Improved Applicability Domain Determination of QSAR Models Using Local Mapping: RDN

Natalia Aniceto, Alex Freitas, Andreas Bender, Taravat Ghafourian

PhD student • University of Kent

University of Kent

medway school of pharmacy

University of Sussex

University of Cambridge

BACKGROUND

"Applicability Domain"

Establishing boundaries for **prediction reliability** is arguably as important as demonstrating good predictive performance.

cumulative result of **data noise** and **sparseness**

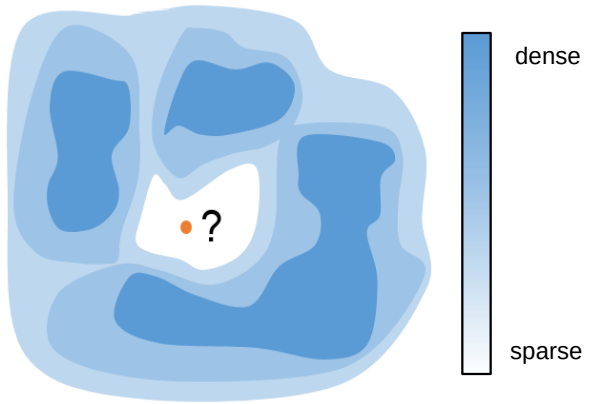
AD methods proposed so far typically address the data as a whole, often focusing on a **single aspect of data** (i.e. similarity to training set, descriptor span, data density, etc).

Combining **different measures** that address data **locally** has shown improved AD characterization. Sahigara et al.[1] and Sheridan[2]

STD + Similarity + Predictions

In order to successfully characterize a model's AD:

Hypothesis: $AD = f(\textbf{local density} \ \& \ \textbf{local reliability})$



1. Sahigara et al. Journal of Cheminformatics 2013, 5:27
2. Sheridan. J. Chem. Inf. Model. 2012, 52, 814−823

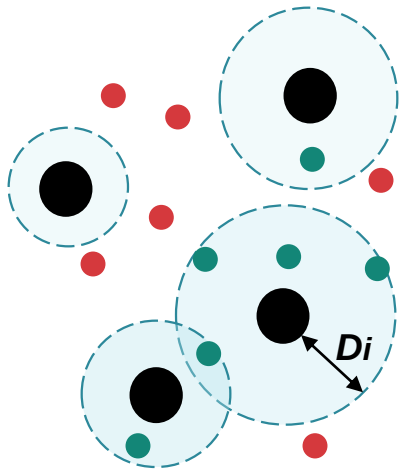
METHODS

$$AD = f\ (\text{local density}\ \&\ \text{local reliability})$$

1 2

1 Local Density

Sahigara et al.[1]



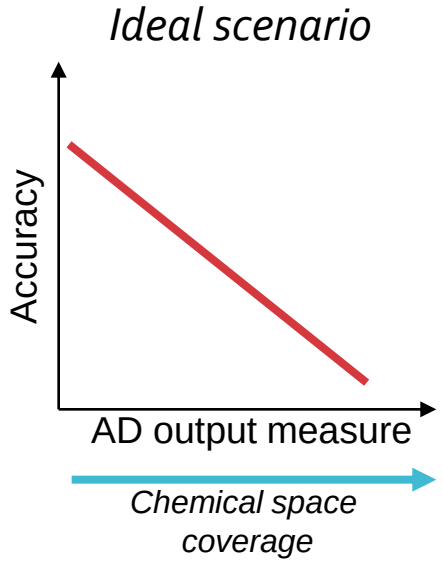
Density Neighbourhood (dk-NN)

Rationale:
Each training instance provides coverage of chemical space according to densely populated is its local vicinity.

A radius of coverage is placed on each instance from the average Euclidean Distance (ED) to its neighbours within the average ED to the k-th NN.

↑**k :** ↑**span of coverage**

scan through the chemical space



However high local ***density*** ***does not imply*** *high* ***reliability***

METHODS

$$AD = f\ (\text{local density}\ \&\ \text{local reliability})$$

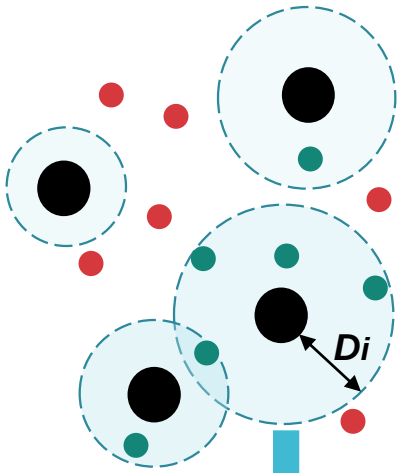
1 2

2 Local Reliability

precision & bias

Deviation within an ensemble of models (STD)
(Tetko et al)

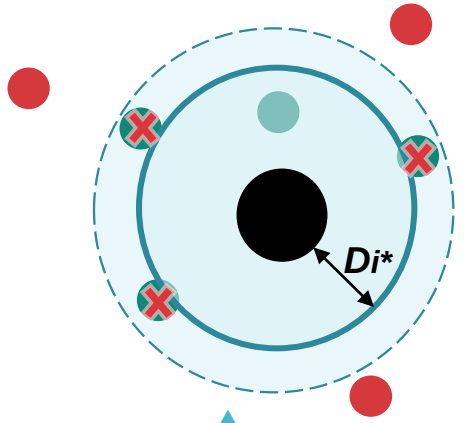
Agreement with observed response



$$W_i = \left(1 - \sqrt{\frac{\sum_{m=1}^{M}\left(\widehat{y_{i,m}} - \bar{y}_i\right)^2}{M-1}}\right) \times \frac{|Y_i \cap \widehat{Y}_i|}{M}$$

STD agreement

$$D_i^* = D_i \times W_i$$



↑ agreement
↑ 1 – STD

↑W

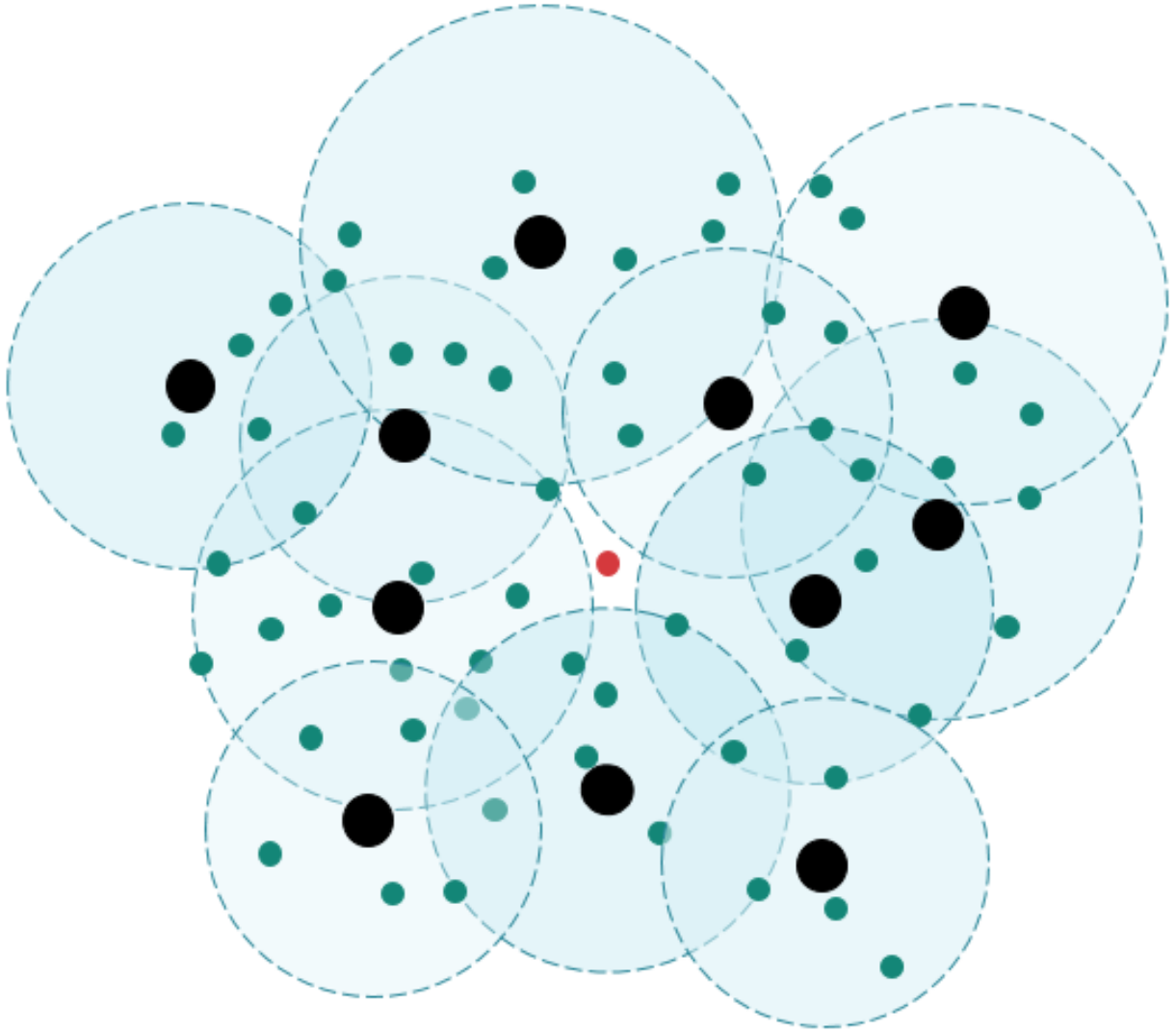
Di is less penalized

Higher reliability = larger coverage

Tetko et al 2008 *J. Chem. Inf. Model.* **2008,** *48,* 1733–1746

METHODS

Scan through the chemical space



Training set

External set

out in

k = 7

By **keeping track of new instances** entering the AD, and **updating in-AD-Accuracy**, we build a **map of reliability** across the model's space

METHODS

Part #1

- Explore the capabilities of the **Reliability-Density Neighbourhood (RDN)** algorithm
- Characterize **Mispredictions**

(❶ “working” dataset)

Part #2

Test the **RDN** algorithm on ❷ benchmark datasets

- ❑ Ames mutagenicity test
- ❑ CYP450 inhibition test *(1A2)*

Class = { + , - }

Part #3

Compare **RDN** vs ❸ other AD methods

METHODS Part #1

❶ Explore the capabilities of RDN

P-gp dataset
Class = {S, NS}

Training set
N = 659

Validation set
N = 194

Test set
N = 195

Build a decision tree model

Ensemble model
(10-fold bagging)

- **STD**
- **Agreement**

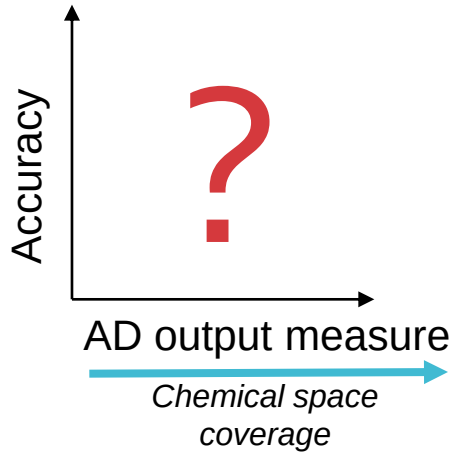
- **Predictions**

Build RDN

Test RDN

Test Hypothesis

$$\mathrm{AD} = \mathbf{f}\left(\textbf{local density} \;\&\; \textbf{local reliability}\right)$$



❷ Characterizing Mispredictions

Global Density?

*Kernel Density Estimation (**KDE**)*

Descriptor span?

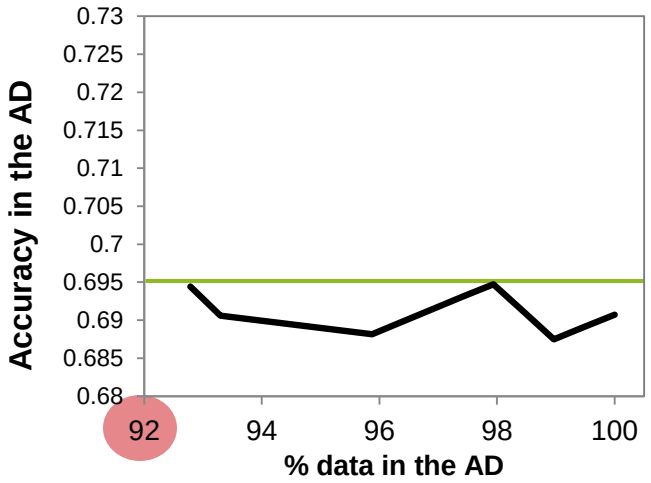
Decision tree descriptor span

RESULTS Part #1

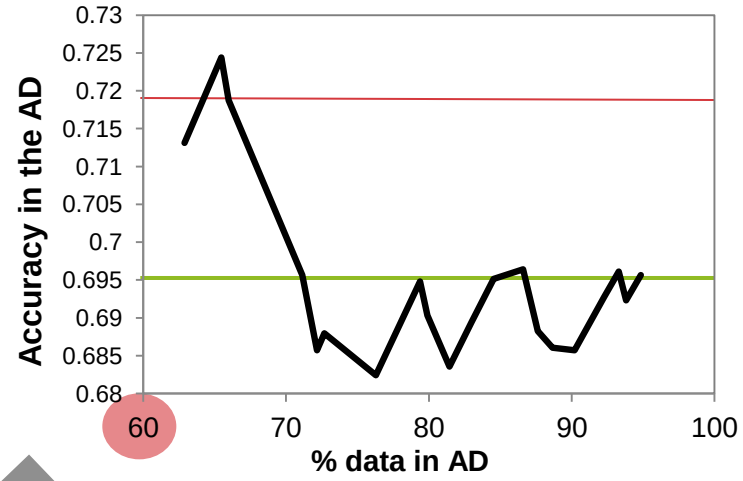
P-gp

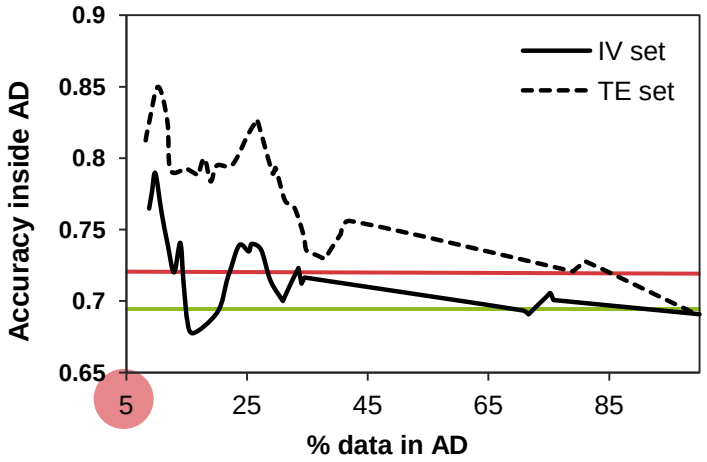
❶ Explore the capabilities of RDN

Original dk-NN

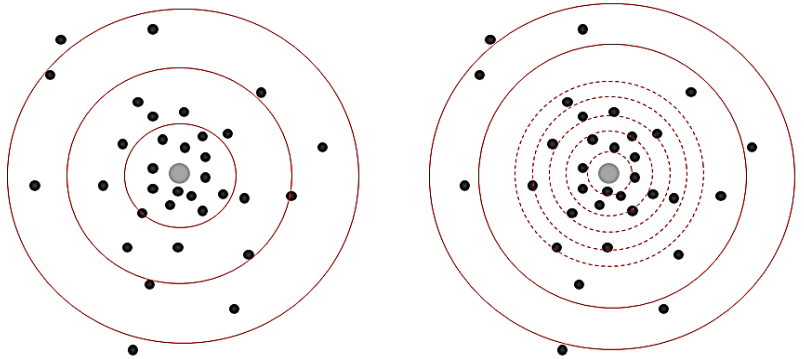


RDN





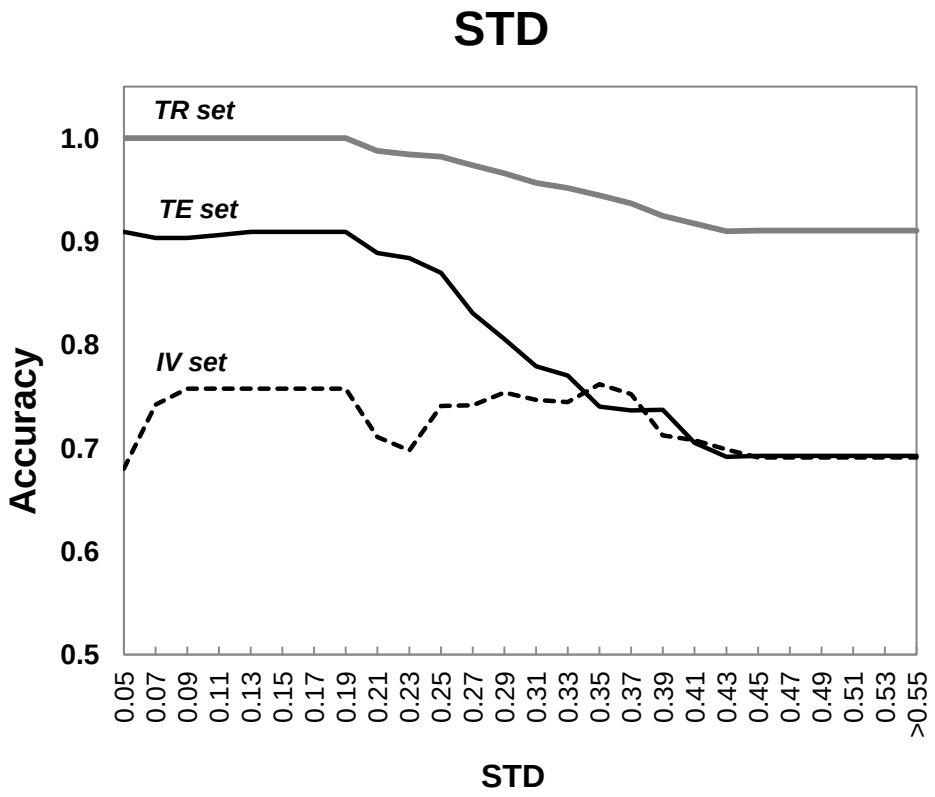
Shrink distances to **1/3** in the beginning → **½ Di** → **Di**

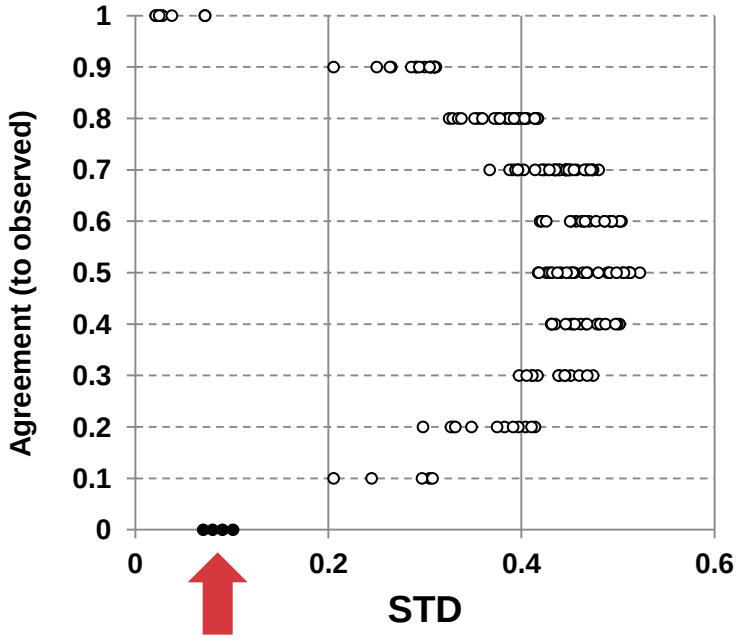


RESULTS Part #1

P-gp

Role of bias-precision correction



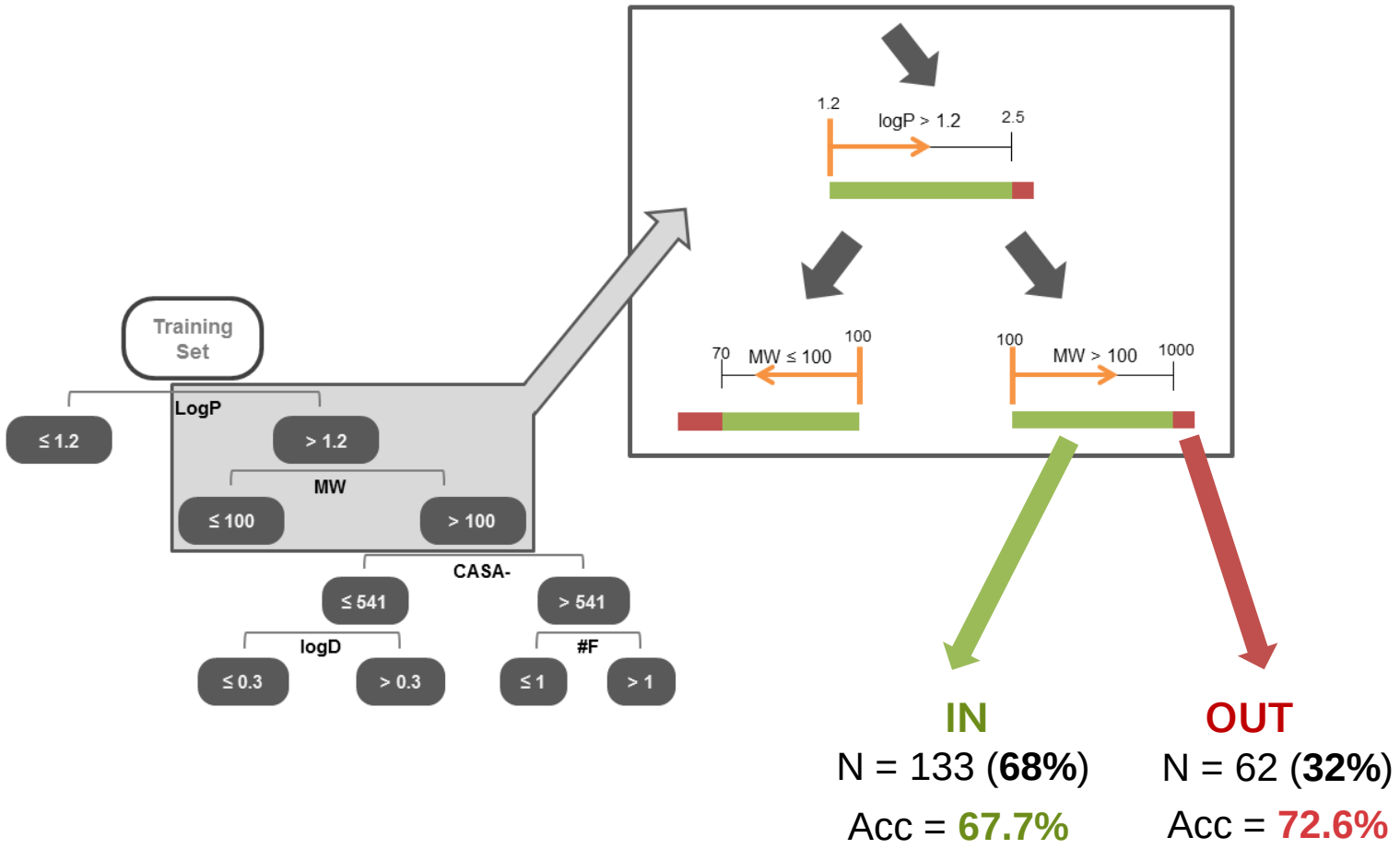


RESULTS Part #1

P-gp

❷ Diagnosing **mispredictions**

Descriptor span ?



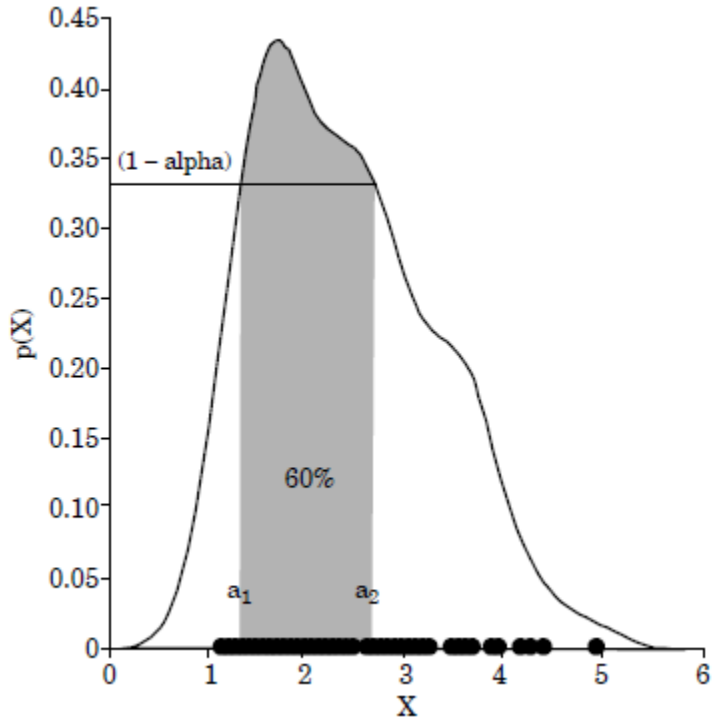
RESULTS Part #1

P-gp

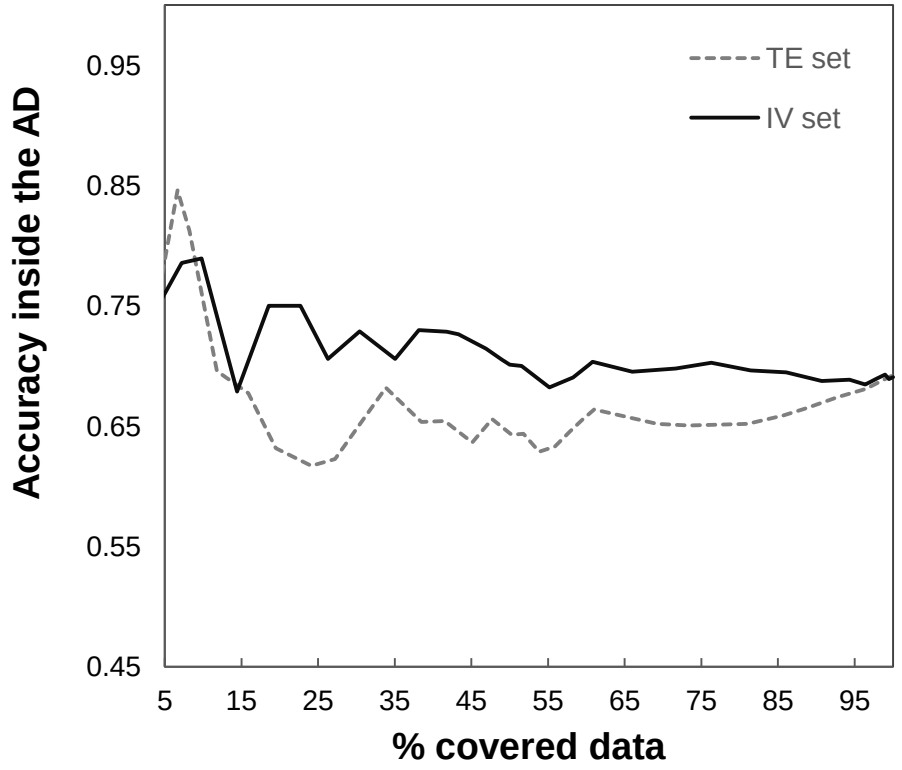
❷ Diagnosing **mispredictions**

Density in feature space ?

Kernel Density Estimation **(KDE)**



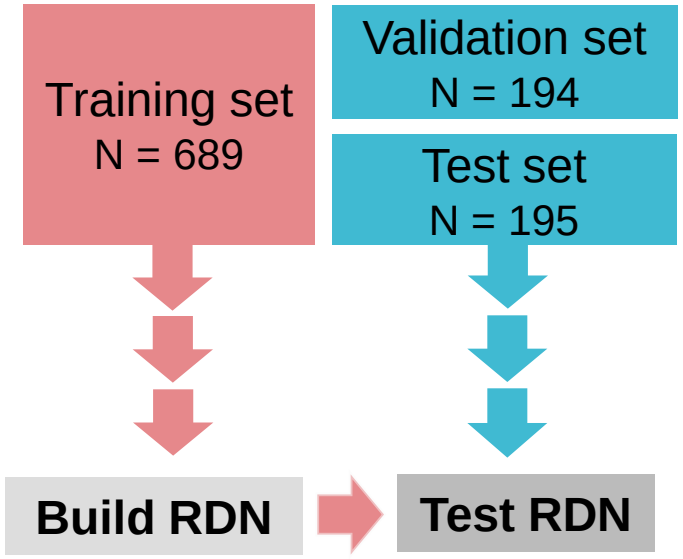
PC1



METHODS Part #2

Test the **RDN** algorithm on ❷ benchmark datasets

P-gp dataset
Class = {S, NS}



Online chemical database
with modeling environment

	training	Test #1	Test #2
Ames dataset	4358	N = 1089	+ 1090
CYP450 dataset	3743	N = 1870	+ 1870

Predictions, **STD** & **Agreement** from **OChem**

Build RDN → **Test RDN**

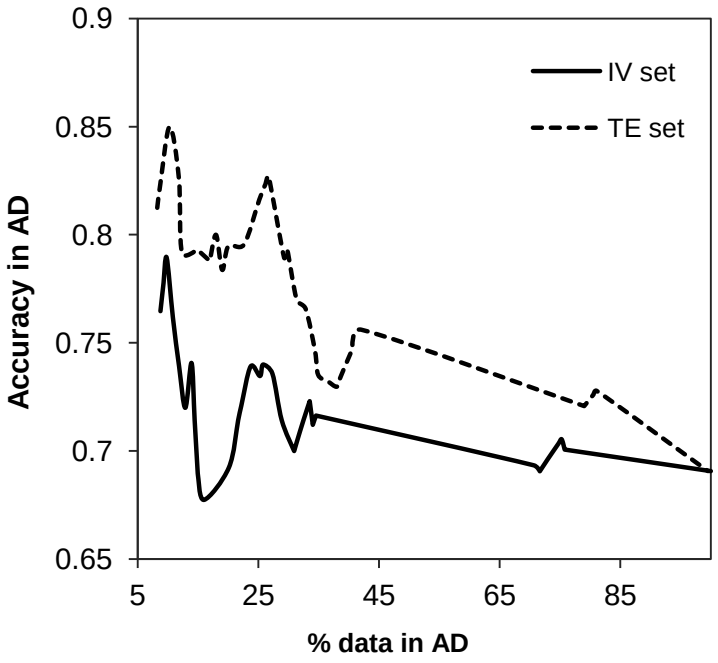
Test Hypothesis

$$\mathrm{AD} = \mathbf{f}\left(\textbf{local density} \;\&\; \textbf{local reliability}\right)$$

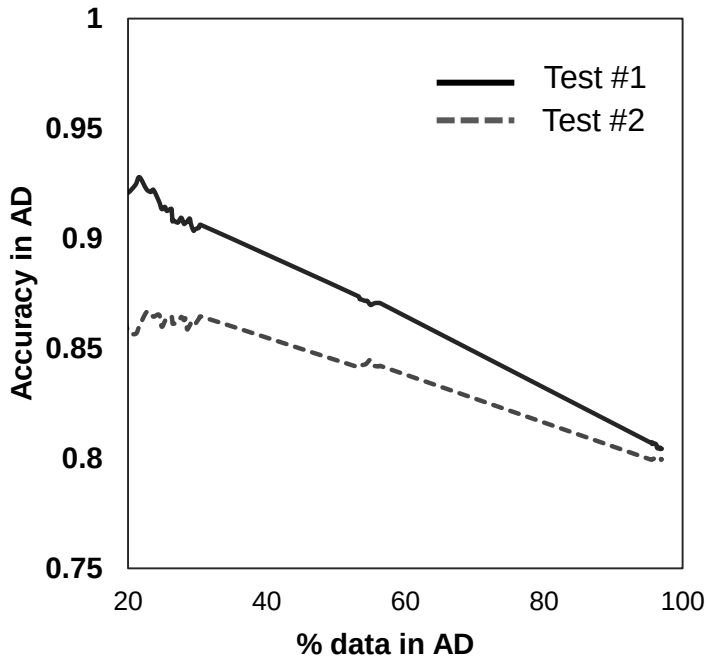
RESULTS Part #2

Test the **RDN** algorithm on ❷ benchmark datasets

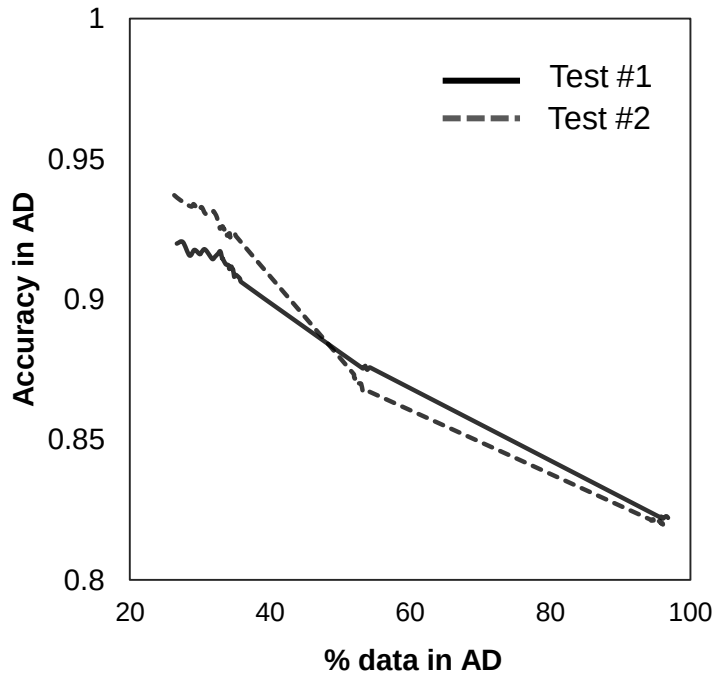
Pgp model



Ames model

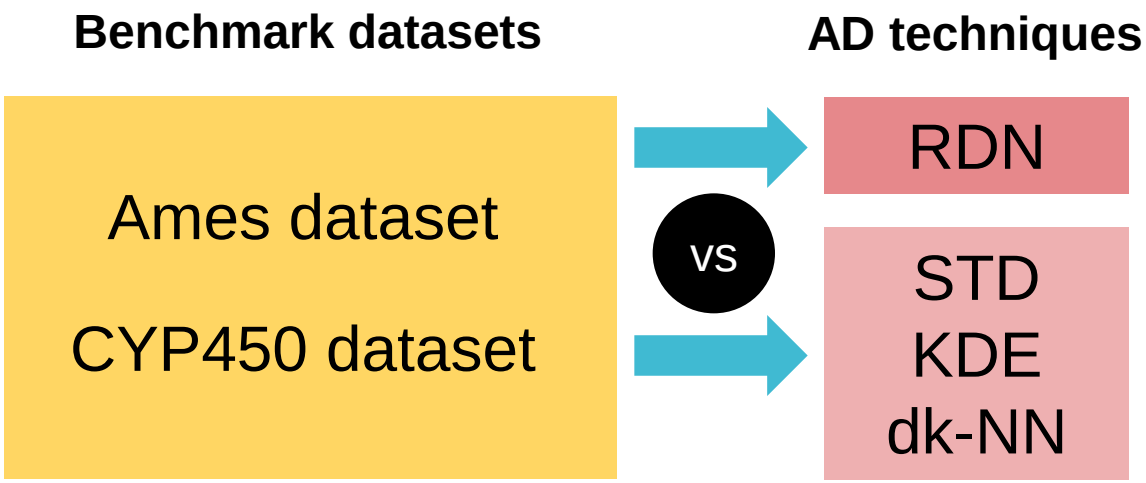


CYP450 model



METHODS Part #3

Compare **Reliability-Density Neighbourhood (RDN)** vs ❸ other AD methods



Test Hypothesis

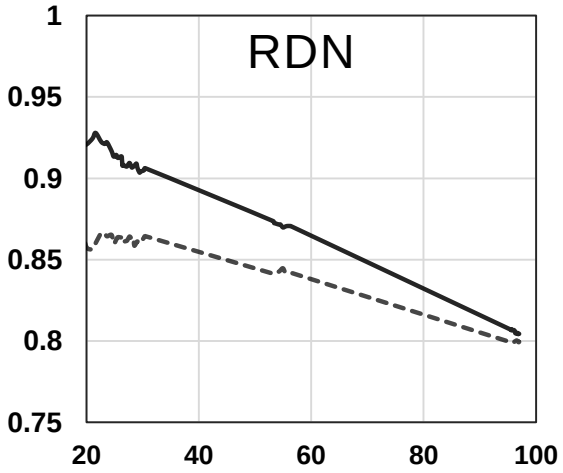
$$\mathrm{AD} = \mathbf{f}\left(\textbf{local density} \;\&\; \textbf{local reliability}\right)$$

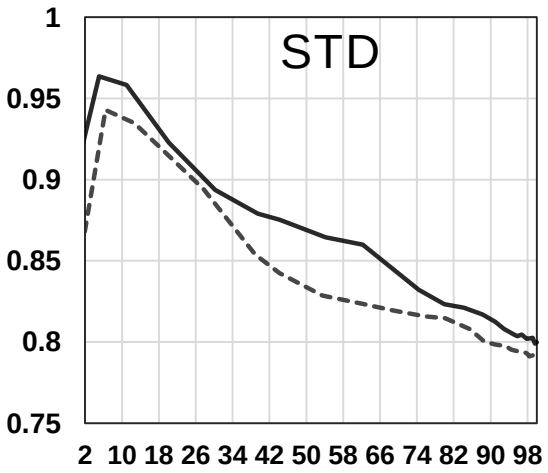
RESULTS Part #3

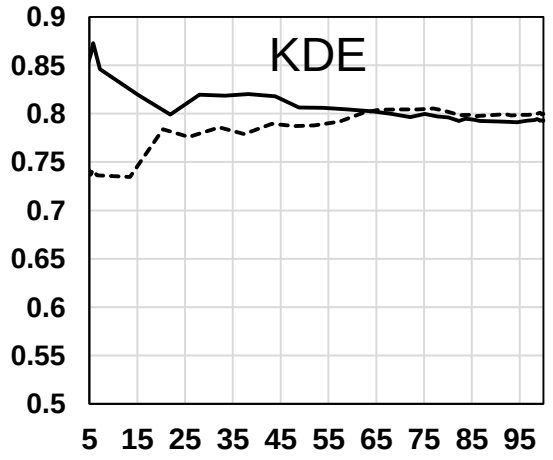
Test #1
Test #2

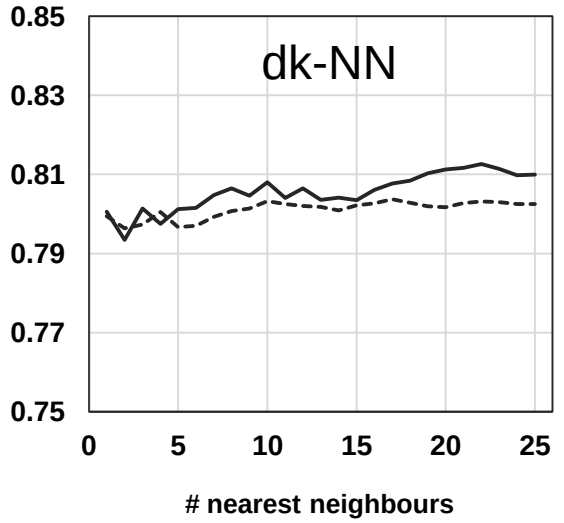
Ames model

Accuracy vs. % coverage (implicitly, **distance-to-model**)

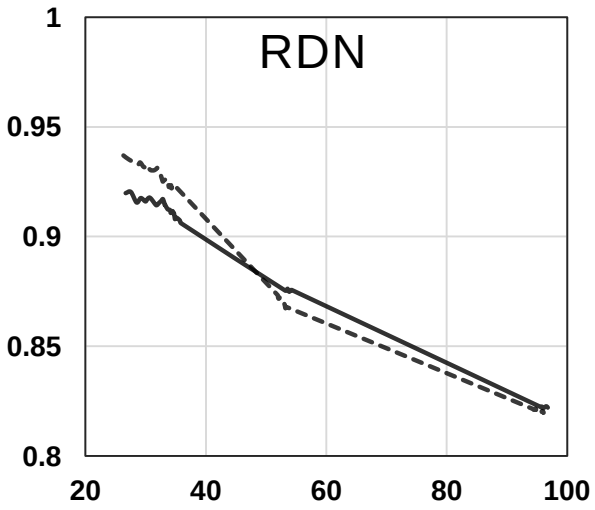


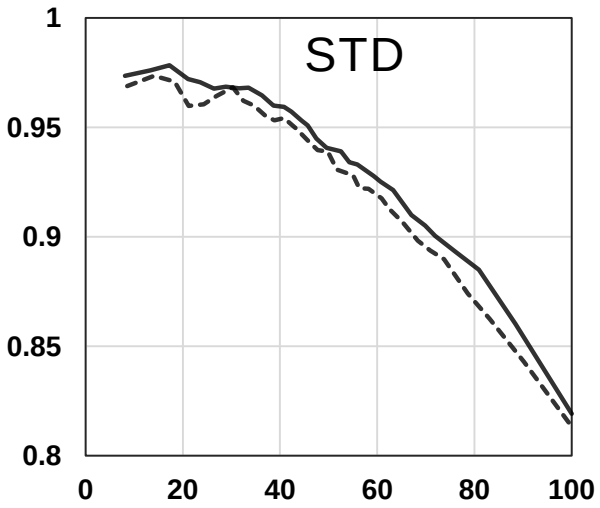


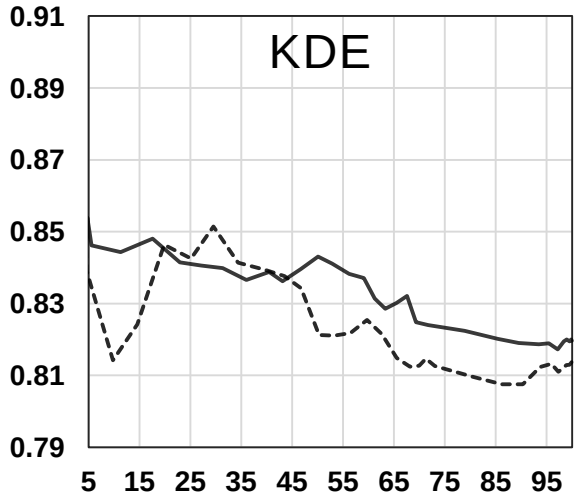


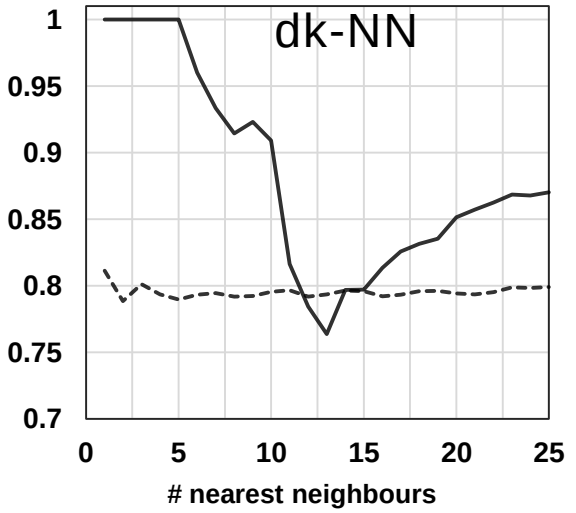


CYP450 model









CONCLUSIONS

- Local density corrected for local reliability (Precision + Bias) is able to successfuly sort new instances according to their predictive performance through the definition of map that identifies regions according to their probability to contain mispredictions

- The RDN method performs robustly in new unseen data (two external datasets show similar profiles of accuracy accross chemical space)

- For the optimal establishment of the AD of a QSAR model:

  RDN + STD → case-by-case selection of the best candidate.

THANK YOU FOR YOUR ATTENTION!

Improved Applicability Domain Determination of QSAR Models Using Local Mapping: **RDN**

Natalia Aniceto, Alex Freitas, Andreas Bender, Taravat Ghafourian

PhD student ● University of Kent

University of Kent

medway school of pharmacy

UNIVERSITY OF SUSSEX

UNIVERSITY OF CAMBRIDGE