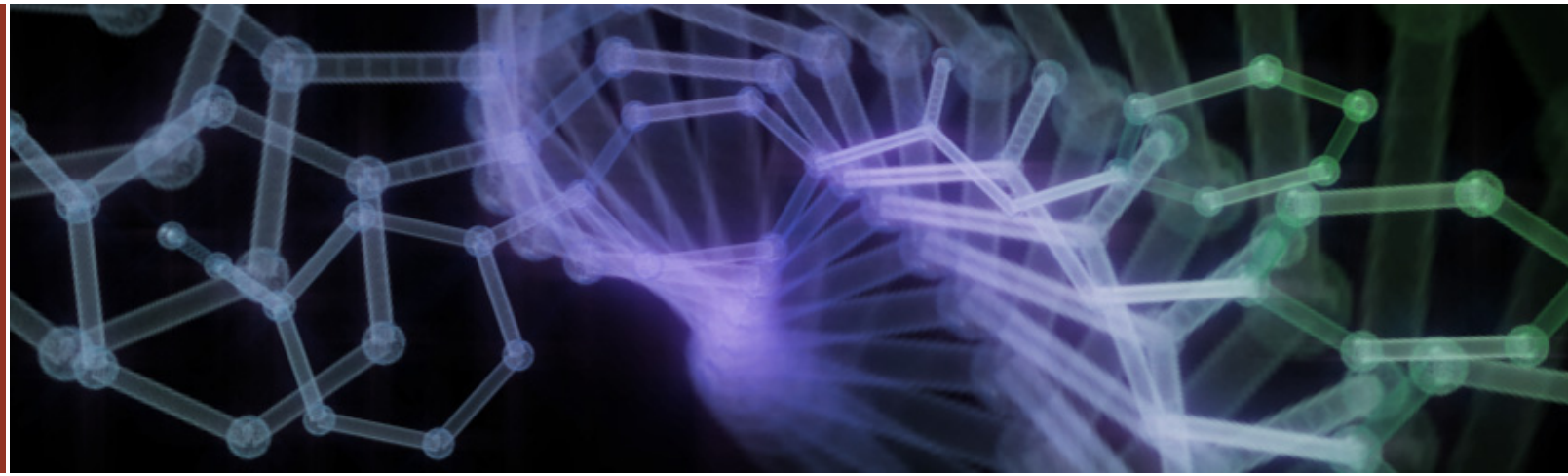


QSAR in Virtual Screening and Lead Optimisation

Dr. Peter Gedeck

Novartis Institutes for BioMedical Research

NOVARTIS
INSTITUTES FOR
BIOMEDICAL RESEARCH

Who are we?

Novartis at a glance

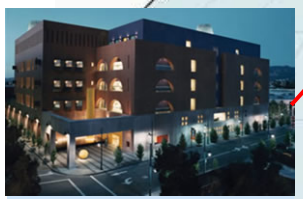
- World's third largest pharmaceutical company by sales
- One of 20 largest companies by market capitalisation
- Ranked among most respected companies worldwide
- Unique portfolio to meet changing healthcare needs:
  - Leading innovative pharmaceuticals
  - High-quality, low-cost generics (Sandoz)
  - Preventative vaccines
  - Eye care and Ophthalmic (CibaVision)
  - Consumer health products

NIBR – A Global Organisation

Horsham UK
Respiratory
Gastro-Intestinal

Cambridge, MA
Diabetes & metabolism
Infectious diseases
Cardiovascular
Oncology
Ophthalmology

Basel, CH
Immunology, Neuroscience
Oncology, Proteases
Musculoskeletal



Emeryville, CA
Oncology

East Hanover, NJ
Chemical development

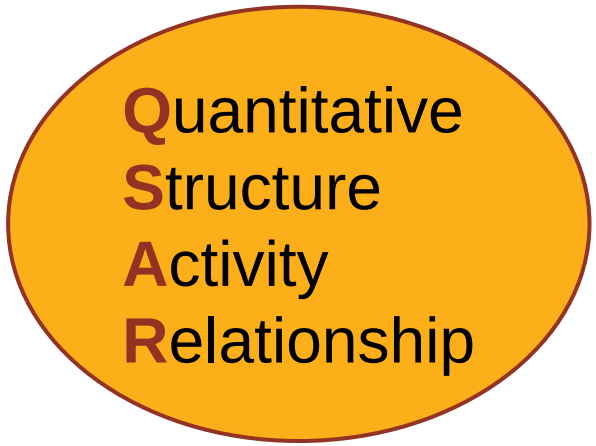
Shanghai, CHN
Cancers of infectious diseases

NOVARTIS

QSAR – definition

QSGR
Genomic

QSPR
Properties



QSXR
… X-omics

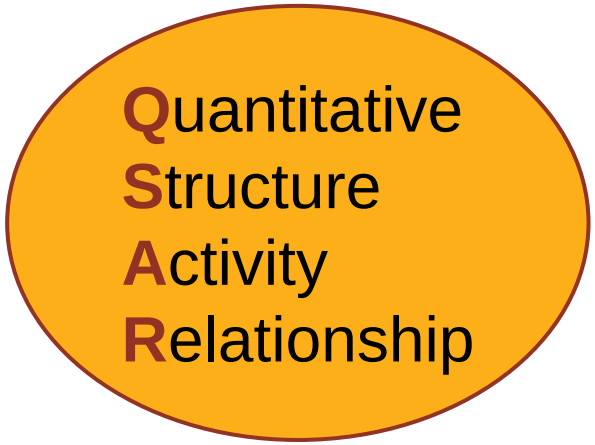
QSTR
Toxicology

- QSAR: Quantitatively relating biological activity to structure.
- However, I will use the term QSAR very general and not only talk about activity, but any possible endpoint

QSAR – definition

Data

Descriptors



Applications

Models

Example: Modeling log*BB* (Blood-brain barrier)

Applications: QSAR in lead optimisation and lead finding

QSAR – definition

- Rational QSAR
  - Use structural properties related to Free Energy Relationships
- Pragmatic QSAR
  - Use similarity, in effect predicting compounds according to similarity to those in the training set.
  - Relevance of descriptors to Free Energy is unknown.

Rational QSAR

- Use structural properties related to Free Energy Relationships
- Non-covalent Drug-Macromolecular interactions are equilibrium processes (as is partitioning)
  - Dissociation constants are determined by the difference in free energy of bound to unbound states.

$$\Delta G = -RT \log K_i$$

K_i is proportional to IC_{50}, therefore ΔG is proportional to $-\log IC_{50}$

$$\log IC_{50} = c\,\Delta G$$

- Use $-\log IC_{50}$ or pIC_{50} to model ΔG

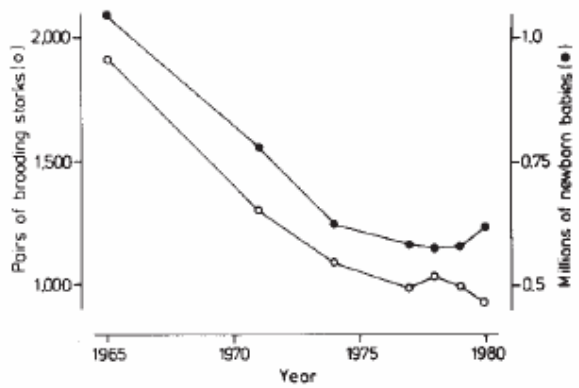
Pragmatic QSAR

- Neighbourhood principle:
  - Molecules with similar structural properties are likely to have similar biological activities
- Don’t worry about theory, just work out combination of properties that give good activity
- Variety of descriptors derived for a structure.
  - Examples: Topological indices, pharmacophores, IR spectra, 2D fragments, structural fingerprints etc.
- However beware: Correlation without causation!

Correlation and Causation

A new parameter for sex education

SIR—There is concern in West Germany over the falling birth rate. The accompanying graph[1,2] might suggest a solution that every child knows makes sense.



HELMUT SIES
Institut für Physiologische Chemie 1, Universität Düsseldorf, Moorenstrasse 5, D-4000 Düsseldorf, FRG

1. *Fachserie Gebiet und Bevölkerung* (Statistisches Bundesamt, Kohlhammer, Stuttgart, 1984).
2. Bauer, S. & Thielcke, G. *Die Vogelwarte* **31**, 183–191 (1982).

- The correlation of storks and babies on the left is 0.99
- Even if a descriptor is highly correlated with activity, it may not be its cause
- This is particularly true for *pragmatic* QSAR models

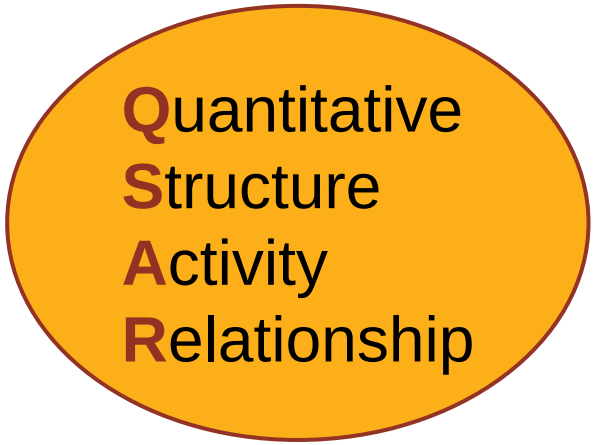
Source: Sies H, Nature 332 (1988) 495. April issue!

NOVARTIS

QSAR – definition

Data

Descriptors



Applications

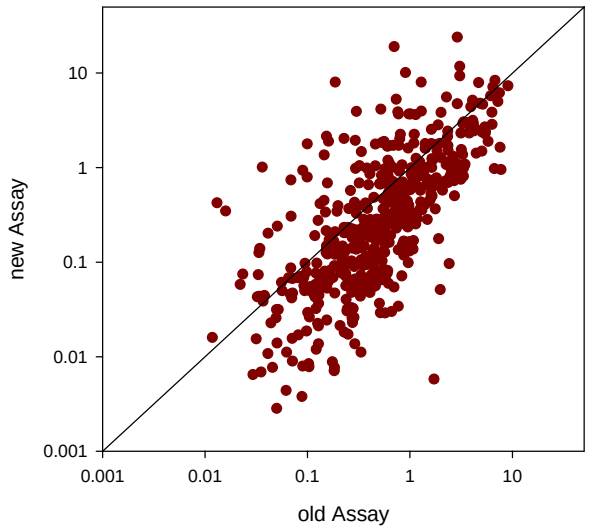
Models

Example: Modeling log*BB* (Blood-brain barrier)

Applications: QSAR in lead optimisation and lead finding

Data quality

- What does this graph show?



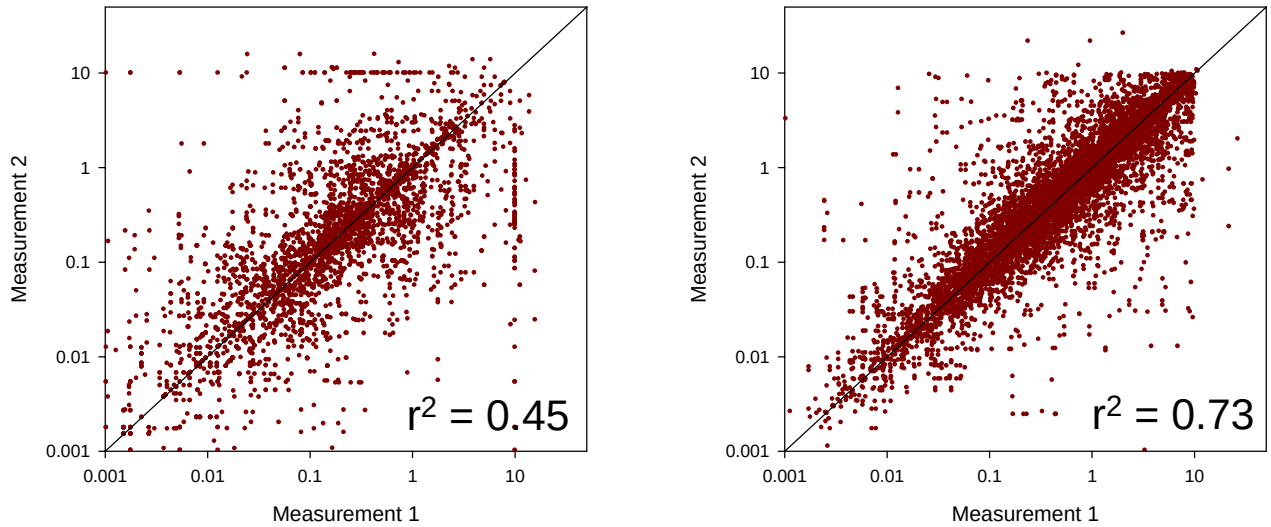
$r^2 = 0.5$!

  - Results for two different in-vitro assays for the same target (drift in assay conditions)

- Be realistic: QSAR models can only be as good as your data

Data quality

- Even within an assay we see variability of the data
  - Comparison of results from repeated measurements
  - Sample concentration, performance of cells

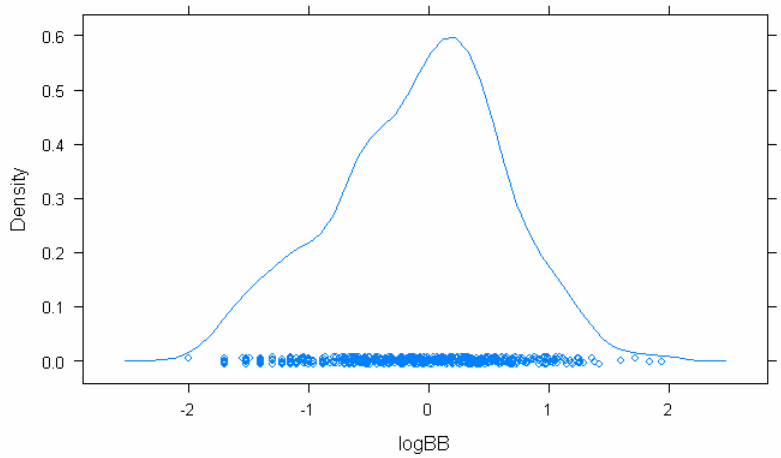


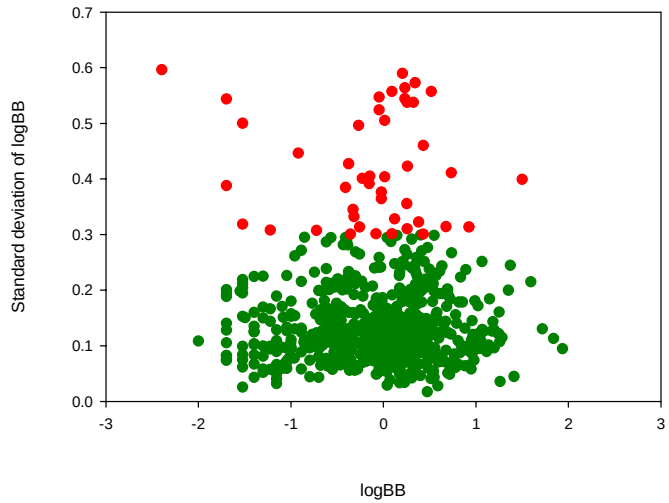
- Be realistic: Garbage in – garbage out

NOVARTIS

Data quality [Example logBB]

- Blood-brain distribution in mice
- Data distribution
  - Most results scattered around centre (logBB~0)
- Data quality
  - Variation of repeated measurements large for some compounds
- Remove compounds with large measurement error



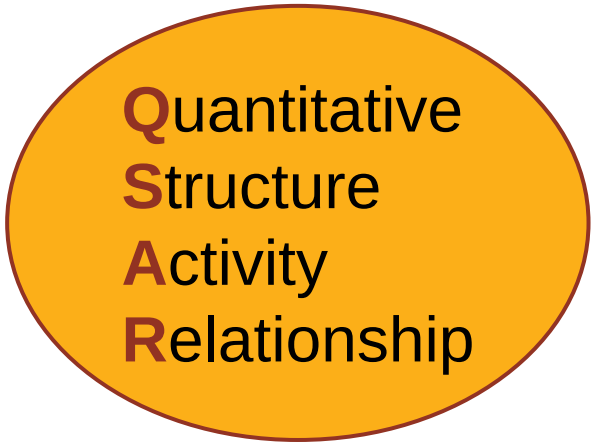


NOVARTIS

QSAR – definition

Data

Descriptors



Applications

Models

Example: Modeling log*BB* (Blood-brain barrier)

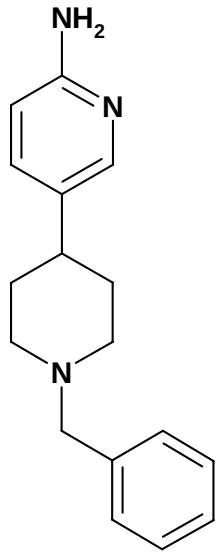
Applications: QSAR in lead optimisation and lead finding

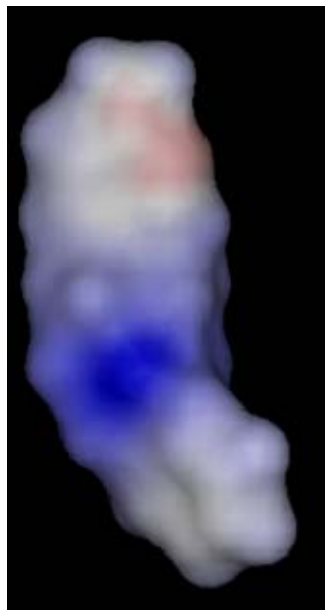
Descriptors

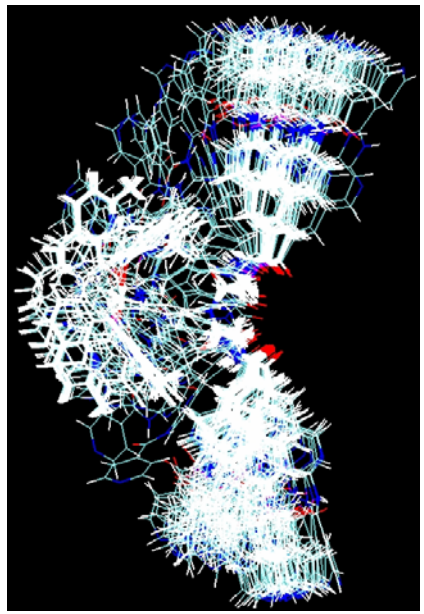
- Classification based on representation of molecule
  - 1D - molecular formula
  - 2D - molecular connectivity / topology
  - 3D - molecular geometry / stereochemistry
  - 4D/5D/… - conformational ensembles

$C_{17}H_{21}N_3$

MW =267.38



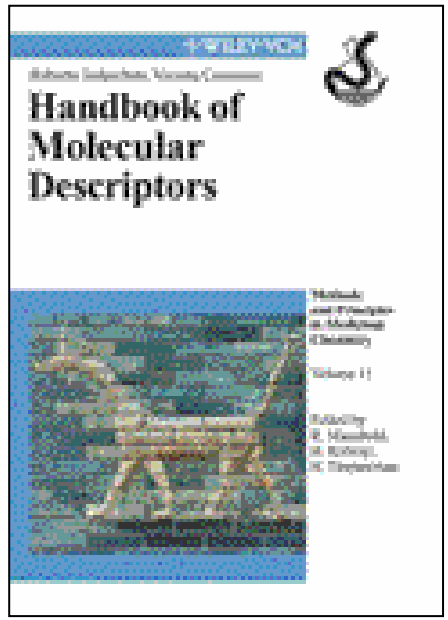




NOVARTIS

Descriptors

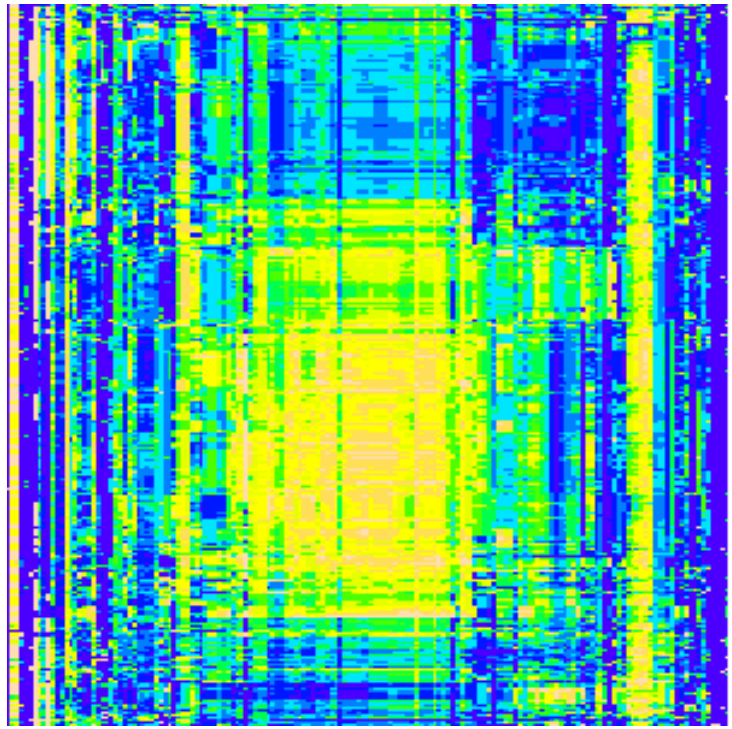
- Descriptors characterize various molecular properties by numerical values.
  - Hydrophobic, electronic, steric / size / shape, hydrogen bonding
- More then 2000 descriptors published
  - R. Todeschini, V. Consonni,
    Handbook of Molecular Descriptors, Wiley, 2000
- There are many approaches to select descriptors (forward selection, genetic algorithms, …)



NOVARTIS

Descriptors [logBB example]

- 1D/2D-Descriptors calculated using MOE [187 descriptors]
- Removed columns which don’t change [160 descriptors]
- Range normalization of descriptors to interval [0,1]

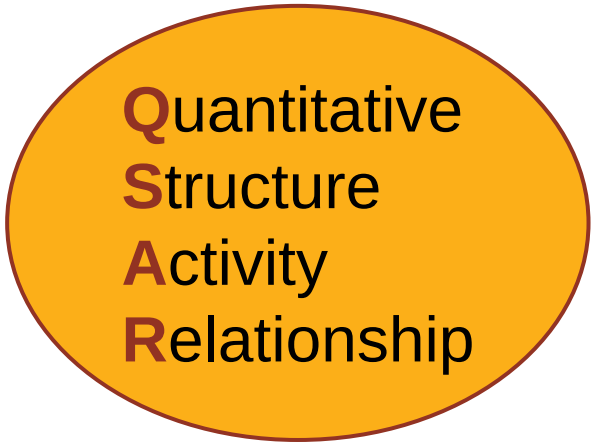


Heatmap showing descriptor values. Rows and columns are ordered using hierarchical clustering

QSAR – definition

Data

Descriptors



Applications

Models

Example: Modeling log*BB* (Blood-brain barrier)

Applications: QSAR in lead optimisation and lead finding

Models

- Model relates activity to the descriptors
  - Regression
  - Classification
- The ideal model
  - avoids overfitting (=perfect fit of training data but no generalization)
  - provides a robust estimate of the reliability and applicability of the model
  - is interpretable
  - works with all datasets
- An ideal model doesn’t exist

Regression

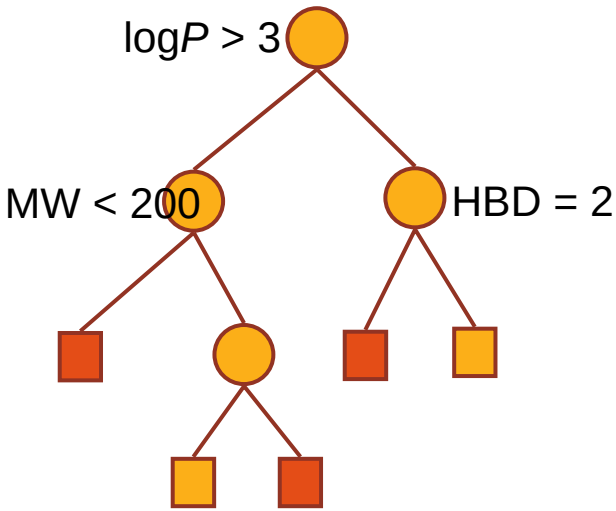
- Model a continuous dependent variable (activity) as a function of independent variables (descriptors)

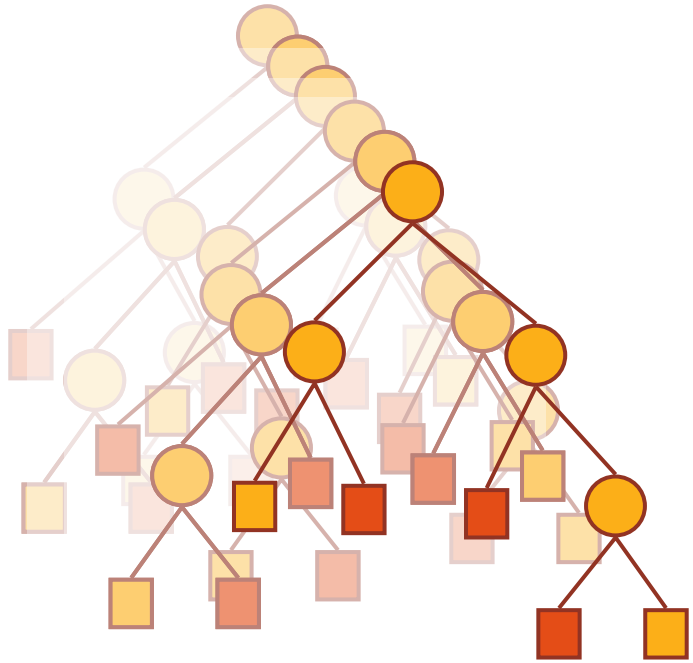
$$y = f(X)$$

- Often used methods
  - Multiple Linear Regression
  - Partial Least Squares regression:
    - Like MLR but prevents overfitting
  - Artificial Neural Networks:
    - “Black Box” generates complex rules relating the activity to the descriptors.
  - Decision Trees, Random Forests:
    - Collection of logical rules that allow prediction of activity
  - Support vector machines, gaussian processes:
    - Novel machine learning approaches

Regression [Decision trees and Random Forests]

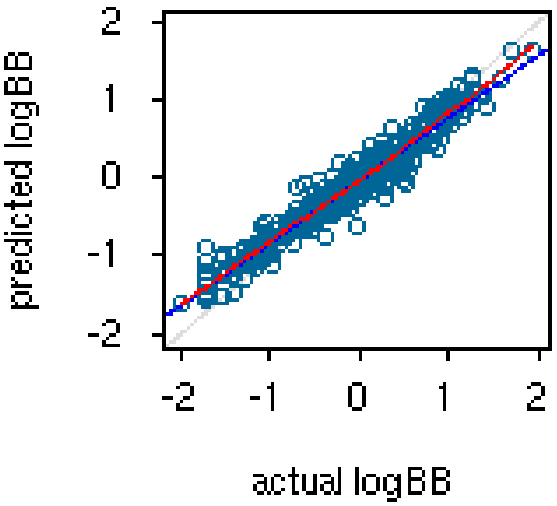
- Decision tree: A sequence of rules links properties with activity
- Ensemble of trees / random forest: Combine prediction of multiple trees

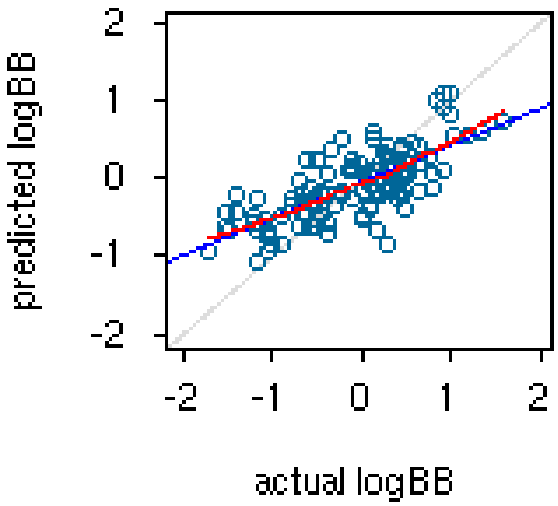




Regression [logBB example]

- Performance of random forest model on training data
  - r^2 = 0.92 [Excellent … really?]
- Performance of model on test data
  - r^2_{pred} = 0.53 [not so good anymore … what went wrong?]





NOVARTIS

Model validation

- **Internal cross-validation r^2_{CV}/q^2 value**
  - Leave one out / leave many out
  - Train several models and predict left out compounds
  - Want q^2 to be >0.4
- **External Validation Training set/Test set r^2_{pred} value**
  - Take a 20% subset of compounds as independent Test set.
  - Other 80% the Training Set. Derive a model with Training set and predict the Test set.
- **Y-Scrambling**
  - Mix up the values for the biological activity (Y, dependent variable) attempt to build a model. If possible to model the scrambled data then we have a problem.

Model validation

- Regression statistics
  - Multivariate predictive r^2_{pred}

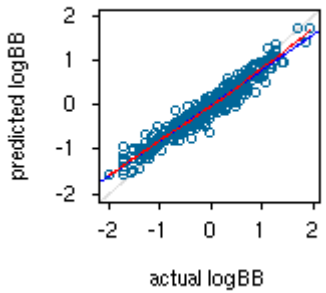
$$r^2_{pred} = 1 - \frac{\sum_{i \in test} \left(y_i^{pred} - y_i^{act} \right)^2}{\sum_{i \in test} \left(y_i^{act} - \bar{y}^{act} \right)^2}$$

  - Correlation actual versus predicted r^2_{corr}

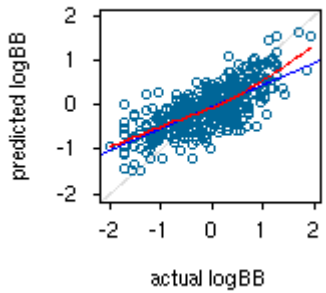
$$r^2_{corr} = \frac{\left(\sum_{i \in test} \left(y_i^{pred} - \bar{y}^{pred} \right) \left(y_i^{act} - \bar{y}^{act} \right) \right)^2}{\sum_{i \in test} \left(y_i^{pred} - \bar{y}^{pred} \right)^2 \sum_{i \in test} \left(y_i^{act} - \bar{y}^{act} \right)^2}$$

Model validation [logBB example]

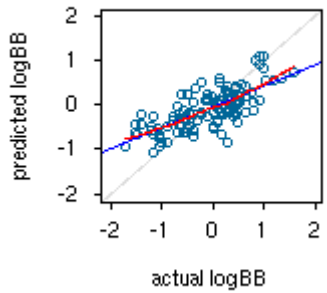
Training performance:
r2=0.92 rmse=0.203
correlation=0.97



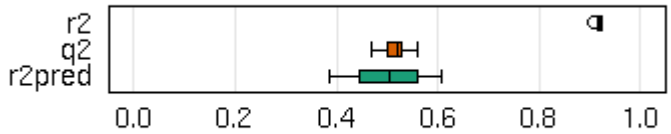
Internal validation:
q2=0.52 rmse=0.491
correlation=0.72



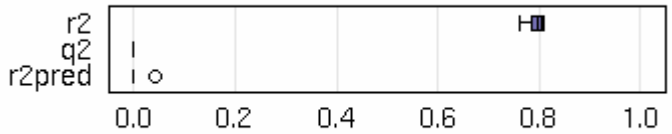
External validation:
pred-r2=0.53 rmse=0.477
correlation=0.73



Repeated training/test splits:
Boxplot shows variation of statistics



Y-scrambling:
r^2 still very good, but q^2 values below 0



NOVARTIS

Classification

- Model a discrete dependent variable (class) as a function of independent variables (descriptors)

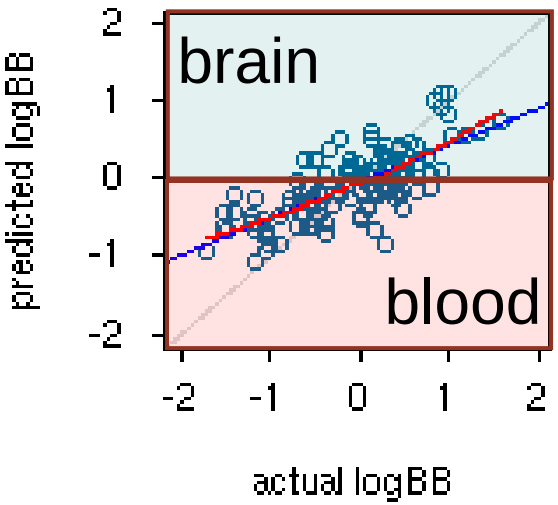
$$c = f(X)$$

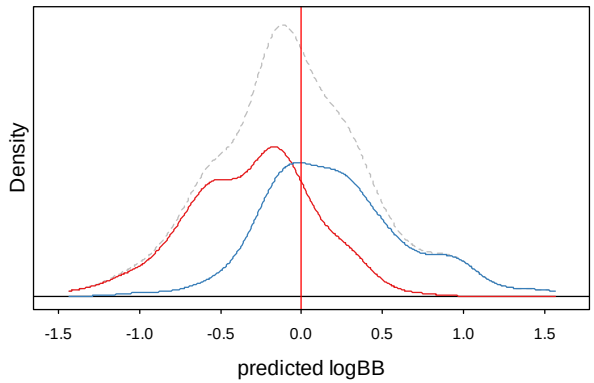
- Many methods exist, e.g.:
  - Linear Discriminant analysis
  - Decision Trees
  - Random Forests
  - Convert regression model into classifier
- Validation
  - Internal cross-validation
  - External validation using test sets
  - Scrambling

Classification [logBB example]

- Prediction from Random Forest Regression converted into Classification

$$c = \begin{cases} blood & \text{if } f(X) < 0 \\ brain & \text{if } f(X) > 0 \end{cases}$$



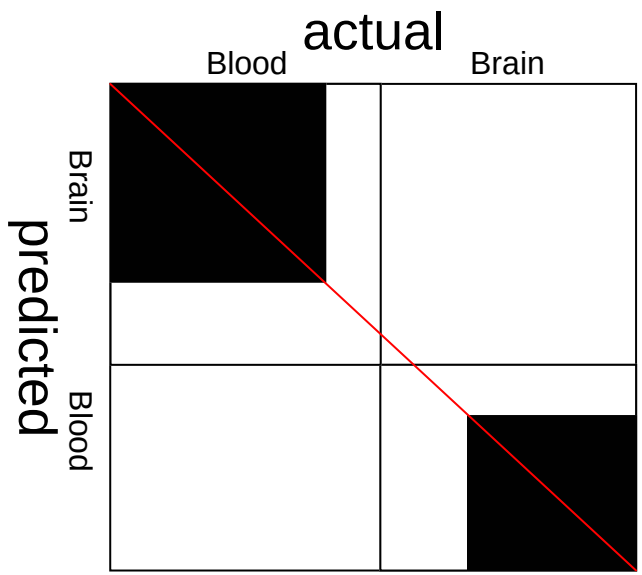


NOVARTIS

Classification [logBB example]

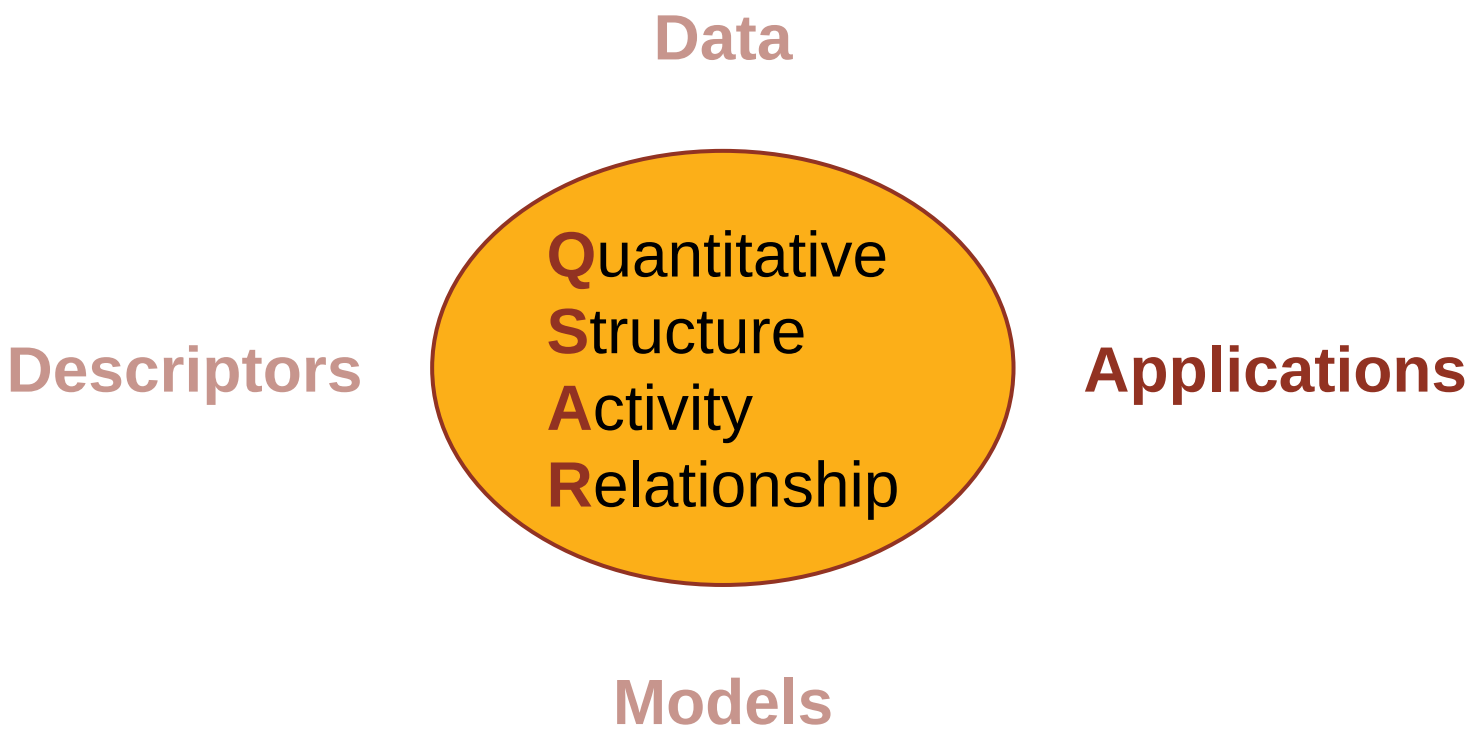
- Confusion matrix
  - Table
  - Graph
- Usual statistics
  - Accuracy: 73% (percent correctly predicted)
  - Recall (percent correctly predicted by predicted class):
    65% brain, 80% blood
  - Precision (percent correctly predicted by actual class)
    76% brain, 71% blood

	Actual	
Predicted	Blood	Brain
Blood	0.41	0.1
Brain	0.17	0.32



NOVARTIS

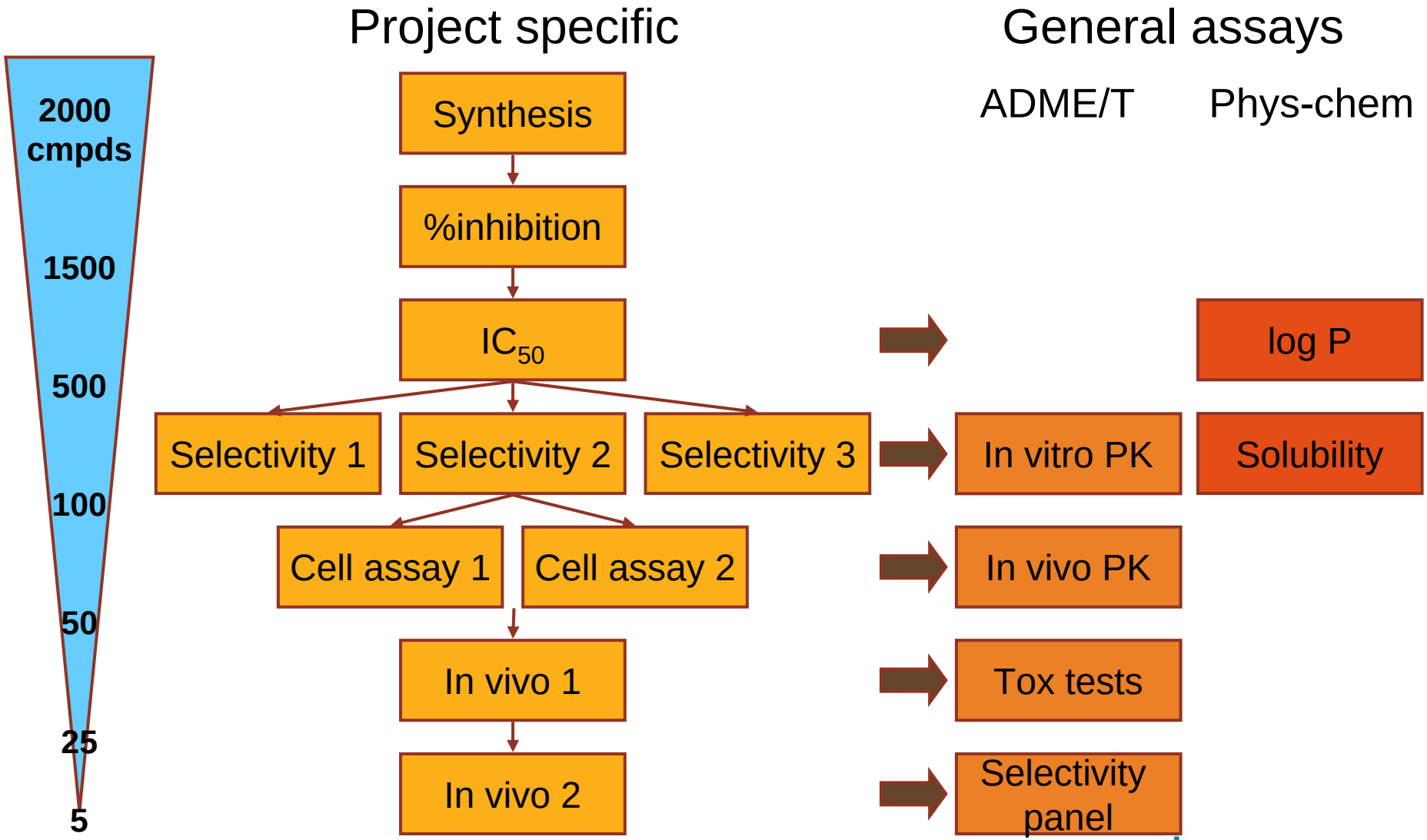
QSAR – definition



Example: Modeling log*BB* (Blood-brain barrier)

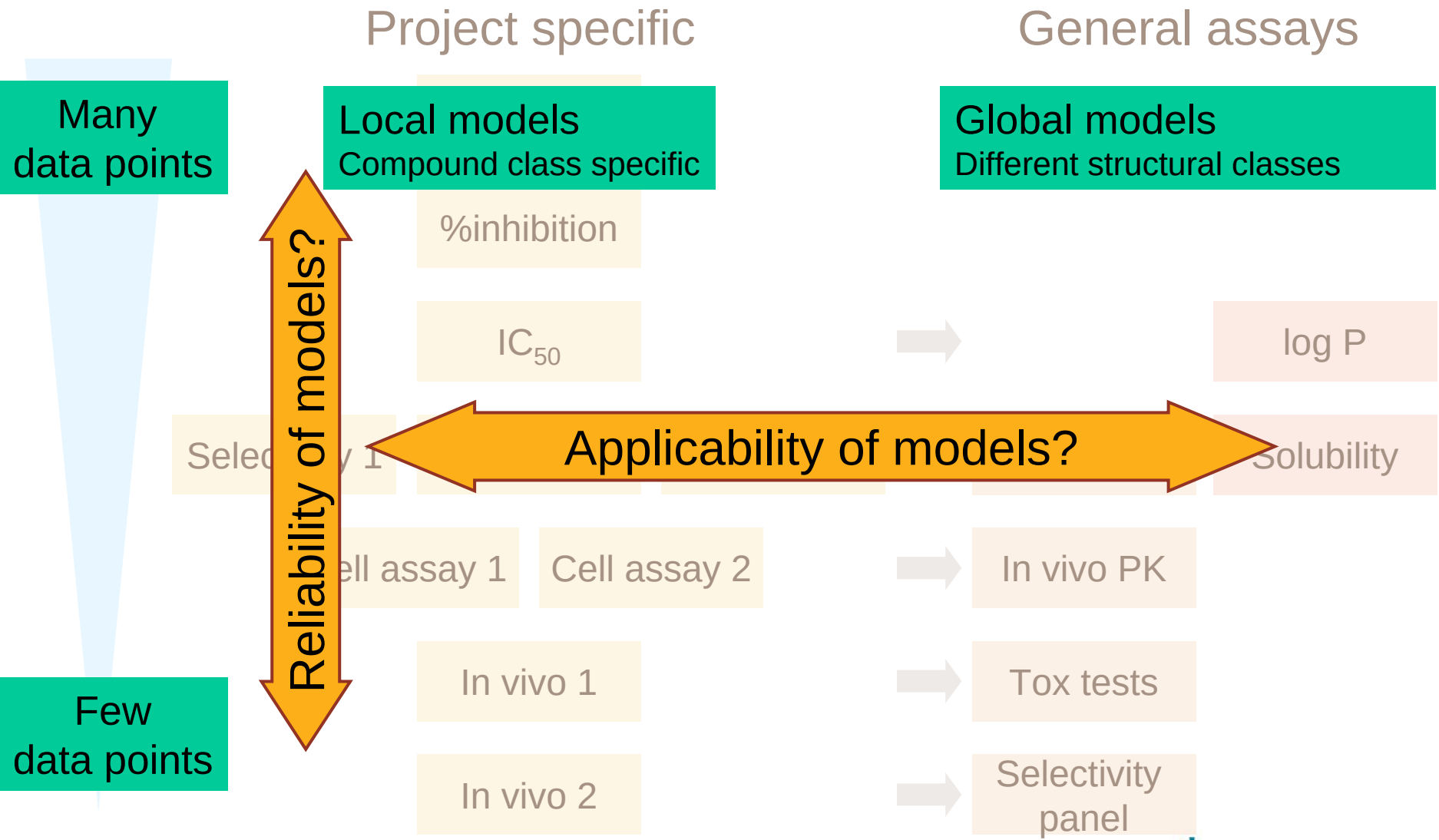
Applications: QSAR in lead optimisation and lead finding

Lead optimisation: a typical flow chart



NOVARTIS

Lead optimisation: a typical flow chart

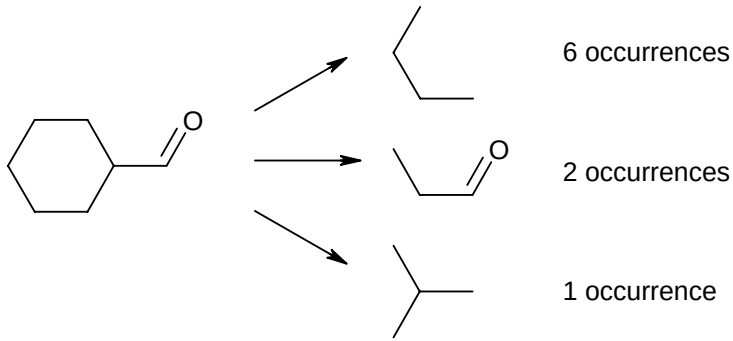


NOVARTIS

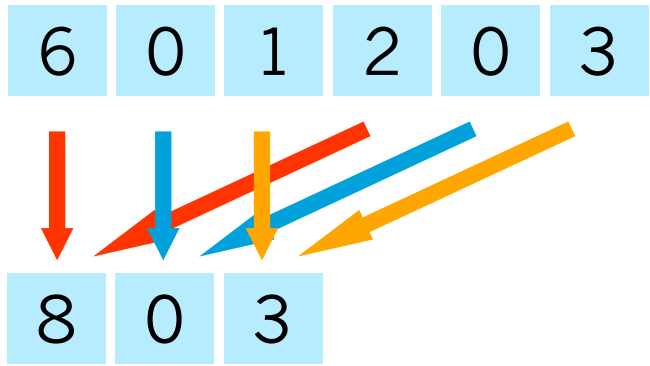
Lead optimisation – a pragmatic approach

HQSAR Descriptors

Break structure into fragments and count occurrences



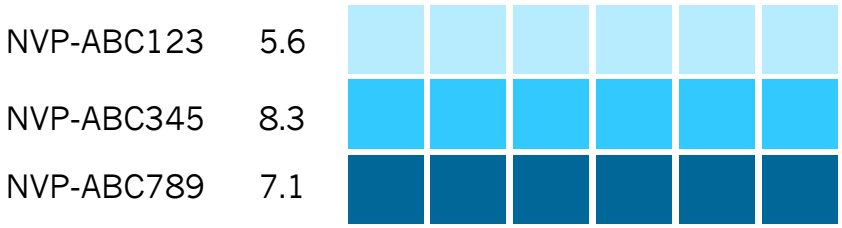
Combine all counts for all possible fragments into a vector of numbers = hologram



Reduce length of vector by folding = reduced hologram

NOVARTIS

Lead optimisation – a pragmatic approach



PLS
Partial **L**east **S**quares

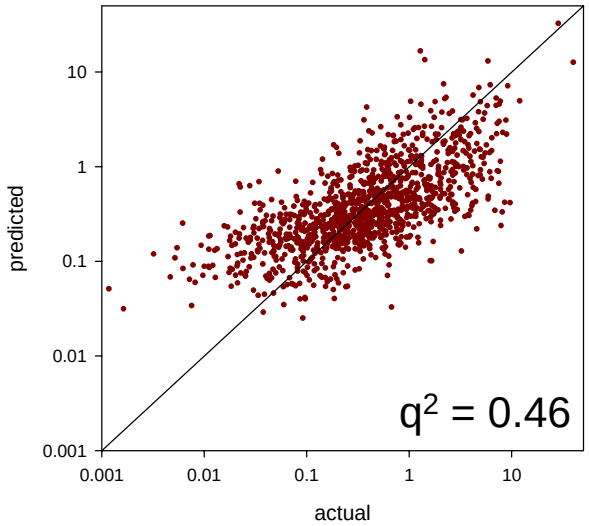
PLS:
Multiple linear regression model suitable for cases with fewer observations than descriptors.

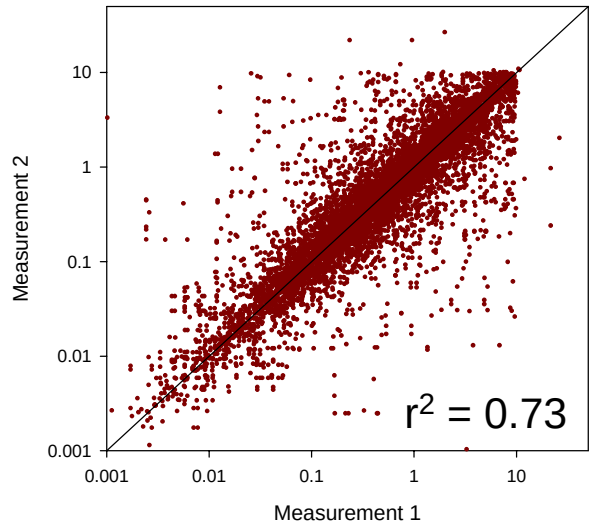
HQSAR-model:

Prediction = f(reduced hologram)

Lead optimisation – a pragmatic approach

- HQSAR model (left: medium/low quality)
  - Potent compounds are underestimated
- Note assay variability as shown on the right





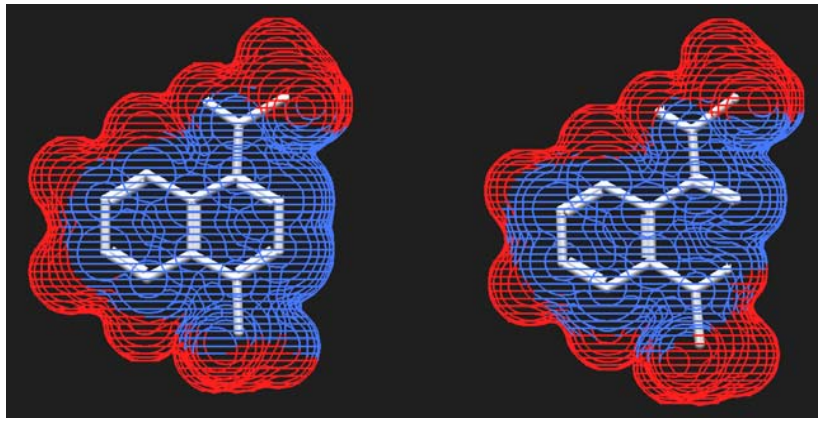
NOVARTIS

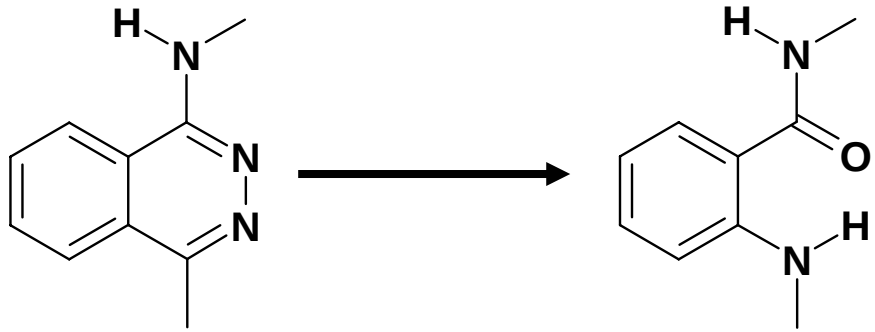
Lead optimisation – a pragmatic approach

- Method is fast and can easily be applied to larger virtual libraries
- Using Novartis in-house data (about 500 assays), we showed that using HQSAR
  - for 35% of assays we get $r^2_{pred}>0.5$ and for
  - for 55% of assays we get $r^2_{pred}>0.4$

Lead optimisation – the more rational approach

- Protein-ligand interactions do not care about the specific molecular structure
- The receptor only recognises the electron cloud, rather than the bonding structure, allowing us to propose new chemotypes
- But how can QSAR capture this?

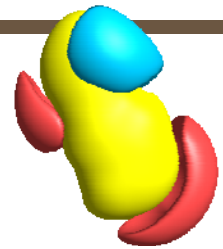


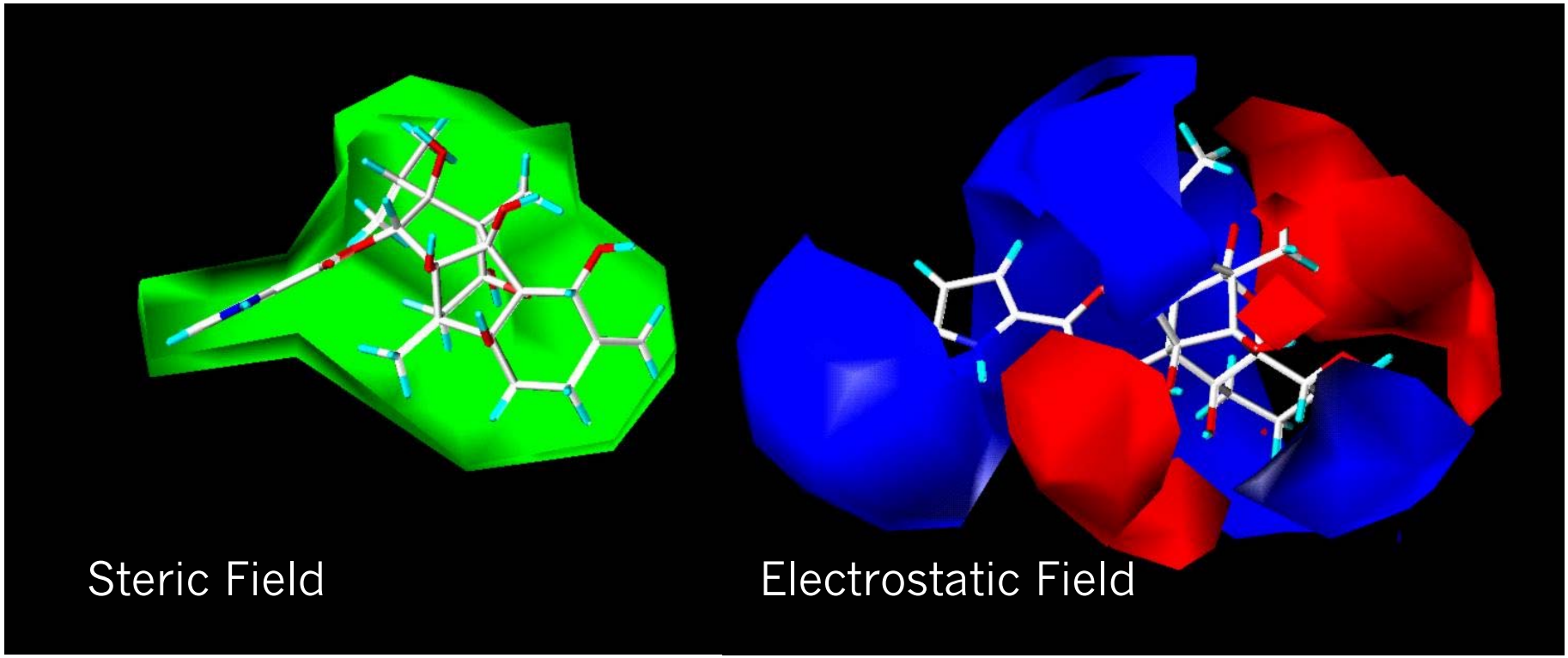


NOVARTIS

CoMFA

- Comparative Molecular Field Analysis
- What is a molecular field?

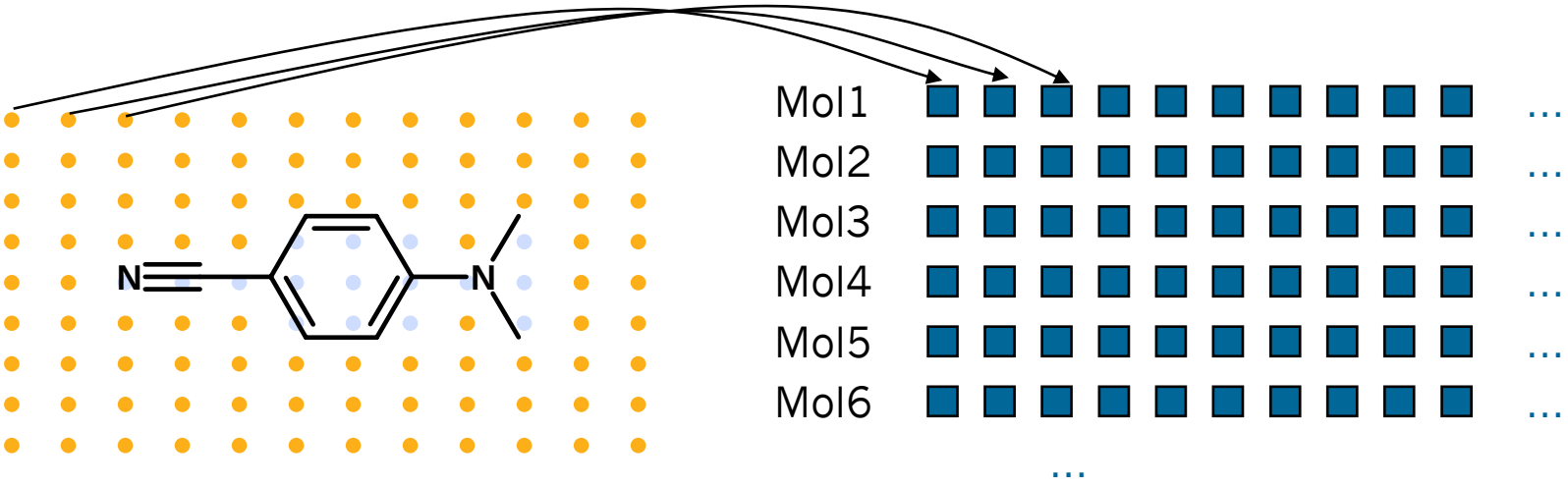




NOVARTIS

CoMFA

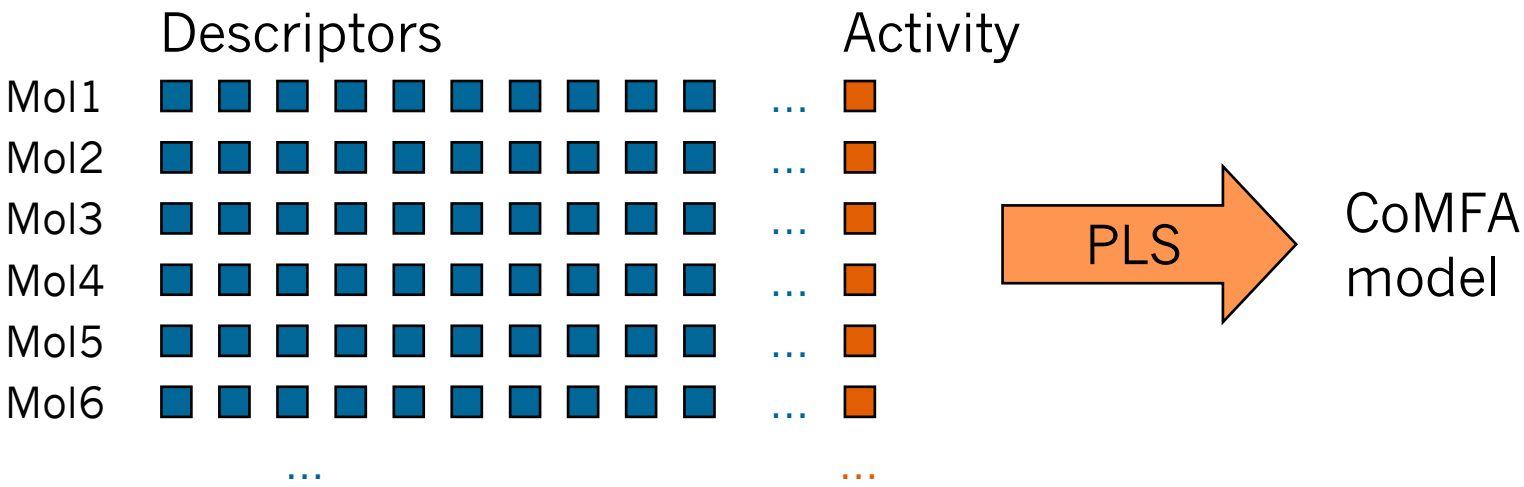
- Align all molecules
- Calculate molecular field at discrete points in space around each molecule individually



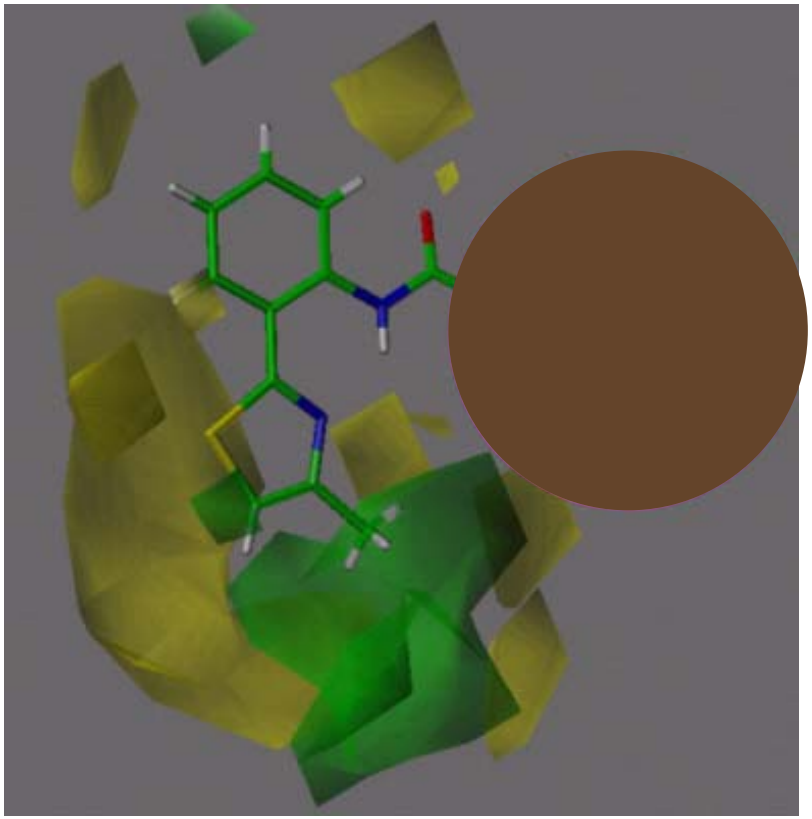
NOVARTIS

CoMFA

- Large number of highly correlated descriptors
- Number of descriptors much larger than number of compounds
- Determine regression with activity using Partial Least Squares (PLS) method

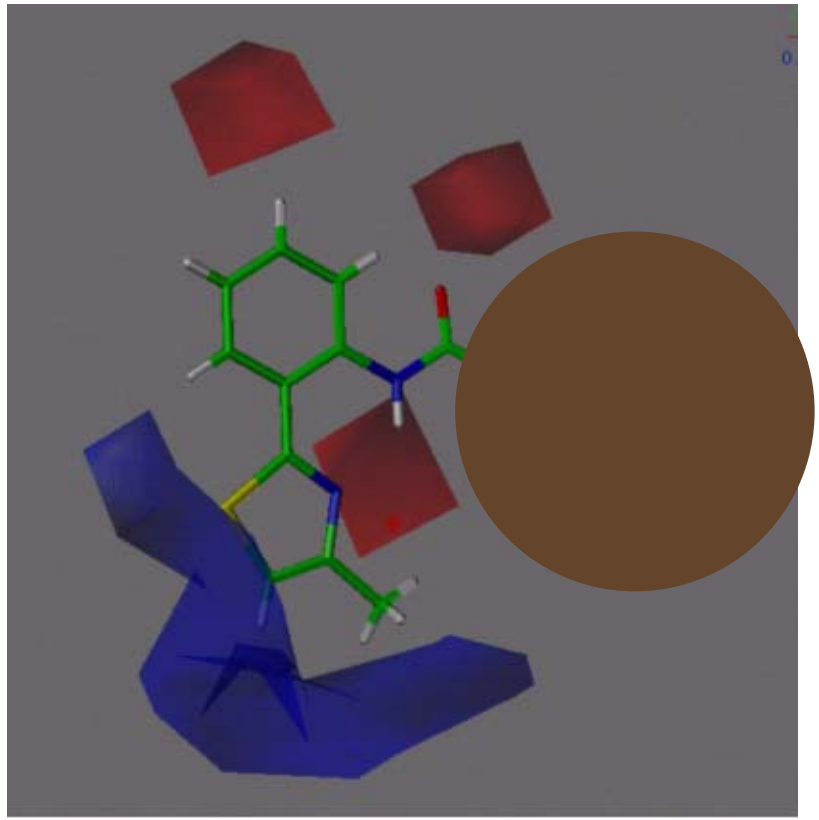


CoMFA



Steric Probe:

Green steric bulk favoured
Yellow steric bulk disfavoured



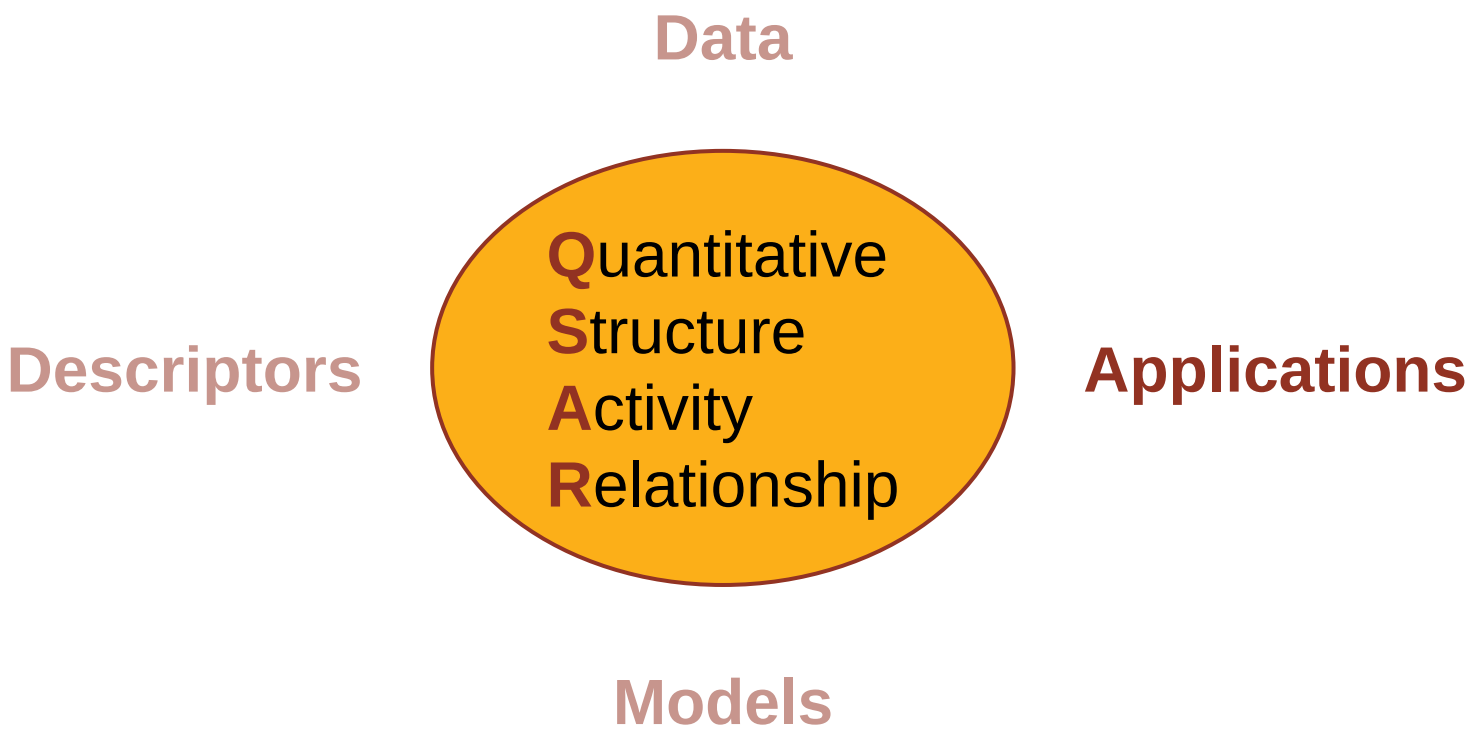
Electrostatic Probe:

Red electronegative field favoured
Blue electropositive field favoured

CoMFA

- CoMFA model identifies points in space that most significantly determine activity
  - It will only learn what controls levels of activity
  - If you don't test a possible interaction, it will not highlight it as positive or negative (e.g. hinge binder in kinases)
- Analysis of CoMFA models allows to suggest regions for modifications
  - For example where changing the steric bulk, would increase or decrease binding
- Structural alignment of compounds is crucial
  - It can impact the interpretation of model

QSAR – definition



Example: Modeling log*BB* (Blood-brain barrier)

Applications: QSAR in lead optimisation and lead finding

QSAR in Lead finding

Lead optimisation

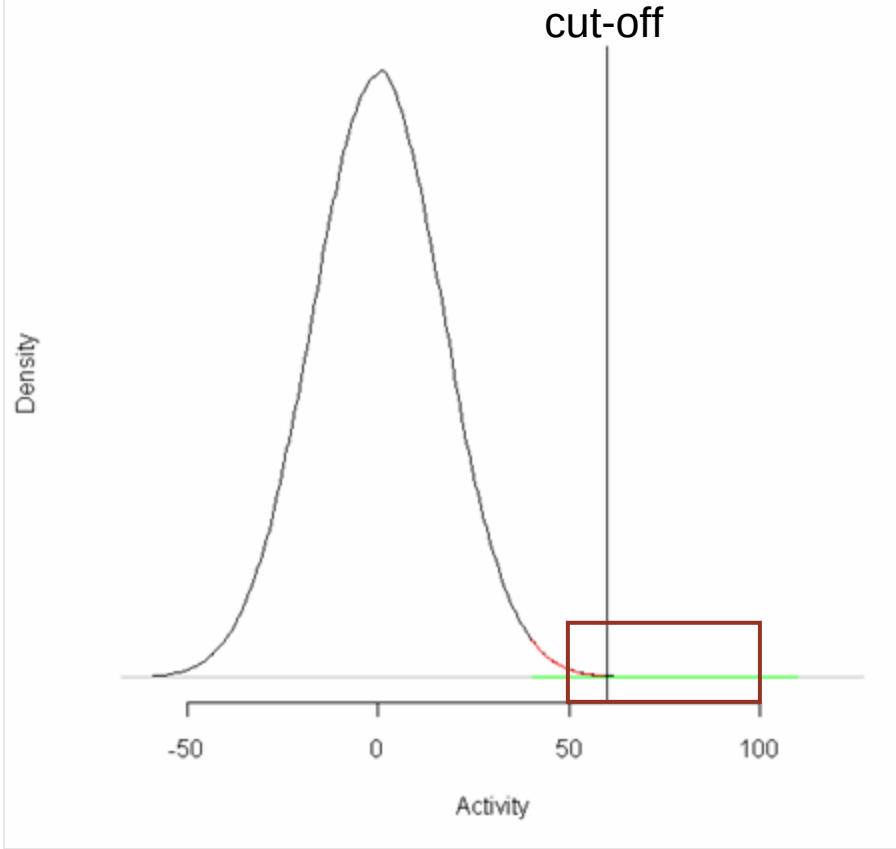
- Few data points (100s)
- High quality assay
- Biased towards actives

Lead finding (HTS)

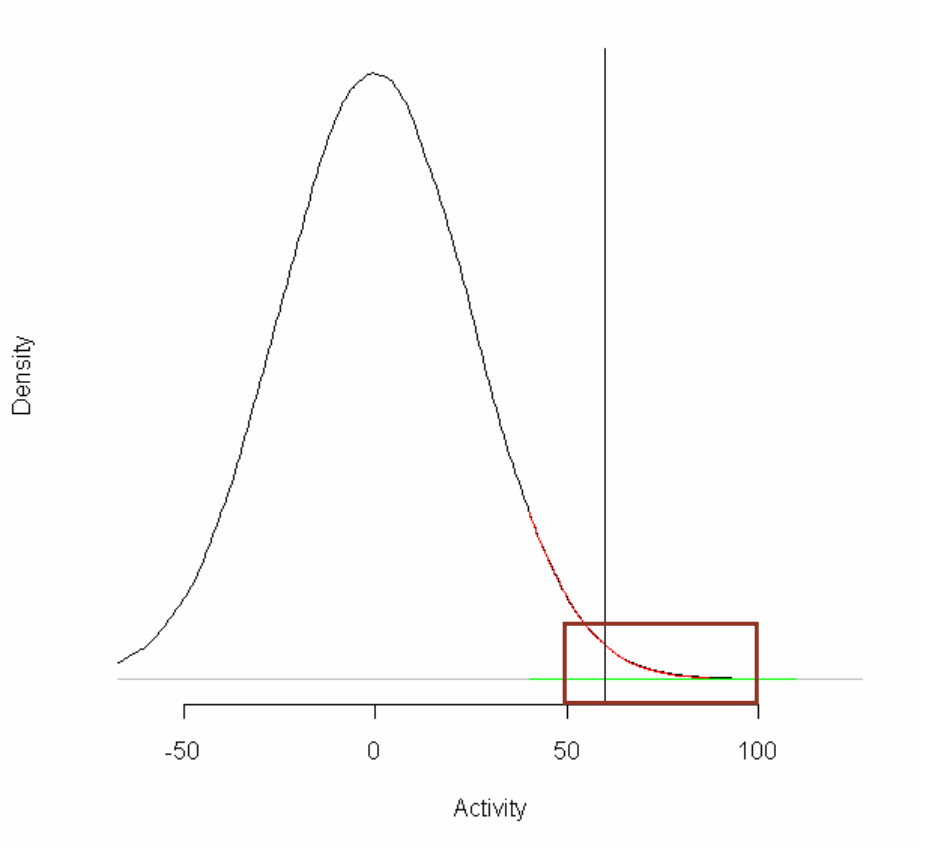
- Huge number of data points (1.000.000s)
- Low quality assay
- Majority of compounds is inactive (0.1-1% hit rate)

The problem in lead finding

- 20 % false positives
- 0.1% hit rate



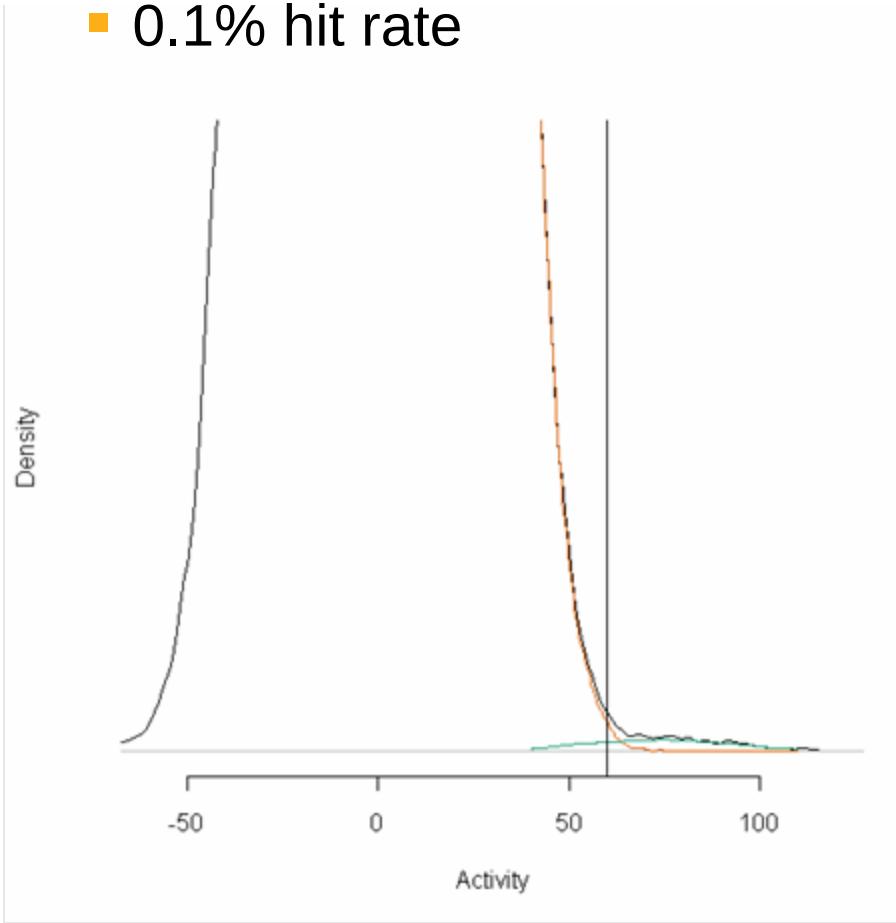
- 90% false positives
- 1% hit rate



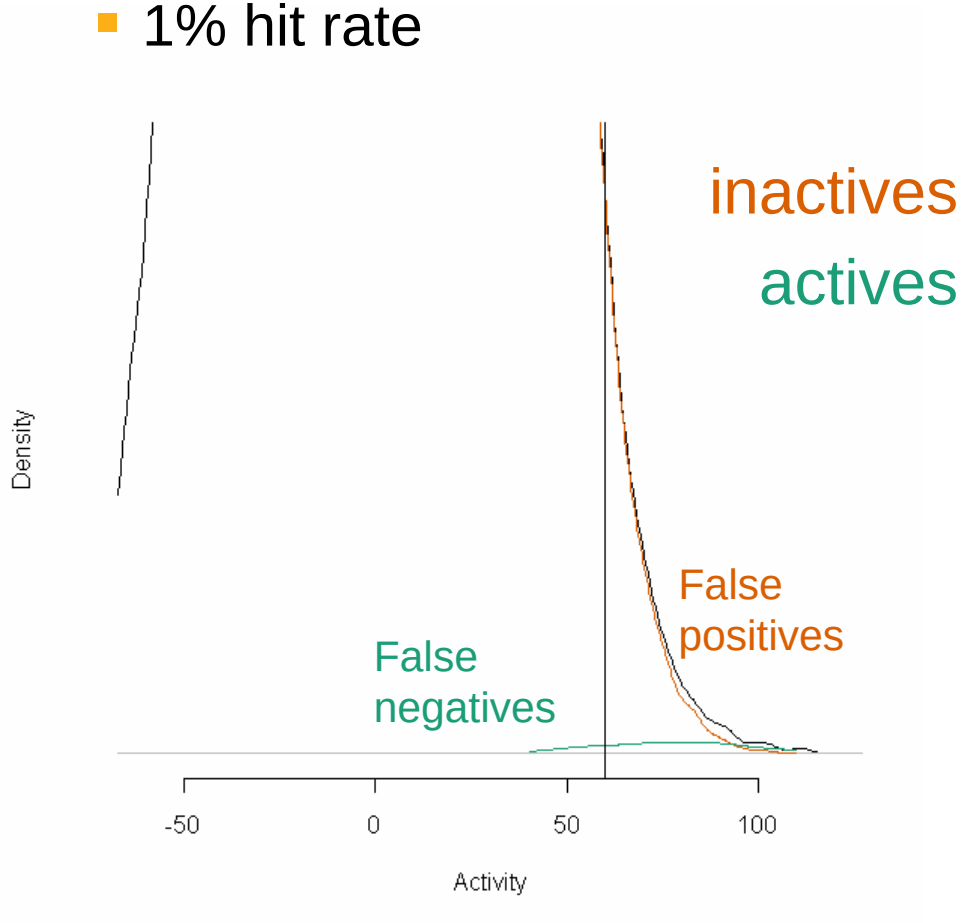
NOVARTIS

The problem in lead finding

- 20 % false positives
- 0.1% hit rate



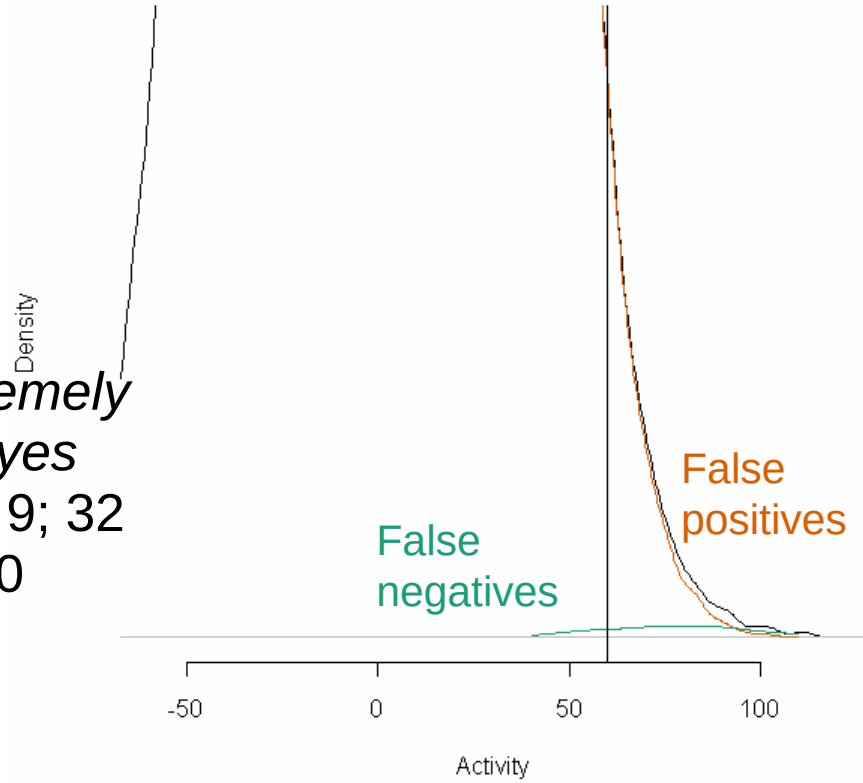
- 90% false positives
- 1% hit rate



NOVARTIS

The problem in lead finding

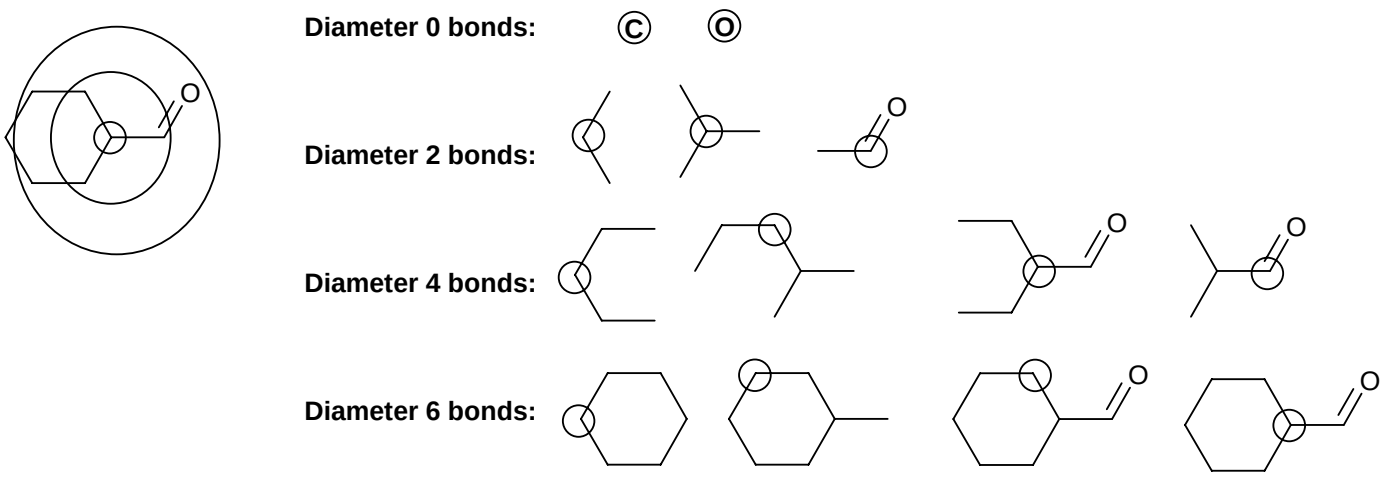
- Can we reduce the number of false positives and rescue false negatives?
- A very pragmatic approach:
  - Build a classification model to predict compounds to be active or inactive
  - Glick M, et al. *Enrichment of extremely noisy HTS data using a naïve Bayes classifier*. *J Biomol Screen* 2004; 9; 32 DOI: 10.1177/1087057103260590



NOVARTIS

Extended connectivity descriptors

- Extended connectivity descriptors (ECFP)
  - aka circular substructure fingerprints
  - Implementation in PipelinePilot (SciTegic / Accelrys)
- Fragment based descriptor scheme
  - Atoms and their environment



NOVARTIS

Naïve Bayesian Classifier

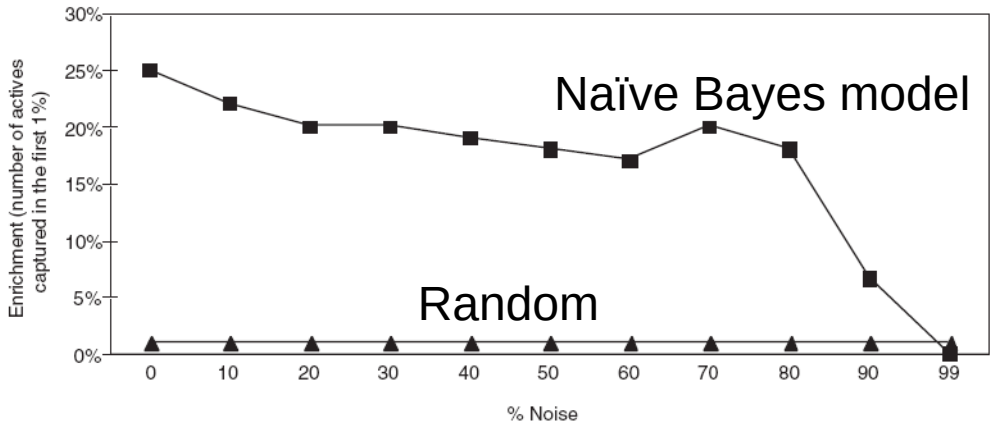
- Simple classifier based on Bayes' theorem

$$p(C|F_1,\ldots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$$

- Probability that compound with features F_i belongs to class C is determined by the product of probabilities to find Feature F_i when it is in class C
- The features used in this study are the occurrence of the extended connectivity fragments

Results

- Classification models successfully improved hit rate
- Amazing robustness of models against noise levels

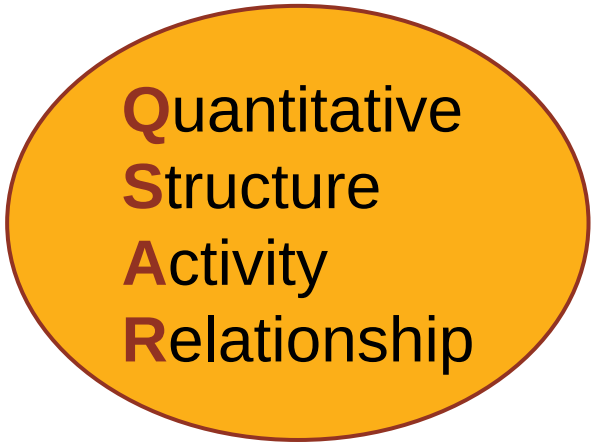


- Similar publications on the use of Naïve Bayes classifier to improve virtual screening / high-throughput docking results

Summary

Data

Descriptors



Applications

Models

Example: Modeling log*BBB* (Blood-brain barrier)
Applications: QSAR in lead optimisation and lead finding

Summary

- Data quality
  - It’s important to understand your data and their quality
- Descriptors
  - Correlation means not necessarily causation
- Methods
  - No methods works for all cases
  - Validation is an important aspect of model building
- Applications
  - Activity and property prediction (best when rational QSAR)
  - Assistance in compound and library design
  - Understanding of changes required to gain good properties
  - Lead finding
- Importantly we must not be surprised to fail

Acknowledgments

- Richard Lewis
- Peter Hunt
- Katrin Spiegel
- Lilya Sviridenko
- Michael Carter
- Meir Glick
- John Davies