Pharma Algorithms
WWW.AP-ALGORITHMS.COM

Assessing and Improving the Reliability of *In Silico* Property Predictions by Incorporating In-house Data

Pranas Japertas

Outline

- Why do *in silico* models fail?
- Knowing when *in silico* models fail – assessing Model Applicability Domain (Reliability Index)
- Relating RI to accuracy of predictions
- Improving models with in-house data
- Case studies of model improvement with in-house data
- Connecting measurement and prediction

Why do *in silico* models fail?

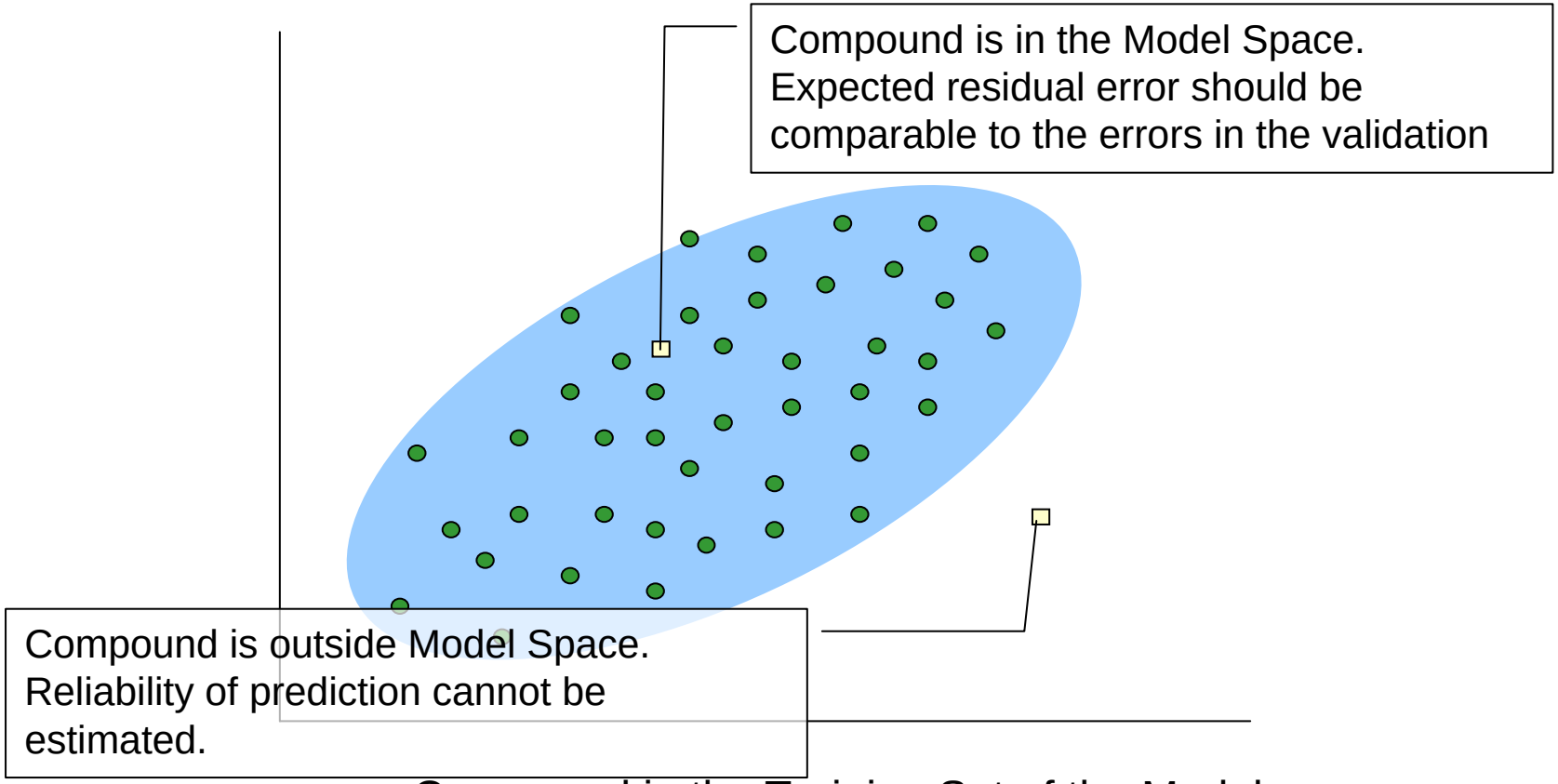
Why do *in silico* models fail?

- Irrelevant descriptors?
- Statistical techniques not sophisticated enough?
- Limited diversity of the training set?
- Improper usage of statistical tools?
- Poor data quality of the training set?

Any model, no matter which descriptors or statistical methods were used in its development cannot be better than the data it is based on.

Every empirical model works only in certain chemical space, where the compounds from the training set are located – boundaries of Model Applicability Domain

Model Space (Applicability Domain)

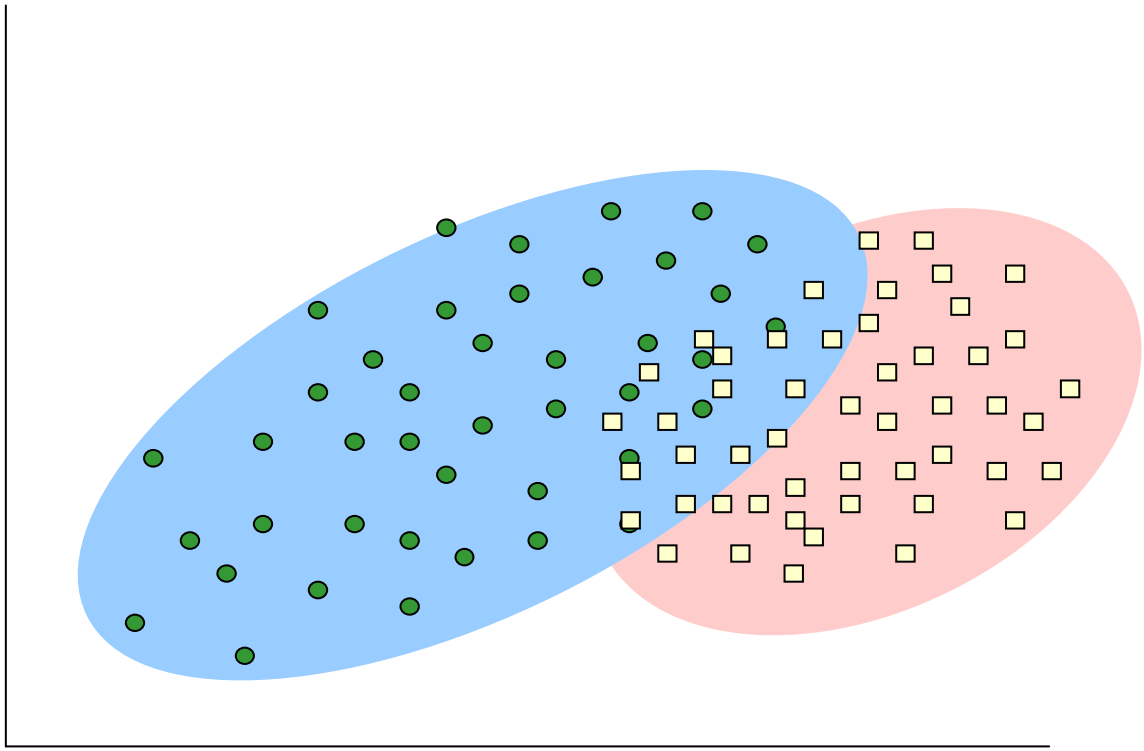


– Compound in the Training Set of the Model

– Model Space (Applicability Domain)

– New Compound, we are making prediction for

Model Space and Chemical Space of proprietary Compounds



– Compound in the Training Set of the Model

– Model Space (Applicability Domain)

– New Compound, we are making prediction for

– Chemical Space of in-house compounds

The very FIRST question *in silico* model should answer:

- Is a compound in the Model Applicability Domain (Model Space)? Can we trust this prediction?

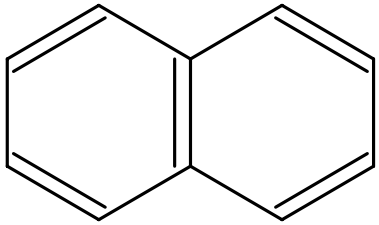
What is the predicted value for property X is only the second question.

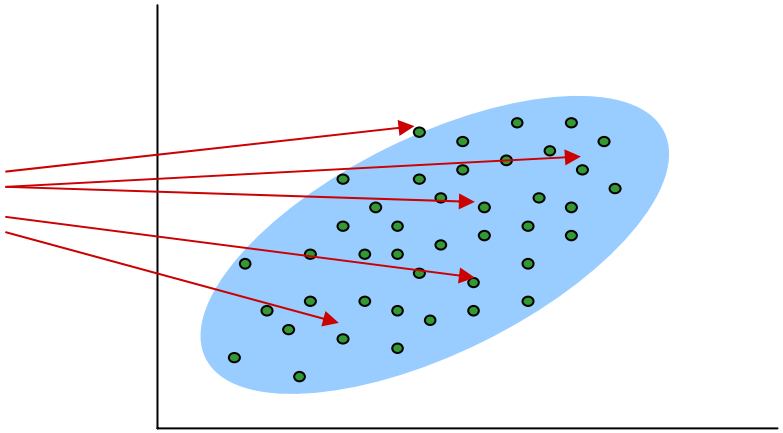
Assessing Model Applicability Domain. Reliability Index (RI) as a measure of the quality of a prediction

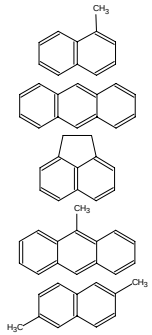
Pathway to the Reliability Index

- Similarity Index
- Data-Model Consistency Index
- **Reliability Index**

The First Step – Finding most similar compounds





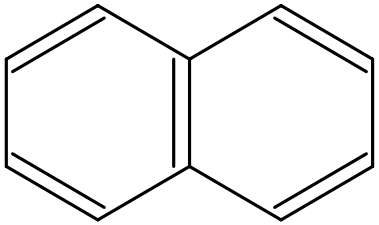


Compound we are making prediction for

Compounds in the training set

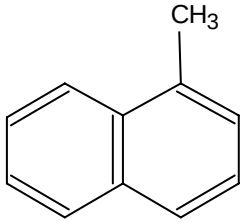
The most similar compounds in the training set

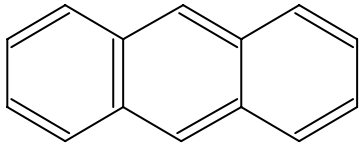
Calculating Similarity Index

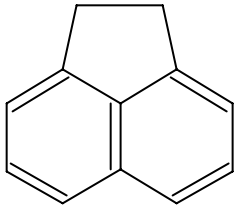


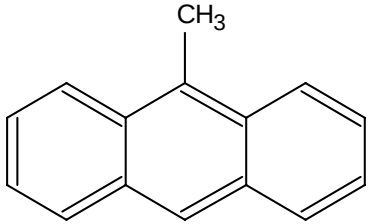
$SI = 0.89$

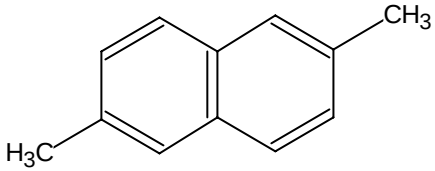
n most similar compounds from the database:











SI_1=0.97 SI_2=0.92 SI_3=0.88 SI_4=0.87 SI_5=0.80

Similarity Index (SI) is calculated as weighted average pair wise similarity to n most similar compounds in the training set:

$$SI = \sum_{i=1}^{n} a^{i-1} \cdot SI_i \Big/ \sum_{i=1}^{n} a^{i-1}$$

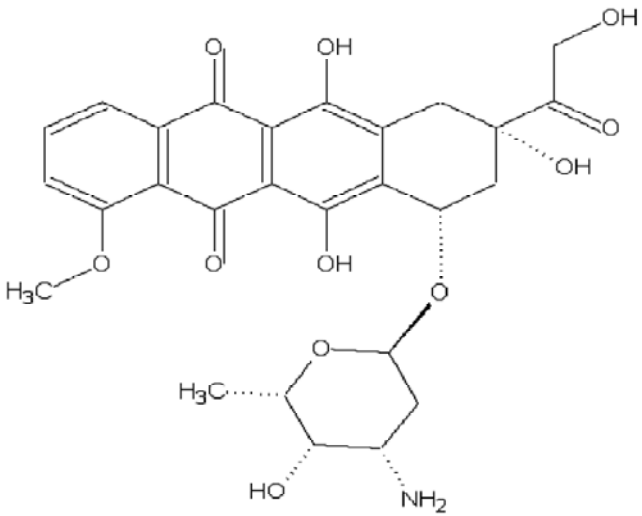
- Similarity Index obtains values in the range from 0 (nothing similar exists) to 1 (n completely similar compounds exist), making it easily understandable and usable

- Similarity Index is a simple but efficient criterion identifying compounds that DO NOT belong to Model Applicability Domain

- But having similar compounds does not necessary means that predictions will accurate and reliable

Let's consider the following example with prediction of Acute Toxicity (LD50) values for the doxorubicin…

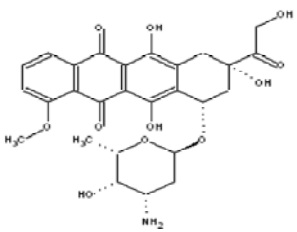
Doxorubicin, Rat/IP



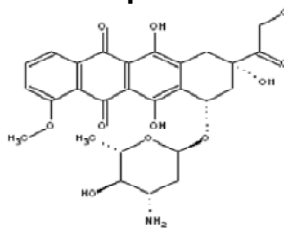
Rat, Intraperitoneal administration:

Estimated LD_{50} = 13 mg/kg
RI = **0.71**

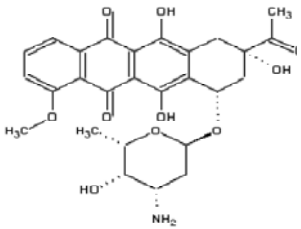
Five most similar compounds with experimental values from the library:



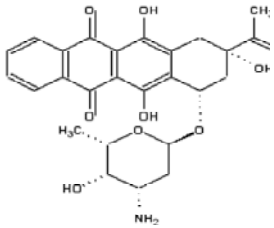
LD_{50} = 10mg/kg
SI = 1.00



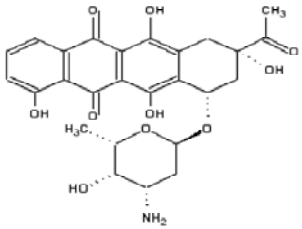
LD_{50} = 31 mg/kg
SI = 1.00



LD_{50} = 8.6 mg/kg
SI = 0.93

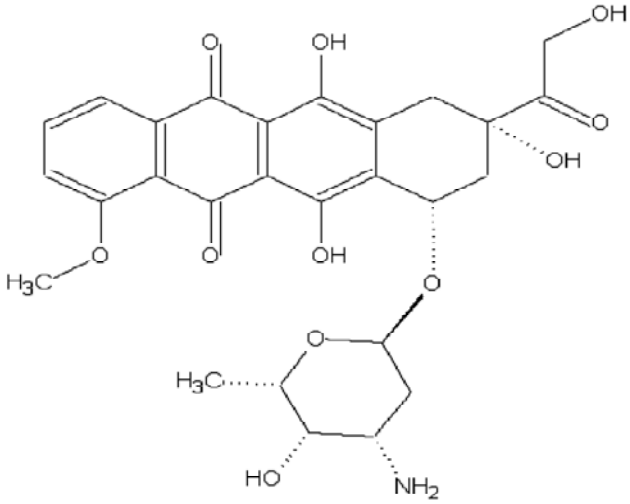


LD_{50} = 4.1 mg/kg
SI = 0.92



LD_{50} = 0.32 mg/kg
SI = 0.83

Doxorubicin, Mouse/Oral

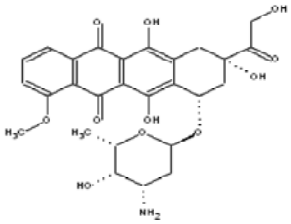


Mouse, Oral administration:

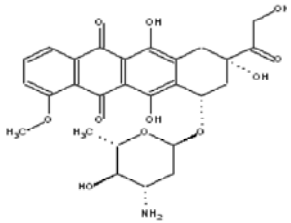
Estimated LD_{50} = 65 mg/kg
RI = **0.04**

?

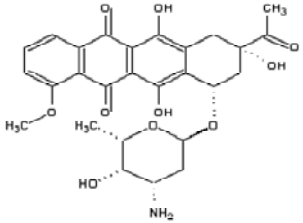
Five most similar compounds with experimental values from the library:



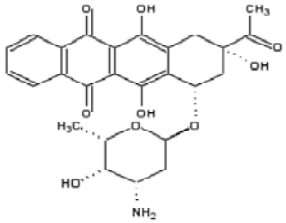
LD_{50} = 570 mg/kg
SI = 1.00



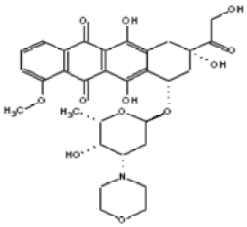
LD_{50} = 698 mg/kg
SI = 1.00



LD_{50} = 205 mg/kg
SI = 0.98



LD_{50} = 16 mg/kg
SI = 0.98

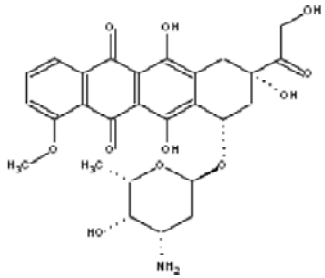


LD_{50} = 0.5 mg/kg
SI = 0.93

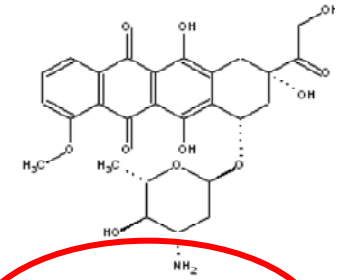
Experimental LD50s for similar compounds

Closer look at the experimental values of LD50s for similar compounds in the training set:

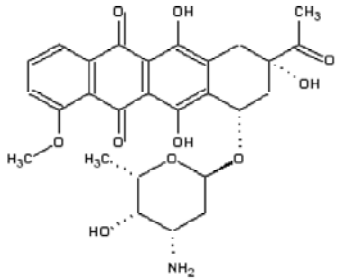
:



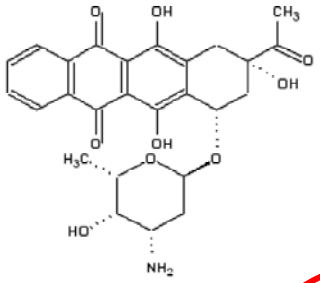
LD_{50} = 570 mg/kg
SI = 1.00



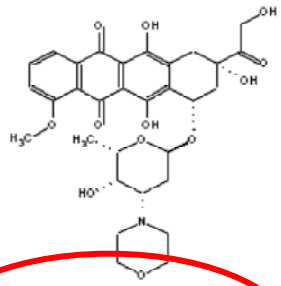
LD_{50} = 698 mg/kg
SI = 1.00



LD_{50} = 205 mg/kg
SI = 0.98



LD_{50} = 16 mg/kg
SI = 0.98

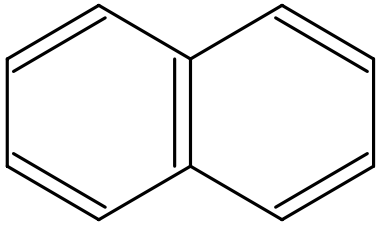


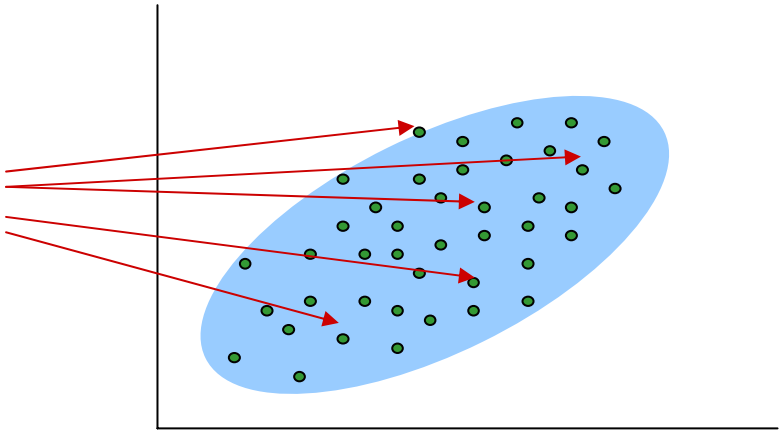
LD_{50} = 0.5 mg/kg
SI = 0.93

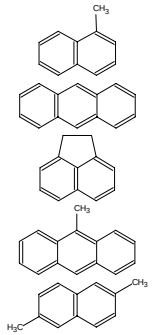
LD50 ranges from 0.5 mg/kg to 700 mg/kg!

- Can we have an idea how well model will perform by looking at similar compounds?

How well will we predict? (1)





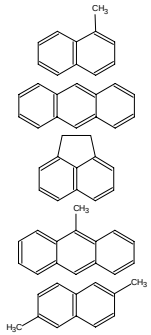


Compound we are making prediction for

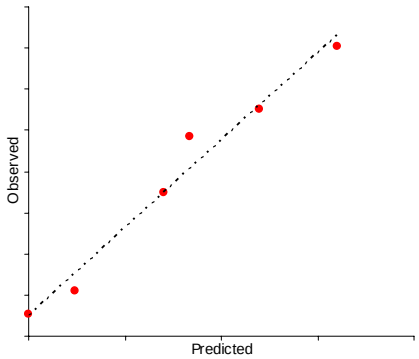
Compounds in the training set

The most similar compounds in the training set

How well will we predict? (2)



1.97	1.75
0.92	0.73
2.28	1.69
1.47	1.93
0.80	0.49



The most similar compounds in the training set

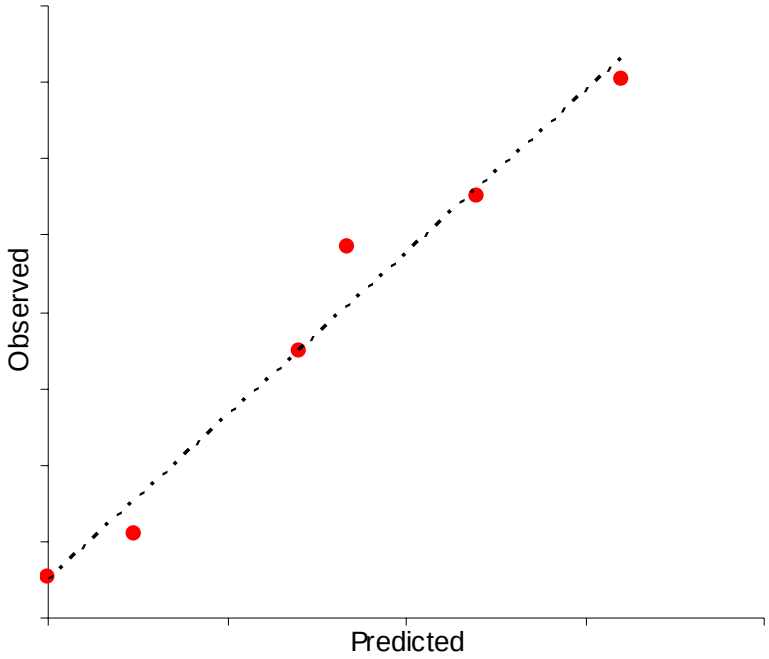
Retrieve measured values

Predict values using model

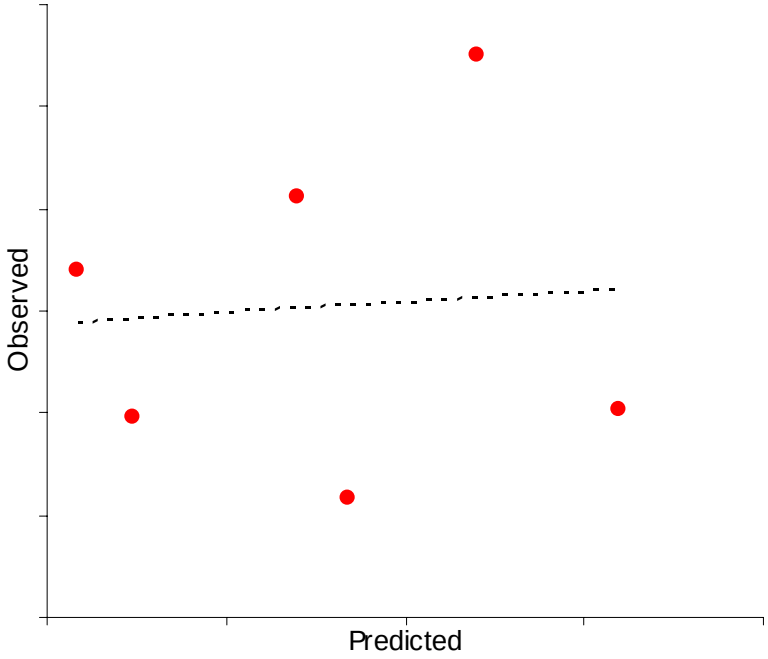
Analyse model performance for those compounds

Model performance for similar compounds (1)

Two compounds, with the same Similarity Index for both. Scatter plots illustrating model performance for similar compounds.



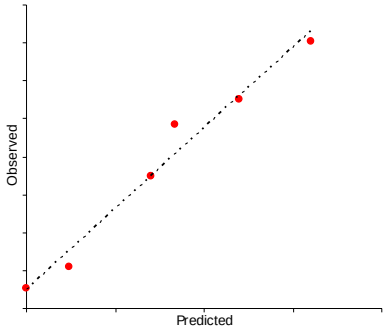
Similar compounds of compound A

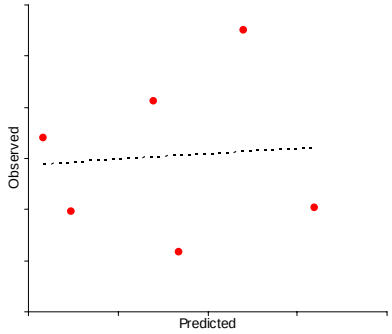


Similar compounds of compound B

Which prediction shall we trust more? For compound A or B?

Model performance for similar compounds (2)





Compound A:

Model predictions for similar compounds agree with experimental data

Compound B:

Model doesn't agree with experimental data

Reasons:

1. Model doesn't work
   1. Model doesn't work in general
   2. Model doesn't work for this particular class of compounds
   3. Similarity doesn't work – selected compound are not similar
2. Data quality is bad
3. Both model and data are bad

Model-Data Consistency Index

Considerations from the previous slides were reflected in the developed Model-Data Consistency Index, which is calculated looking at the predicted and measured values for the similar compounds:

$$MDCI = e^{-\left(\sum_{i=1}^{n} a^{i-1} \cdot SI_i \cdot (\Delta_i - \overline{\Delta})^2 \Big/ \sum_{i=1}^{n} a^{i-1}\right) / b}$$

Scaling to [0;1] range

Summation of residual errors between observed and predicted values for similar compounds

Δ_i - Difference between estimated and experimental value for the ith nearest neighbor

$\overline{\Delta}$ - Average difference for the neighbors

SI_i - Similarity to the ith nearest neighbor

a, b - Scaling parameters

Reliability Index

Reliability Index is a product of Similarity and Model-Data Consistency indices:

$$RI = SI \cdot MDCI$$

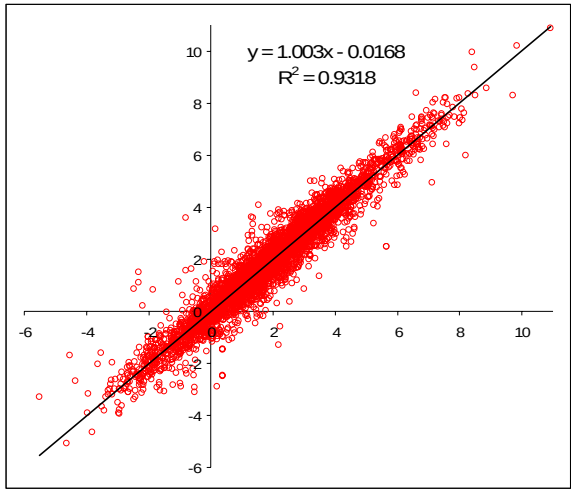
- RI will be low if SI is low (no similar compounds) OR MDCI (model doesn't agree with the experiment for the similar compounds) is low

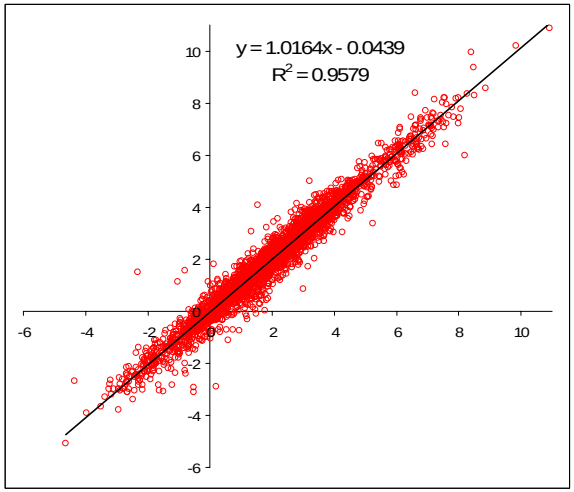
- RI will be high only if we have similar compounds AND model performs well compared to measured values on those compounds

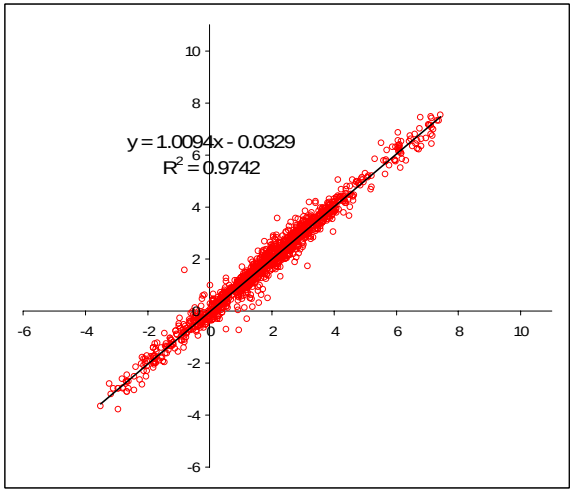
Demonstrating relationship between accuracy of the predictions and Reliability Index (Validation)

RMSE vs. RI (1)

System: LogP (training set size: 10593 compounds)
Results on the validation set:







R^2 = 0.9318

RMSE = 0.503

N = 5296

R^2 = 0.9579

RMSE = 0.404

N = 3486

(compounds with estimated RI > 0.75)

R^2 = 0.9742

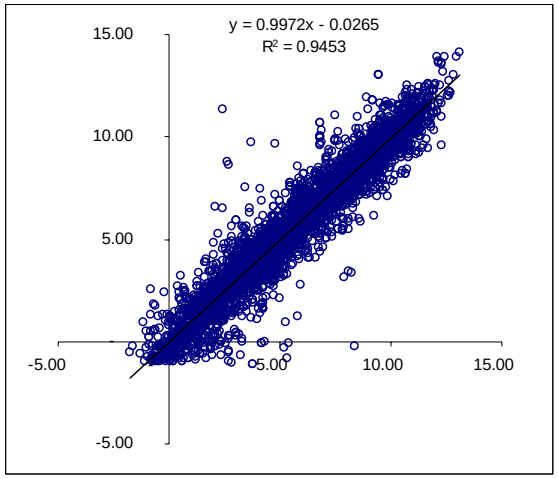
RMSE = 0.313

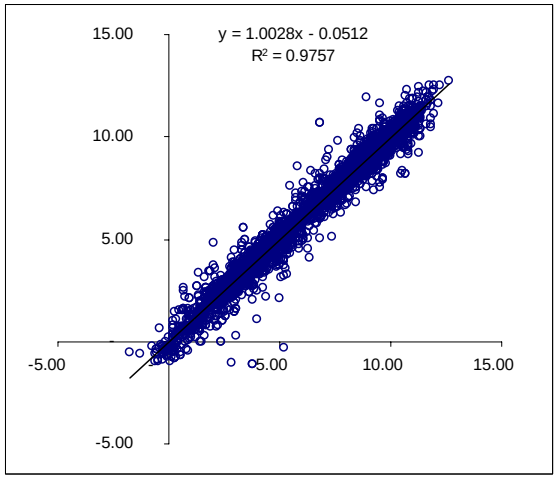
N = 1220

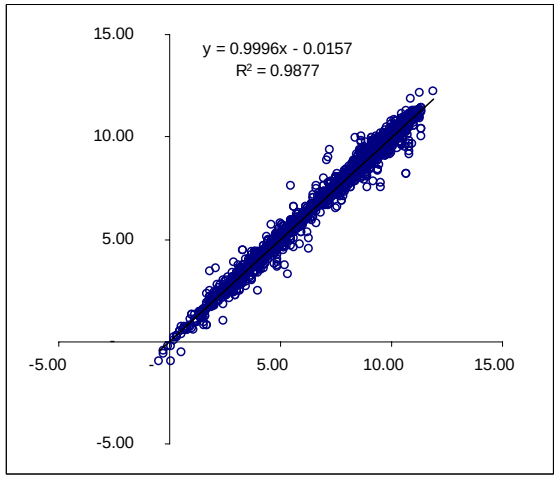
(compounds with estimated RI > 0.85)

RMSE vs. RI (2)

System: pK_a(base) (training set size: 8335 compounds)
Results on the validation set:







R^2 = 0.9453

RMSE = 0.723

N = 7950

(compounds with estimated RI > 0.5)

R^2 = 0.9757

RMSE = 0.460

N = 5498

(compounds with estimated RI > 0.8)

R^2 = 0.9877

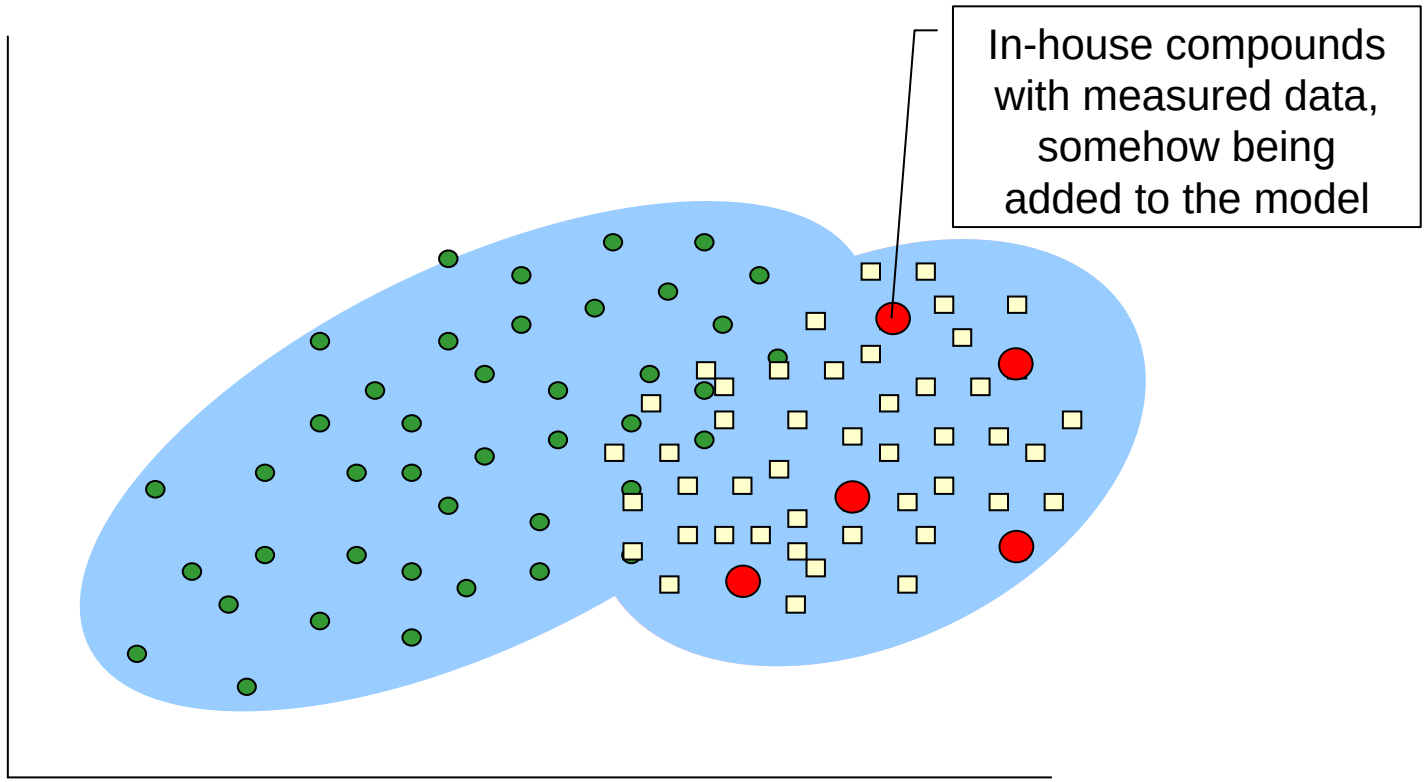
RMSE = 0.302

N = 3010

(compounds with estimated RI > 0.9)

Improving models using in-house data

Expanding Model Applicability Domain



– Compound in the Training Set of the Model

– Model Space (Applicability Domain)

– New Compound, we are making prediction for

– Chemical Space of in-house compounds

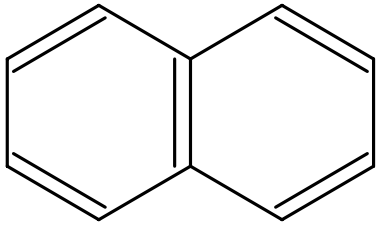
Algorithm for Trainable Models

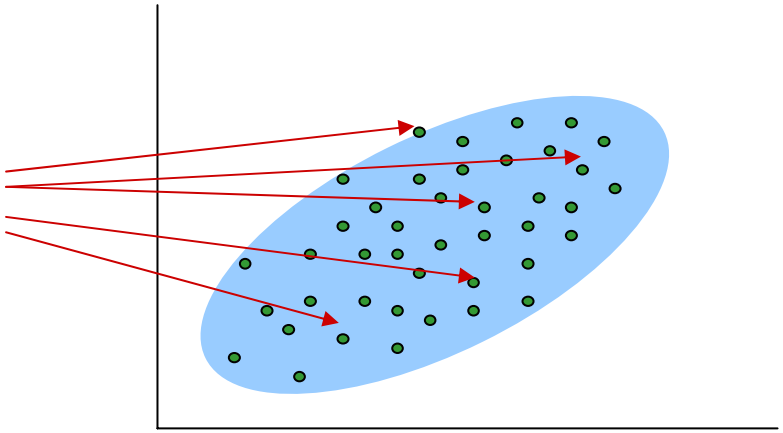
Prediction algorithm behind Trainable Models consists of two parts:

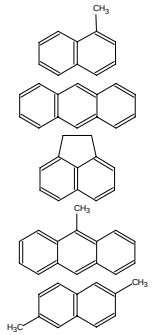
PLS + Correction according to similar structures

- Linear, additive method
- Learns global trends and what's "similar" for particular property
- It's constant, doesn't change

- Adds non-linearity
- Makes correction to the original prediction from PLS by analysis of local environment
- It's changing, when new data is added – trainable part of the algorithm



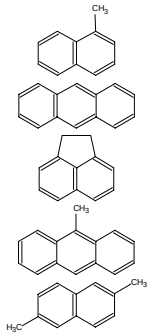




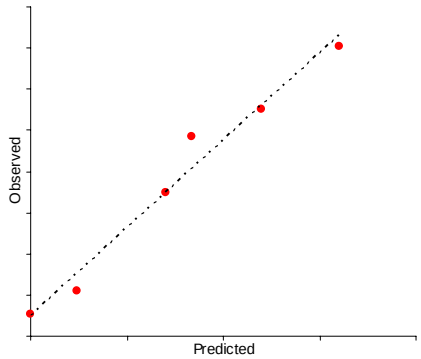
Compound we are making prediction for

Compounds in the training set

The most similar compounds in the training set



1.97	1.75
0.92	0.73
2.28	1.69
1.47	1.93
0.80	0.49



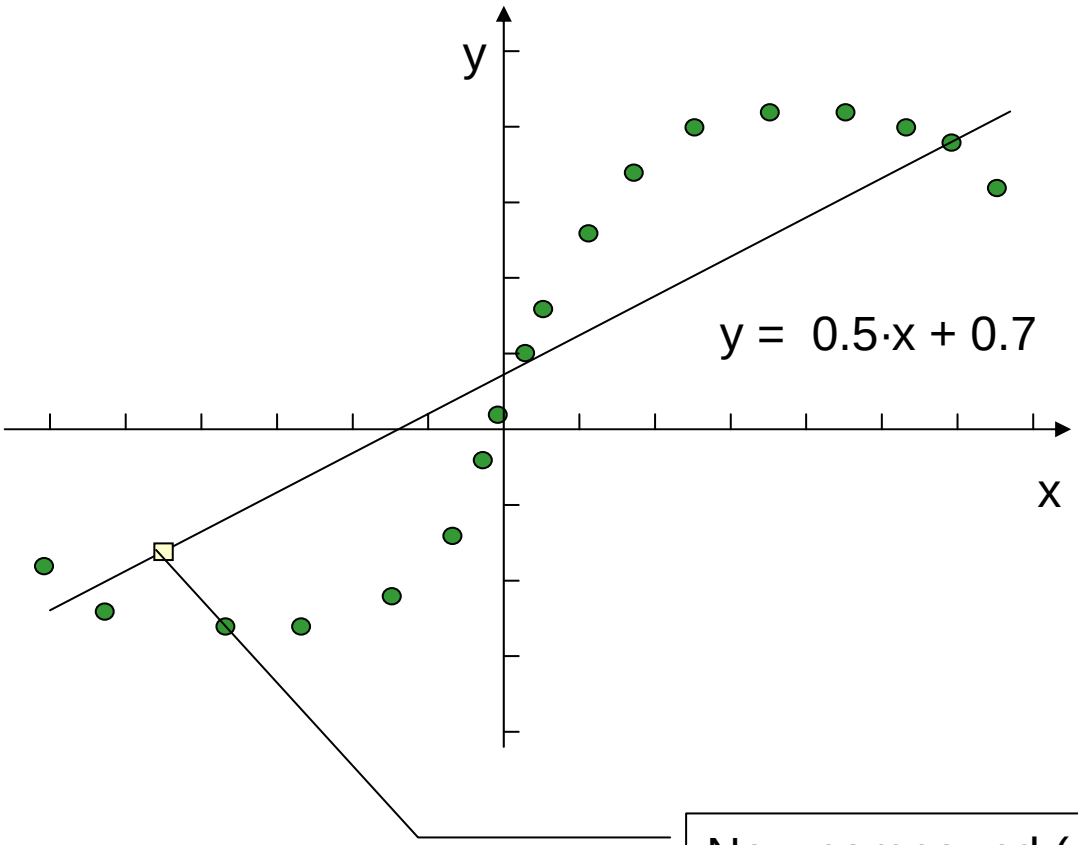
The most similar compounds in the training set

Retrieve measured values

Predict values using model

Do **local** modeling

PLS Illustration

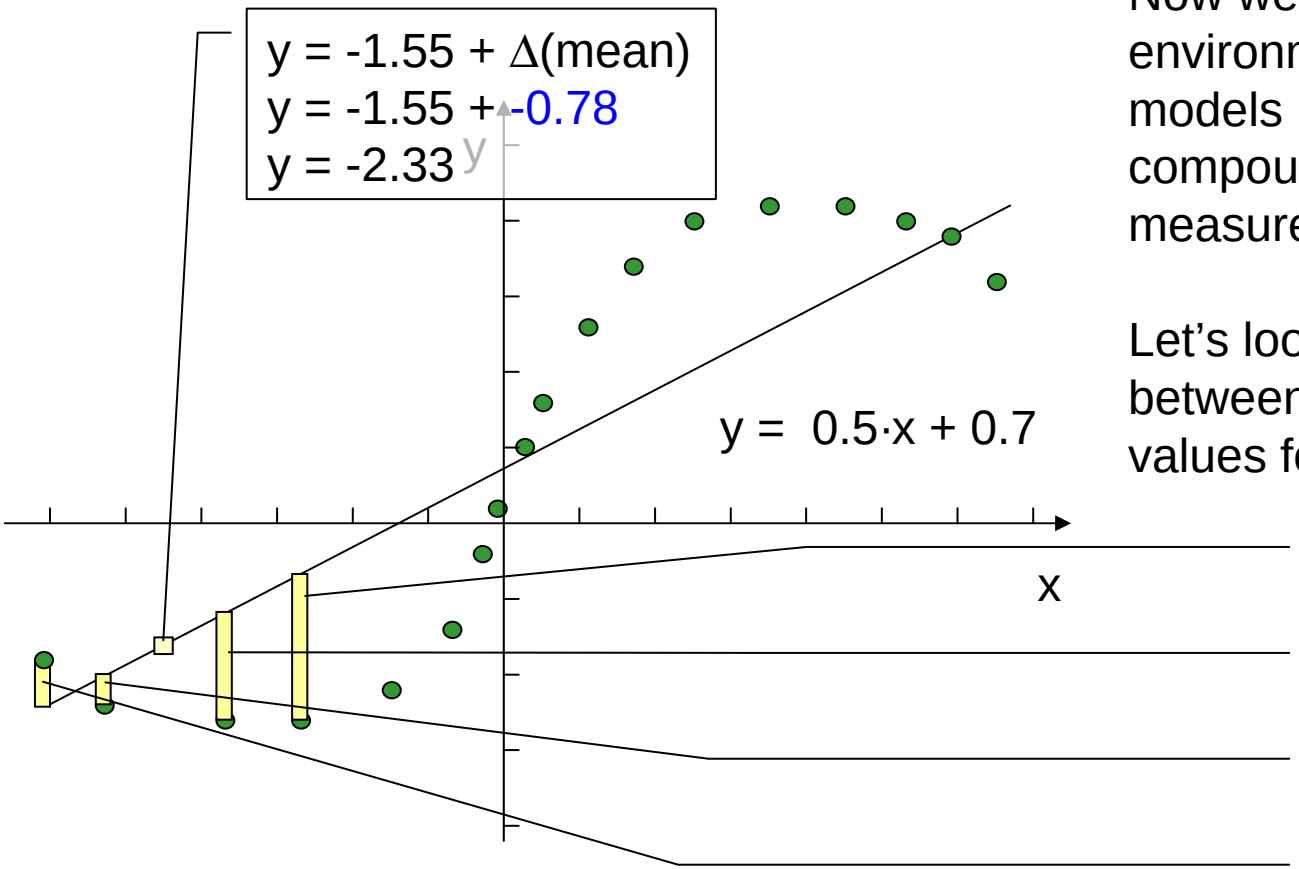


Simple model relating property y and descriptor x

PLS analysis would come up with linear model y = 0.5*x + 0.7

New compound (x = -4.5),
using equation y = 0.5 * x + 0.7,
we get y = -1.55

PLS + Similarity Illustration



Now we also explore local environment, analyzing how models performs for similar compounds with known measurements

Let's look at the difference between Observed and Predicted values for those compounds:

Δ = -1.93

Δ = -1.42

Δ = -0.40

Δ = +0.63

On average: Δ(mean) = -0.78

Validation Studies of Trainable Models

Validation case:

Take existing model of aqueous solubility (based on publicly available data).

Add portions of in-house measured data retraining the model.

Check predictivity against validation set of in-house compounds after each addition of experimental data.

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
Size of the training set	No of cmpds	MAE	**No of cmpds**	**MAE**	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	**270**	**0.841**	48	0.695		
Built-in + 100	400	0.894	**291**	**0.873**	57	0.710		
Built-in + 250	400	0.868	**303**	**0.841**	76	0.703		
Built-in + 500	400	0.821	**327**	**0.774**	144	0.586	58	0.469
Built-in + 750	400	0.697	**365**	**0.658**	256	0.534	118	0.434
Built-in + 913	400	0.624	**382**	**0.616**	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
Size of the training set	No of cmpds	MAE	No of cmpds	MAE	**No of cmpds**	**MAE**	No of cmpds	MAE
Built-in	400	0.900	270	0.841	**48**	**0.695**		
Built-in + 100	400	0.894	291	0.873	**57**	**0.710**		
Built-in + 250	400	0.868	303	0.841	**76**	**0.703**		
Built-in + 500	400	0.821	327	0.774	**144**	**0.586**	58	0.469
Built-in + 750	400	0.697	365	0.658	**256**	**0.534**	118	0.434
Built-in + 913	400	0.624	382	0.616	**304**	**0.530**	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	**400**	**0.894**	**291**	**0.873**	**57**	**0.710**		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
Size of the training set	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	**400**	**0.868**	**303**	**0.841**	**76**	**0.703**		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	**400**	**0.821**	**327**	**0.774**	**144**	**0.586**	**58**	**0.469**
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

Size of the training set	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	**400**	**0.697**	**365**	**0.658**	**256**	**0.534**	**118**	**0.434**
Built-in + 913	400	0.624	382	0.616	304	0.530	154	0.433

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

External Validation at Syngenta, Solubility Model

Pharma Algorithms
WWW.AP-ALGORITHMS.COM

System: Aqueous solubility (training set size varies)
Results on the validation set:

	Whole test set		Unreliable removed (RI>0,3)		Moderate and high (RI>0,5)		High (RI>0,7)	
Size of the training set	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE	No of cmpds	MAE
Built-in	400	0.900	270	0.841	48	0.695		
Built-in + 100	400	0.894	291	0.873	57	0.710		
Built-in + 250	400	0.868	303	0.841	76	0.703		
Built-in + 500	400	0.821	327	0.774	144	0.586	58	0.469
Built-in + 750	400	0.697	365	0.658	256	0.534	118	0.434
Built-in + 913	**400**	**0.624**	**382**	**0.616**	**304**	**0.530**	**154**	**0.433**

MAE ranges
(0,8 - 0,9]
(0,7 - 0,8]
(0,6 - 0,7]
(0,5 - 0,6]
(0,4 - 0,5]

Any model, no matter which descriptors or statistical methods were used in its development cannot be better than the data it is based on.

Every empirical model works only in certain chemical space, where the compounds from the training set are located – boundaries of Model Applicability Domain

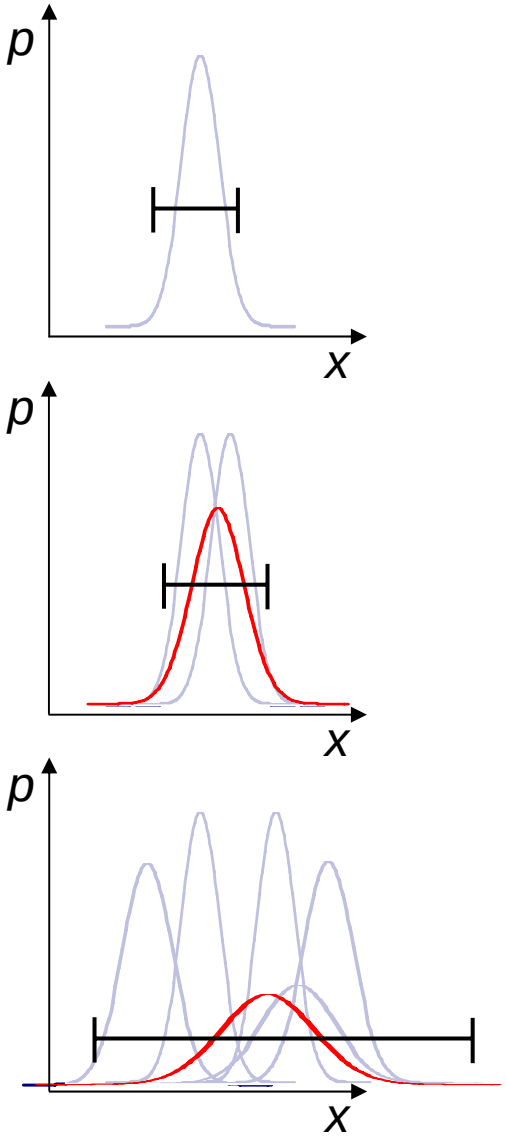
Even within Model Applicability Domain model cannot get more accurate than the unexplainable variation in measured values.

Unexplainable variability in measured data

1. Measurements from the same laboratory

2. Same protocol from several laboratories

3. Compilation of publicly available data



Models incorporating in-house data

Better accuracy is achieved because:

- Chemical space is much better represented, model learns structure-property relationships for new chemical classes
- Several layers of data “noise” are removed as we use much more consistent data – model “learns” your methodology, your protocol

Improving models using in-house data. How many compounds do you need to improve a model?

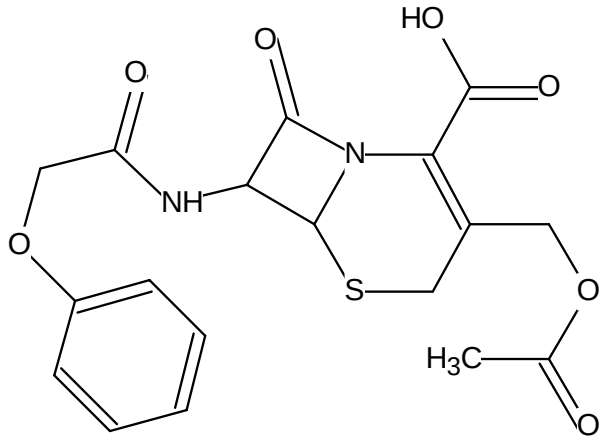
Validation case:

Special built LogP model – no beta-lactam antibiotics in the training set!

Add beta-lactam antibiotics with measured LogP one after another to improve the model

Check predictivity against validation set and selected compounds (beta-lactam antibiotics).

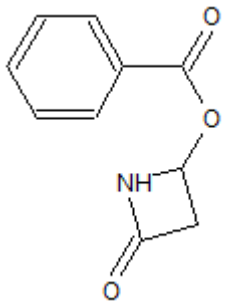
Phenoxymethylcephalosporin example



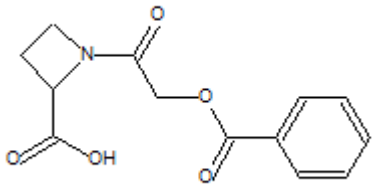
Measured LogP = 0.26

Calculated LogP = 1.46
Estimated RI = 0.26

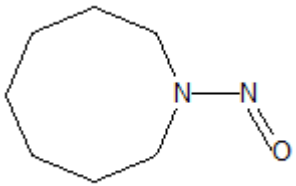
Five most similar compounds from the library:



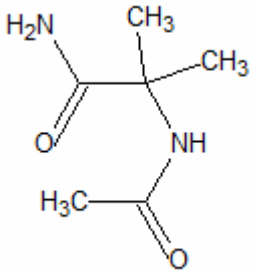
SI = 0.49



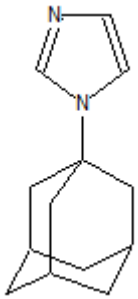
SI = 0.17



SI = 0.12

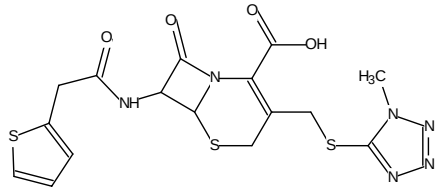
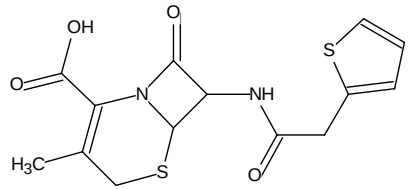
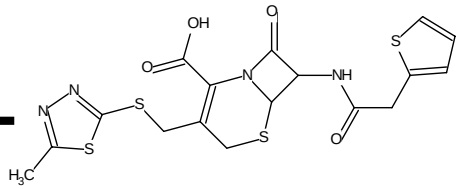


SI = 0.11



SI = 0.10

Prediction results for Phenoxymethylcephalosporin:

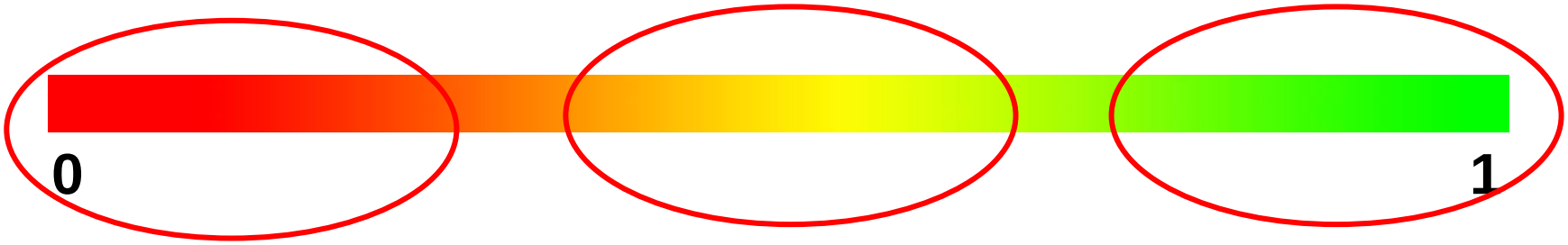
	Predicted LogP	RI	Predicted - Observed
Original prediction (no beta lactam antibiotics in the training set)	1.46	0.26	1.2
+ 	1.05	0.53	0.79
+ 	0.62	0.61	0.36
+ 	0.32	0.71	0.06

Applications of the Reliability Index and Traiinable Models – Connecting measurement and prediction

- As it was shown in the previous slides, RI indeed reflects Model Applicability Domain and can be used as indicator of usability of predicted values

- RI can also be used in experiment planning and prioritization of the measurements – most knowledge will be gained measuring compounds for which estimated RI is the lowest.

Scale of Reliability Index:



Low RI values, these compounds should be measured first and if possible added to model

We expect that amount of compounds with estimated intermediate Reliability Index will decrease after properties of compounds with low RI will be measured and values added to the model for further improvement

High RI values, class of compounds well represented in the training set, model already performs well on those compounds. Measurements would be redundant

Acknowledgements

Pharma Algorithms

- Rytis Kubilius
- Andrius Sazonovas
- Remigijus Didžiapetris
- Kiril Lanevskij

Syngenta

- Eric Clarke
- John Delaney
- Tom Sheldon (Industrial Placement Student, University of Bath)

PhysChem Forum 5 Organizers

Thank You for your attention