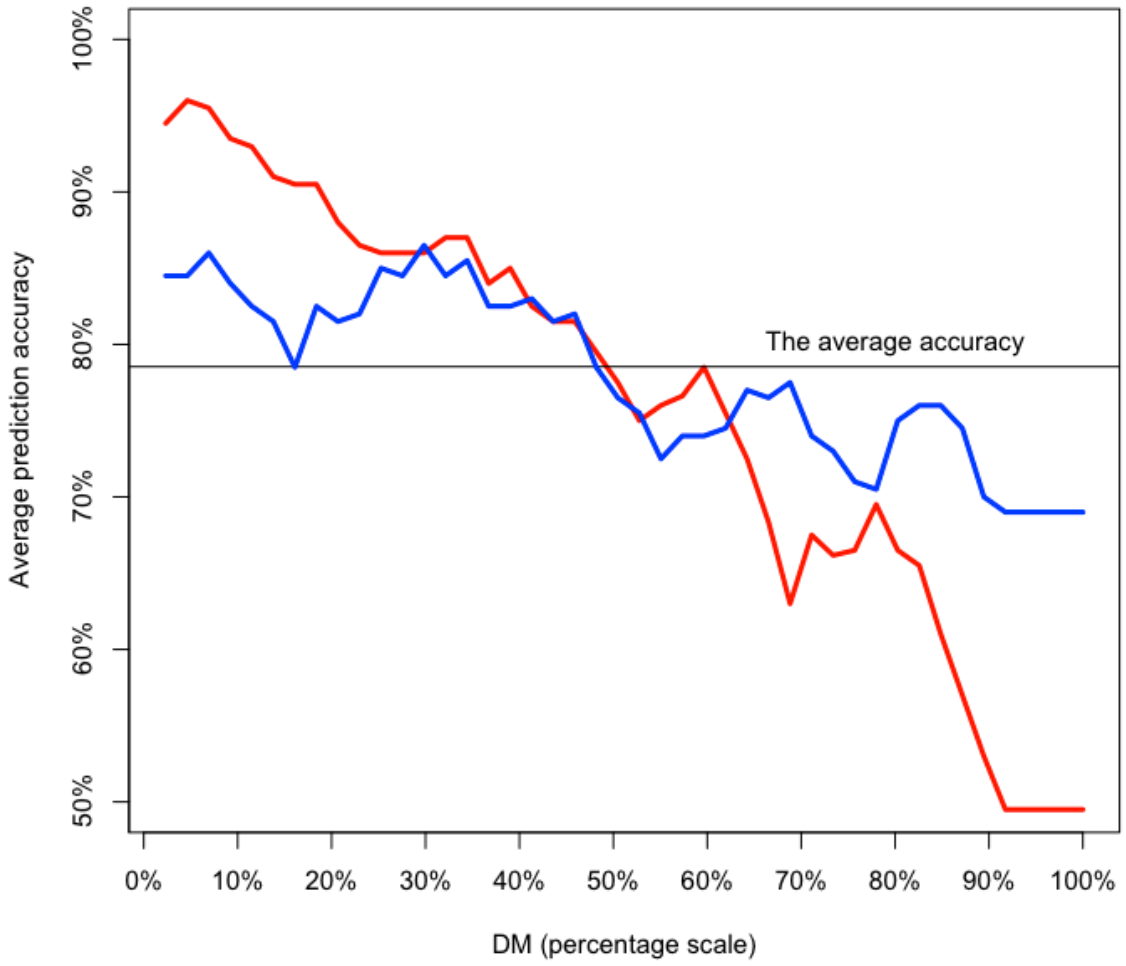
The criteria of a "good" measure:

On average, compounds with low DM should have higher accuracy than compounds with high DM.

A broader criteria:

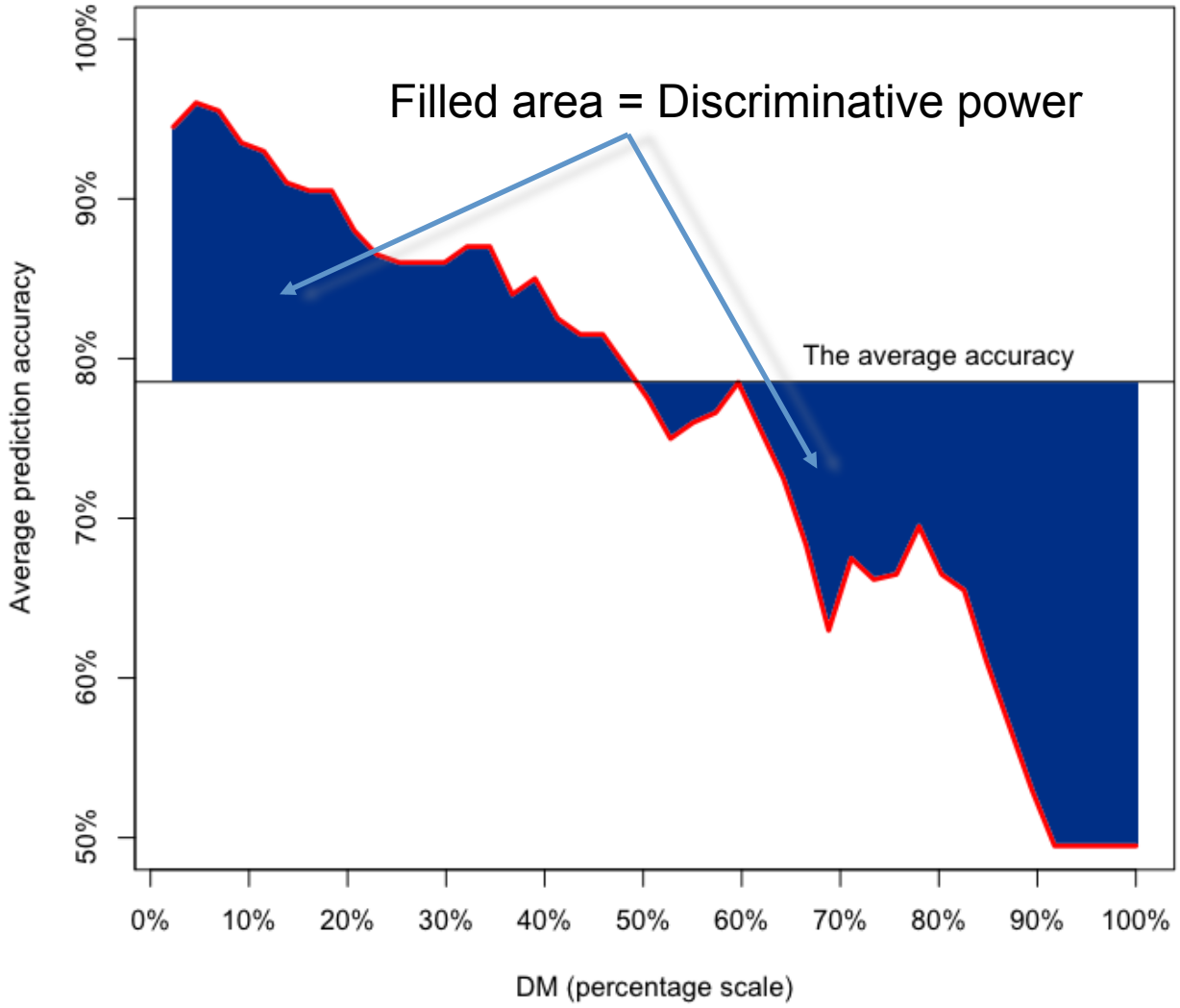
Compounds inside AD should have higher accuracy than compounds outside AD.

DM versus Accuracy



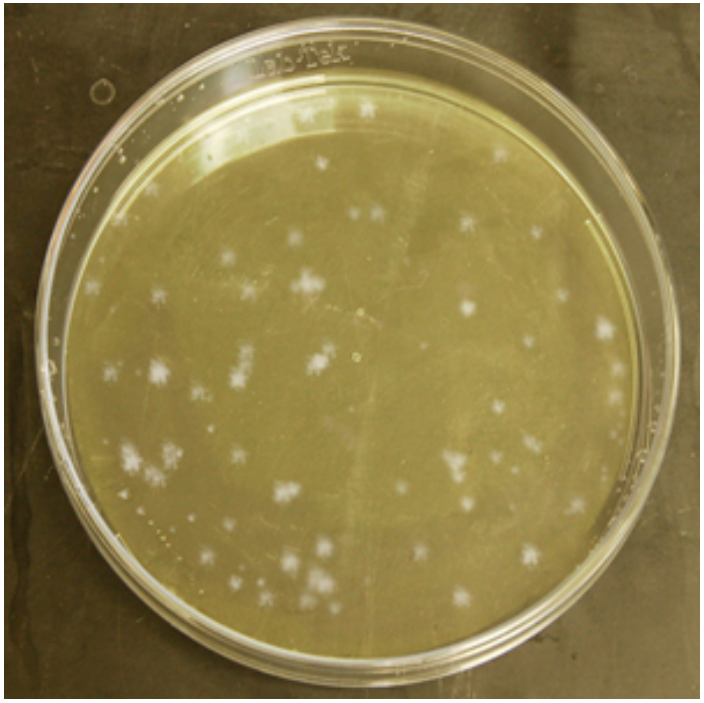
The red DM provides better discrimination of accurate and inaccurate predictions

DM versus Accuracy



Practical application

Ames test



A bioassay to identify mutagenic potential of chemical compounds

An indirect test for carcinogenicity

Can we substitute in-vitro measurements by accurate in-silico predictions?

Ames test

1 dataset

Total compounds measured	6,542
Mutagens	3,516
Non-mutagens	3,026

12 international groups

University of Insubria
Technical University of Berlin
Lanzhou University
Linnaeus University
Helmholz-Zentrum Munich
University of British Columbia
Louis Pasteur University
Moscow State University
Physico-Chemical Institute of the NAS of Ukraine
University Milano-Bicocca
University of North Carolina
Environmental Protection Agency, EPA

29 predictive models + consensus model

Model name	Descriptors used	Training method
CONS	-	-
EPA_2D_FDA	PCID	FDA
EPA_2D_NN	PCID	Neural networks
LNU_Drag_PLS	Dragon	PLS
MSU_FRAG_LR	Fragments	Linear regression
MSU_FRAG_SVM	Fragments	SVM
OCHEM_ESTATE_ANN	E-State indices	Associative neural networks
PCI_Drag_RF	Dragon	Random forest
PCI_SiRMS.Drag_RF	SiRMS	Random forest
PCI_SiRMS_RF	SiRMS	Random forest
TUB_3DDrag_RF	Dragon	Random forest
TUB_3DDrag_SVM	Dragon	SVM
UBC_ID_IWNN	Inductive descriptors	Weighted NN
UBC_ID_NN	Inductive descriptors	NN
UI_Drag_KNN	Dragon	KNN
UI_Drag_LDA	Dragon	LDA
ULP_ISIDA_NB	ISIDA Fragments	Naïve Bayes
ULP_ISIDA_SQS	ISIDA Fragments	SQS
ULP_ISIDA_SVM	ISIDA Fragments	SVM
ULP_ISIDA_VP	ISIDA Fragments	Voted Perceptron
ULZ_3DDrag_KNN	Dragon	KNN
ULZ_3DDrag_SVM	Dragon	SVM
UMB_Drag_DT	Dragon	Decision Tree
UNC_Drag_KNN	Dragon	KNN
UNC_Drag_RF	Dragon	Random forest
UNC_Drag_SVM	Dragon	SVM
UNC_SiRMS.Drag_RF	SiRMS+Dragon	Random Forest
UNC_SiRMS.Drag_SVM	SiRMS+Dragon	SVM

Ames test: accuracy

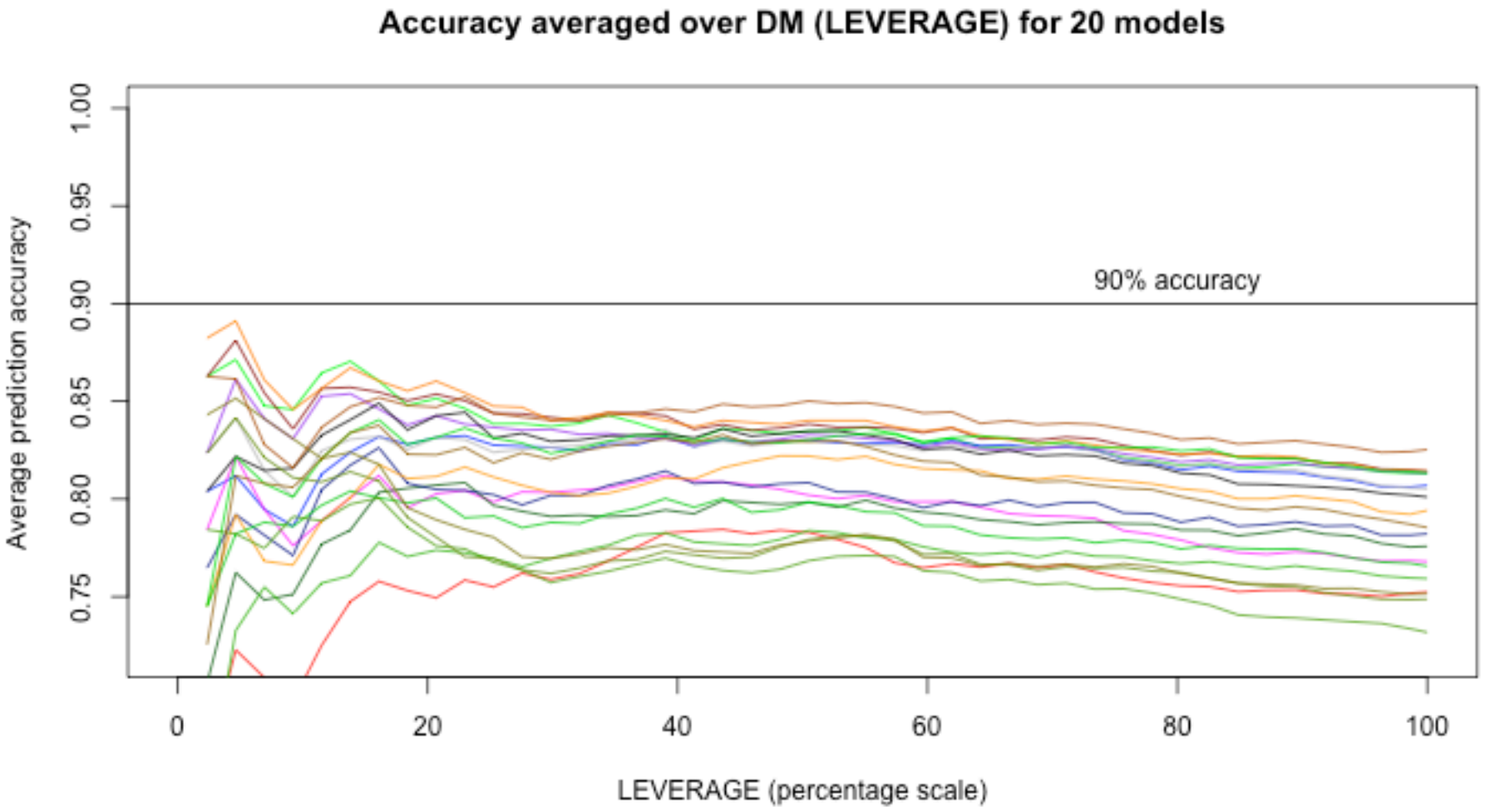
Accuracy of predictions

29 individual models	74%-82%
Consensus model	83%

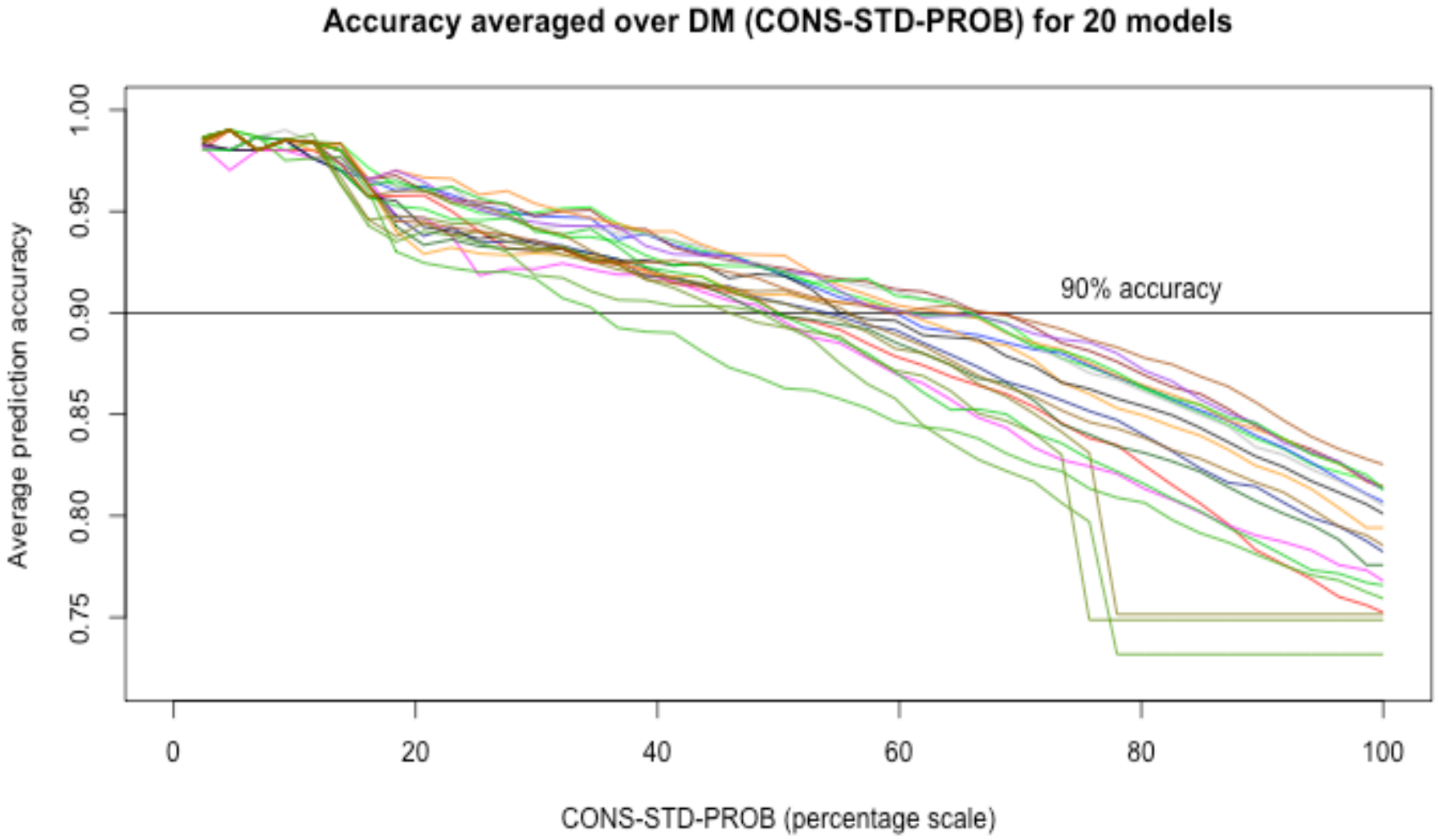
Accuracy of measurements

Inter-laboratory agreement	90%
Intra-laboratory agreement	95%

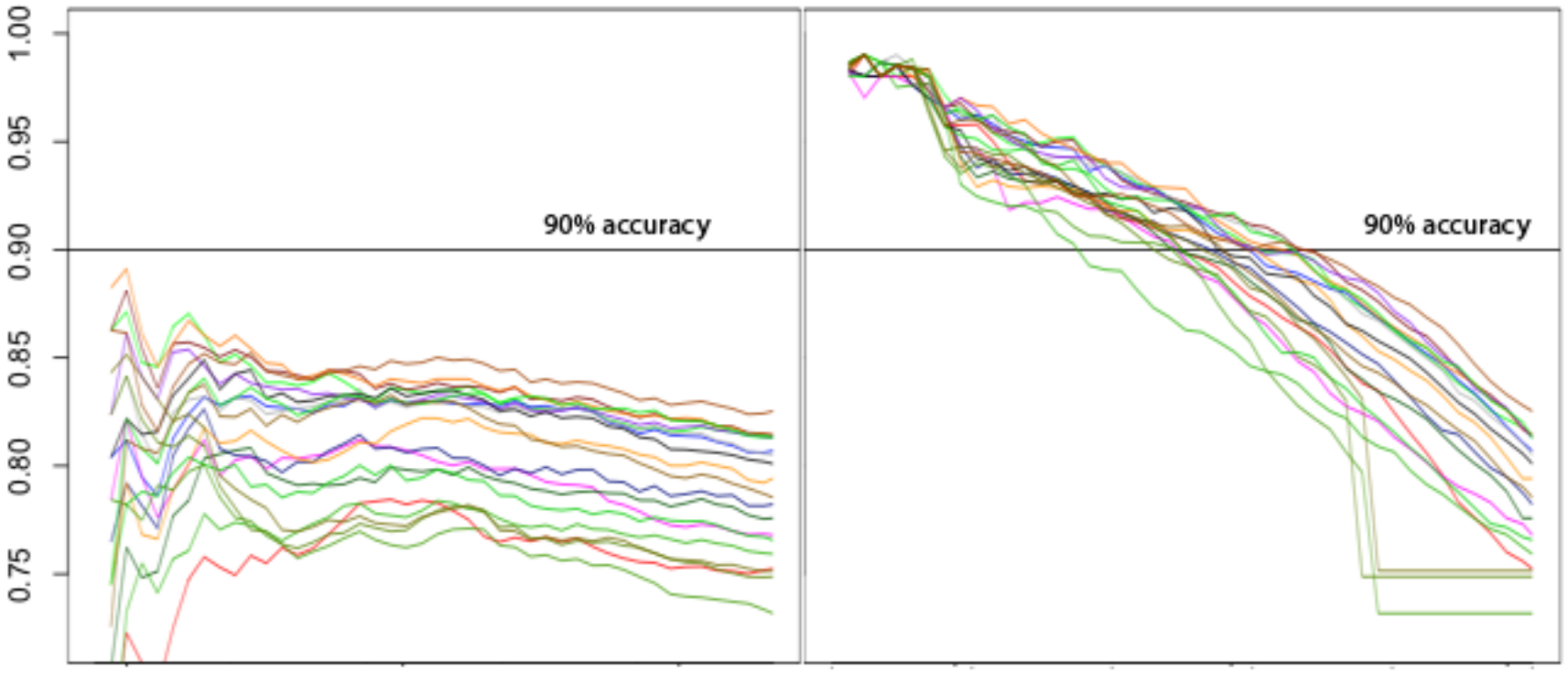
Ames test: DM vs accuracy



Ames test: DM vs accuracy



Ames test: DM comparison

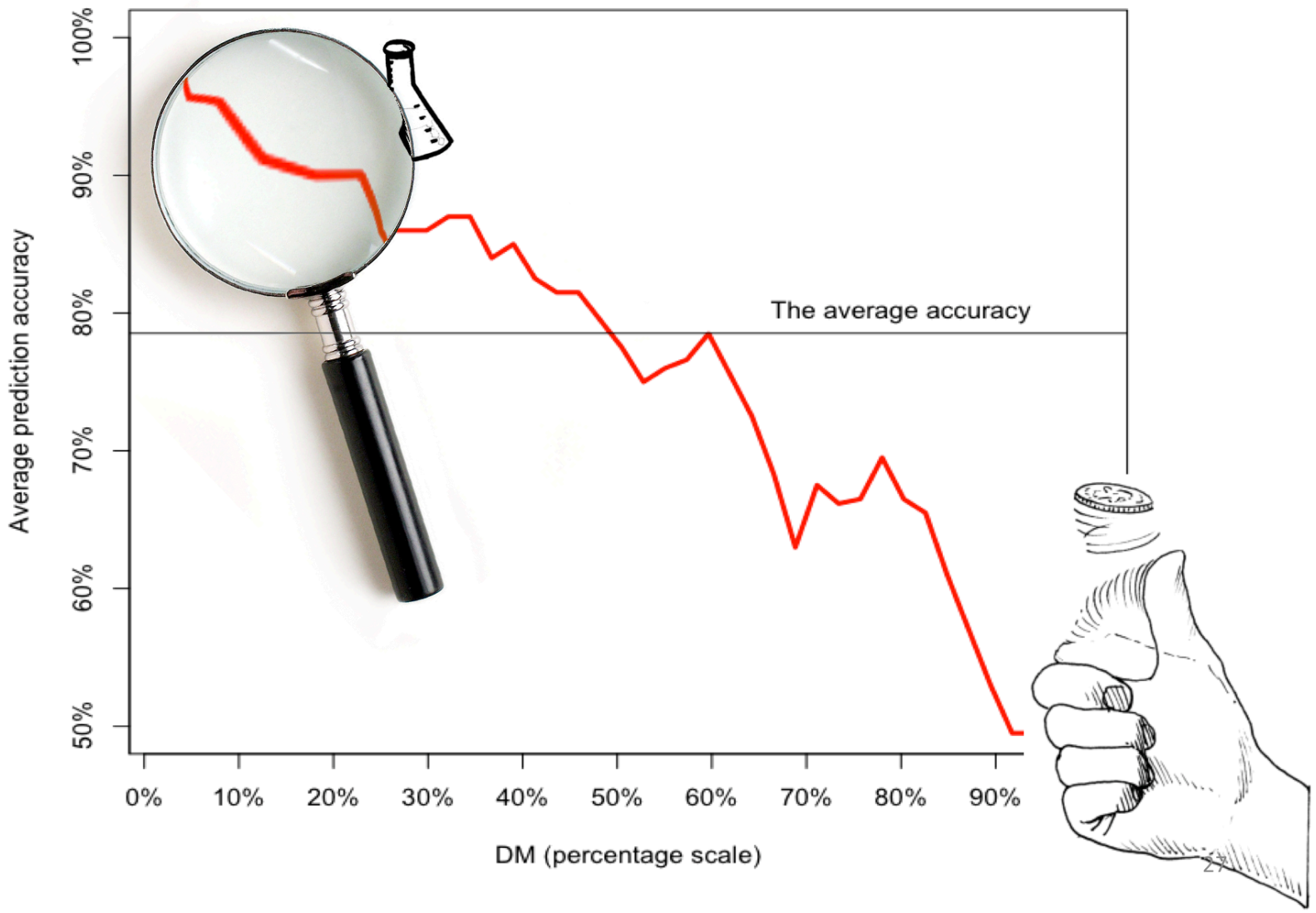


Dissimilarity
(Leverage)

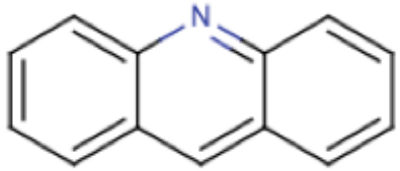
Disagreement
(Consensus voting)

Similar structures does not always mean similar activities
("activity cliffs" phenomemon)

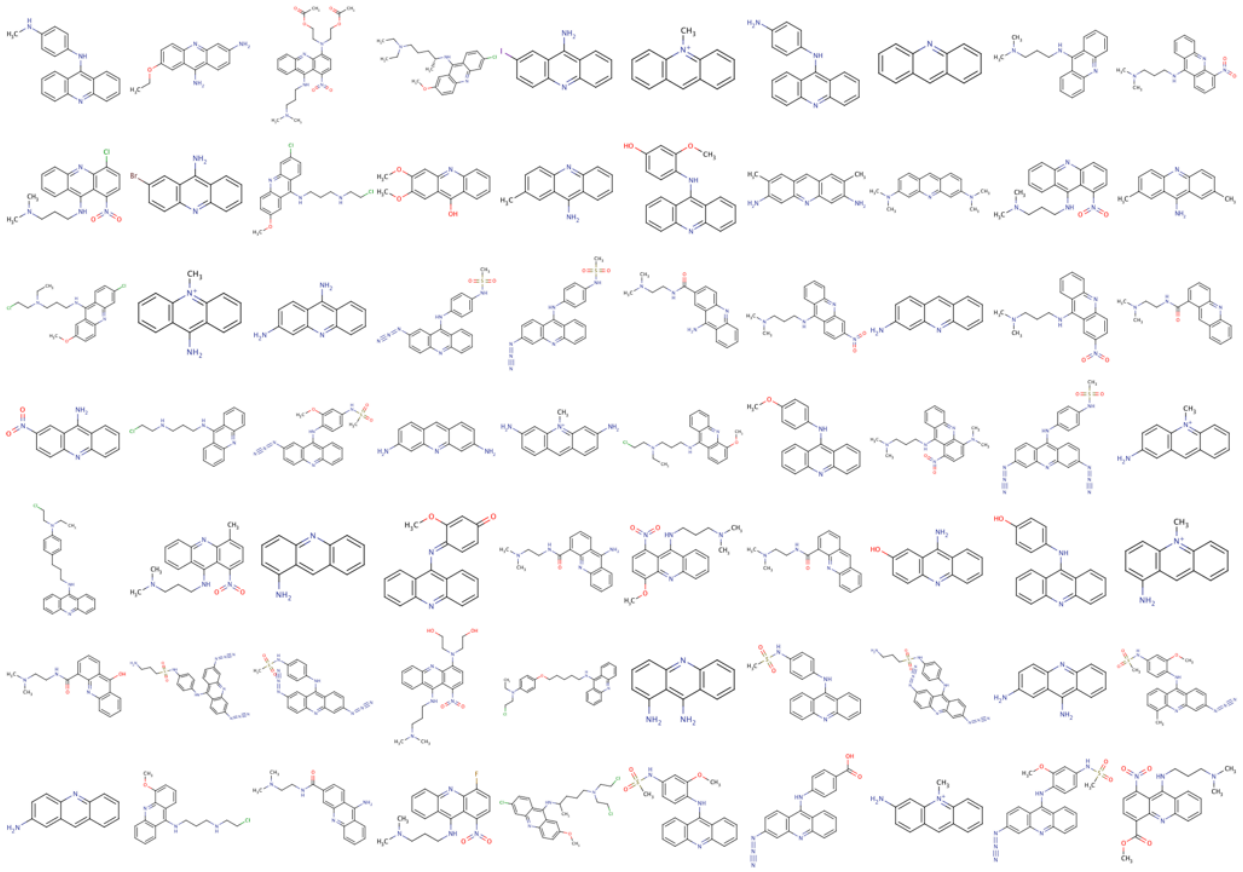
High discriminative power



A look inside the AD

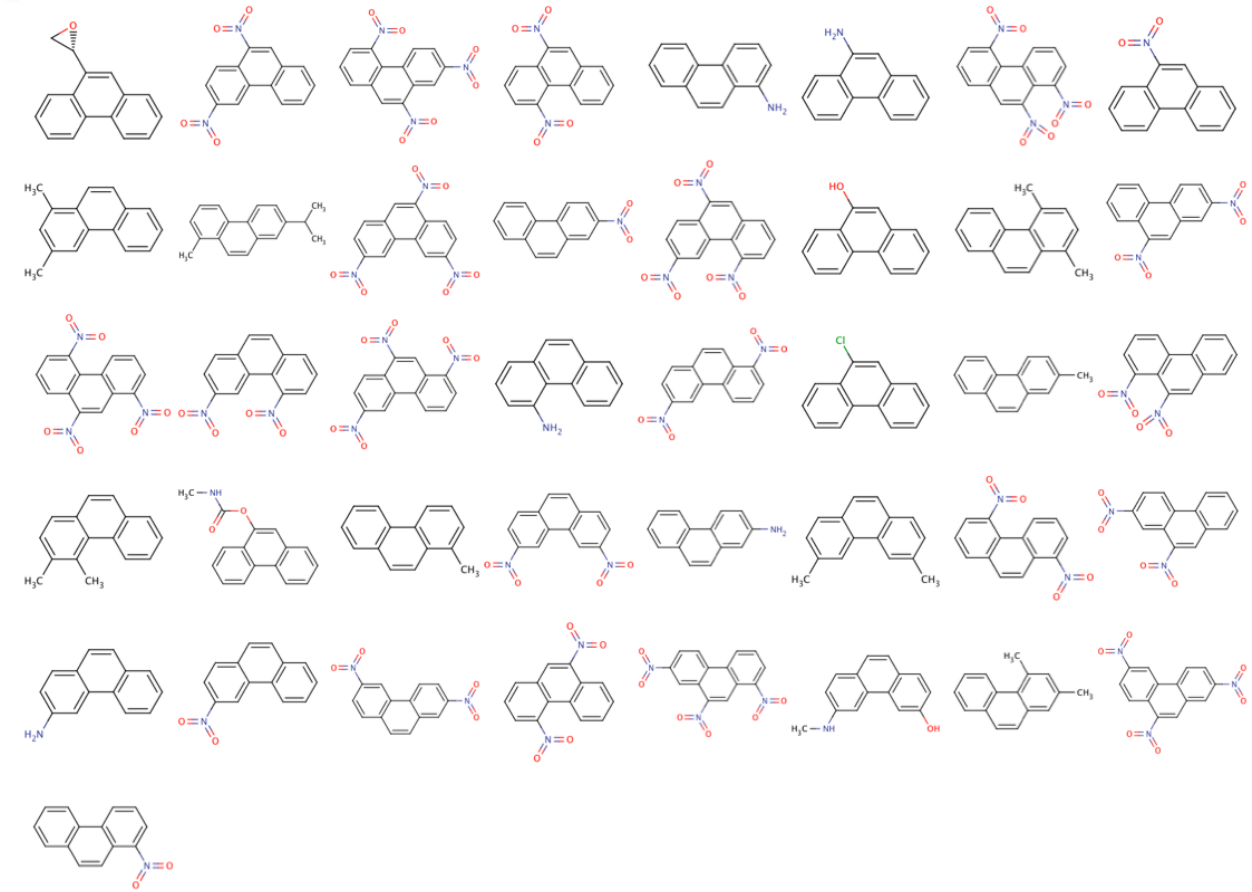


Acridines:
69 out of 74 are mutagens

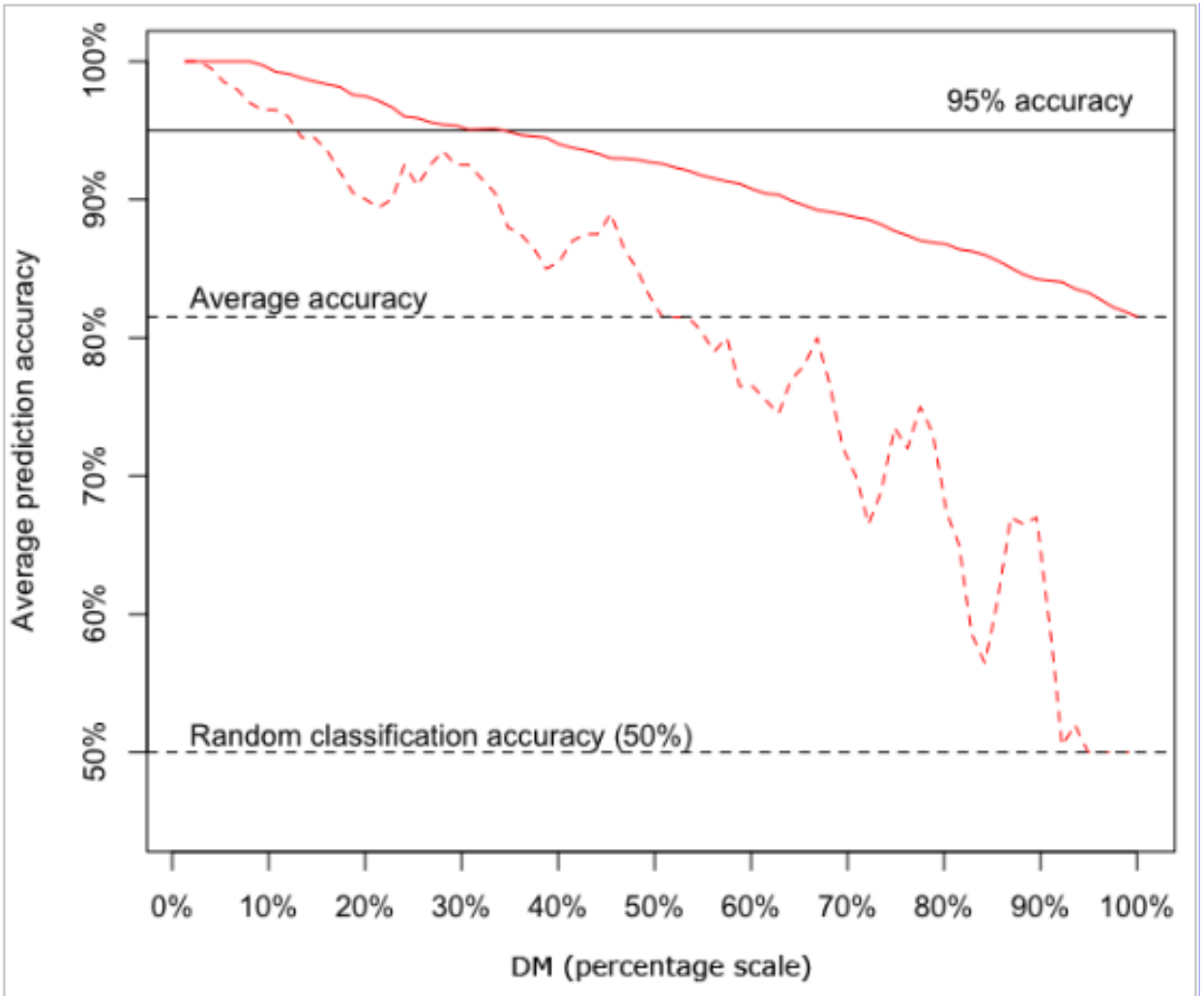


A look inside the AD

Phenathrenes:
38 out of 41 are mutagens



Another example: CYP inhibition

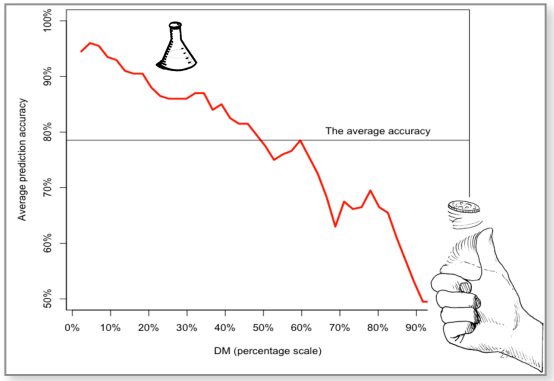


Average accuracy:
81% correct cl-s

Accuracy spread:
50%-100%

Process summary

1. Build and validate an ensemble of QSAR models
2. Calculate several DMs
3. Average accuracy over DMs



4. Evaluate the discriminative power
5. Interpret the AD

Interpretation in terms of:
- subfragments
- molecular properties

Technical implementation

The described approaches are conceptually simple, but reproducing them requires significant technical effort.

- Manage processing of molecular structures
- Run descriptor calculations
- Run machine learning methods
- Develop hundreds of models
- Analyze DMs
- Build accuracy averaging charts
- Store all the models on disk for further application
- Manage format inter-conversions
- etc…

All the steps are automated at QSPR-Thesaurus and OCHEM platforms: http://qspr-thesaurus.eu and http://ochem.eu

Including:

•Calculation of many types of DMs

•Integration of OpenTox DMs (AMBIT)

Summary and Perspectives

Summary:

- DM concept establishes a universal framework for AD assessment
- Structural AD complements it with a broad and interpretable definition

Perspectives:

- The process for AD assessment and validation should become a commonly used practice
- The process should be technically simple and intuitive for end-users: chemists, biologists, regulators

Good models are useful
only if their limitations are known.

Acknowledgements

- Dr. Igor Tetko
- Prof. Paola Gramatica
- Prof. Hans-Werner Mewes
- CADASTER team
- eADMET team

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Thank you

Thank you