QSAR:
an introduction

Wim Dehaen, 30.1.2024

7ADD, Olomouc

Short introduction: Who Am I

- Researcher in D. Svozil's group (UCT Prague)
  - Applied cheminformatics
- Researcher in P. Perlikova's group (UCT Prague)
  - Computational support for medicinal chemistry, mainly SBDD
- My interests:
  - Cheminformatics
  - Medicinal chemistry and CADD
  - Organic chemistry
  - Digital signal processing (of audio)
  - ...

Overview

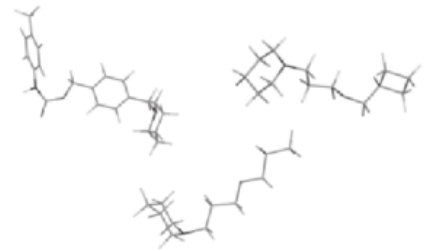
- What is QSAR?
- The history of QSAR
- Structure activity relationships
- Data and featurization in QSAR
- QSAR methods
- QSAR validation
- Applicability domain and interpretability
- QSAR applications

What is QSAR

- QSAR is:
  - Quantitative Structure Activity Relationships
    - QSOR (odorants, olfactory chemicals)
    - QSPR (nonbiological properties, e.g. max. absorption and emission wavelength)
  - Medicinal chemistry + Physical organic chemistry + Statistics (my opinion)
  - "An application of data analysis methods and statistics to developing models that could accurately predict biological activies or properties of compounds based on their structures" - A. Tropsha
  - Predicted biological activity = Function (structural features) + error

What is QSAR

- Overview:
  - Data
    - Data source (molecular structure + end points)
    - Regularization of data
    - Featurization
    - Regularization of features
  - Model-building
    - Appropriate choice of model
    - Appropriate data splits
  - Evaluation
    - Metrics
    - Interpretation
    - Applicability Domain
  - Application
    - Virtual screening
    - Hit optimization
    - ADME(T) filters



Data set of molecules

Curated data set

Calculation of molecular descriptors

Descriptor	MW	Log*P*	ZM1	Mor01m	HOMO
1	171.32	1.11	50	95.1	−8.964
2	185.35	1.58	54	111.94	−8.971
3	199.38	1.97	58	130.144	−8.959
4	213.41	2.38	64	149.713	−8.960
5	227.44	2.77	68	170.646	−8.948
⋮	⋮	⋮	⋮	⋮	⋮
105	287.44	1.98	106	274.86	−8.982

Preprocessing of molecular descriptors

Data set division (train and test sets)

Feature selection

0010010010
0101001001
1001110010

1100010101

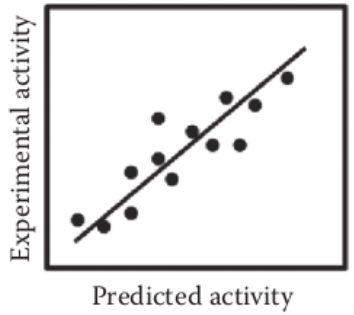
GA
PLS
PCA

QSAR modeling

Y
Response variable

X
Molecular descriptors

Internal and external validation



Selecting the best QSAR model

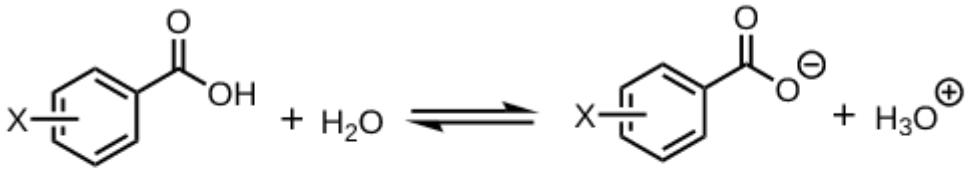
Source: "Quantitative Structure – Activity Relationship: A Practical Approach"

What is QSAR

- Machine learning:
  - QSAR is a form of "traditional" supervised machine learning
    - Non-deep:
      - Small data regime
      - Interpretability
  - "Deep QSAR"
    - Seems mostly a semantic distinction between deep QSAR and deep learning for molecular property prediction
- Virtual screening:
  - QSAR models can be applied for virtual screening:
    - Property filters (solubility, aggregation, toxicity)
    - Activity
  - But Applicability Domain issues!

The history of QSAR

- Legendre, Gauss (1805):
  - Method of least squares used to determine orbits of spatial bodies based on astronomical observations
- Crum-Brown and Fraser (1865):
  - Relation between structure and physiological action
- Hammett equation (1935):
  - Empirically derived substituent constants derived from substituted benzoic acid hydrolysis rate



$$\log\left(\frac{k}{k_H}\right) = \sigma\rho \quad \text{or} \quad \log\left(\frac{K}{K_H}\right) = \sigma\rho$$

σ = substituent constant
ρ = reaction constant

159 years ago!

Although we cannot obtain a rational explanation of the connection between the chemical and physiological characters of a substance until we know more of the *modus operandi* of poisons, it might be supposed that a careful examination and comparison of known facts would lead to the discovery of some empirical law or laws by means of which we could deduce the action from the chemical constitution. Unfortunately, however,

The history of QSAR

- Hantsch equation

$$\log \frac{1}{C} = k_1 D_1 + k_2 D_2 + ... + k_i D_i$$

with k_i and D_i the ith constant and descriptor respectively

- Does not have to be linear, for example Hantsch famous plant hormone work had this form:

$$\log \frac{1}{C} = -k_1 (logP)^2 + k_2 (logP) + k_3 \sigma + k_4$$

The history of QSAR

- Free-Wilson formalism
- Biological activity (log scale) is determined by the sum of the activity of the reference compound and substituent contributions
- With μ the biological activity of the unsubstituted analog and sum(a_i) the sum of substituent contributions.

$$\log \frac{1}{C} = \sum a_i + \mu$$

The history of QSAR

- CoMFA
  - Comparative Molecular Field Analysis
  - "The first 3D QSAR"
  - Shape,electrostatic,H-bonding, … 3D features
  - placing aligned conformations in grid and probing with e.g. lipophilic probe
  - Resulting model is somewhat like a pharmacophore in that it captures spatial distributions of features

The history of QSAR

- OECD and QSAR use in regulatory context
  - Hazard assessment of chemicals
  - OECD QSAR toolbox
  - e.g. Genotoxicity
  - e.g. Biodegradation
  - e.g. Skin sensitization
  - Intended use: filling gaps in (eco)toxicity data

OECD QSAR principles

- A defined endpoint
- An unambiguous algorithm
- A defined domain of applicability
- Appropriate measures of goodness-of-fit, robustness, and predictivity
- A mechanistic interpretation, if possible

Structure activity relationship

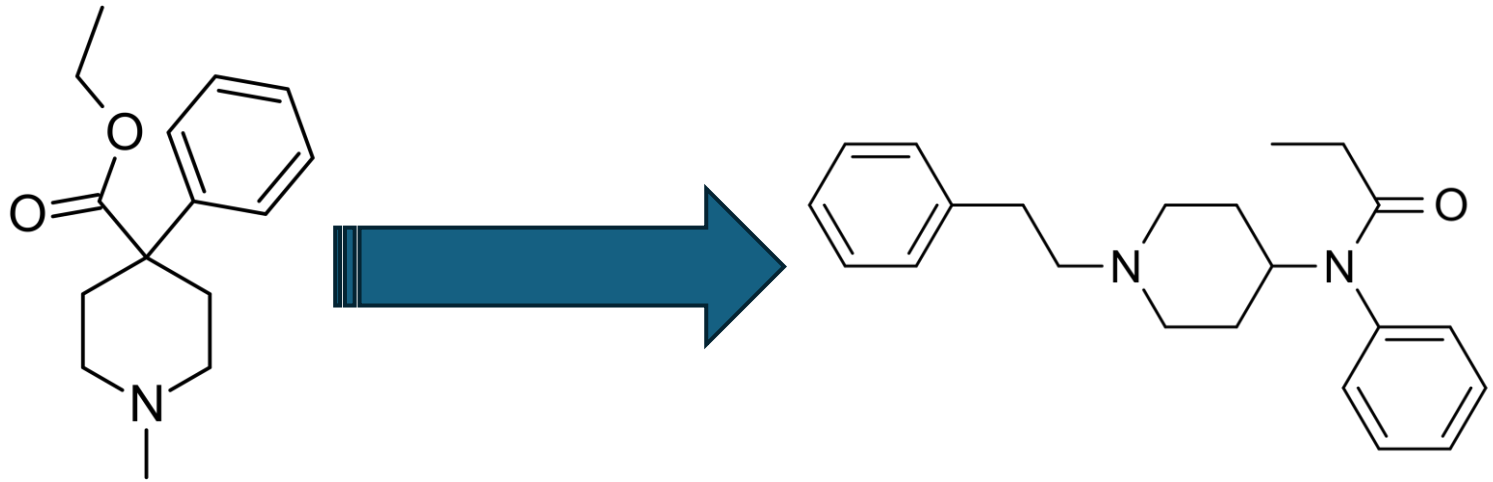
- Very important concept in medicinal chemistry
- Intuitive and ad hoc (non quantitative!)
- "An understanding of the SAR for a set of molecules allows one to rationally explore chemical space, which in the absence of “sign posts” is essentially infinite" R.Guha, In Silico Models for Drug Discovery, Methods in Molecular Biology, vol. 993

Structure activity relationship

- Who recognizes this man?
- Voted "3rd Greatest Belgian"
    1. Father Damian (Catholic missionary)
    2. Eddy Merckx (Cyclist)
    3. ???

Paul Janssen of "Janssen" fame

- Brought 80 drugs to market
  - Fentanyl
  - Haloperidol
- Emblem of the trial-and-error analog exploration SAR approach taken by medicinal chemists
- Fentanyl is the result of SAR exploration and optimization of meperidine

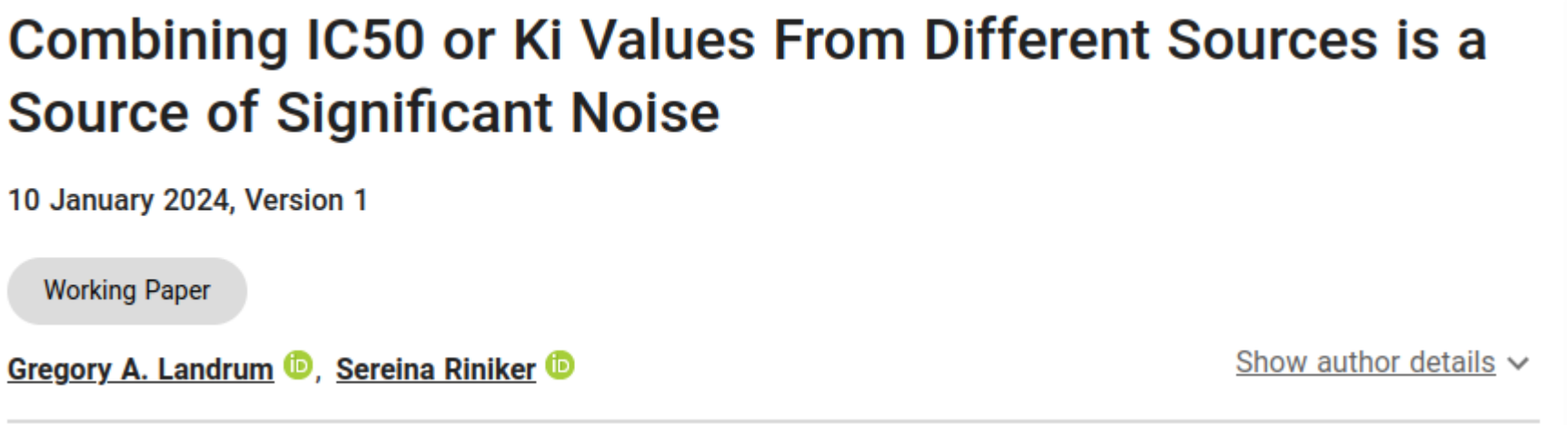


Data source

- Publicly available:
  - ChEMBL
  - PubChem
  - Community benchmarks such as SAMPL
  - Ad hoc data sets in the literature
- Commercial databases:
  - Reaxys
  - Scifinder
    - "Please do not use our product in this way" - Scifinder Rep when i asked them if I can use the result of Scifinder Queries in open source cheminformatics workflows

Data cleanup

- Endpoints (e.g. Assay data)
  - Experimental noise
  - Annotation inconsistencies and errors
  - Fundamental incompability between different measurements (e.g. different assay conditions)
  - Some publications more trustable than others
    - (e.g. HTS hits are more noisy and have more chance to have assay artefacts)



Combining IC50 or Ki Values From Different Sources is a Source of Significant Noise

10 January 2024, Version 1

Working Paper

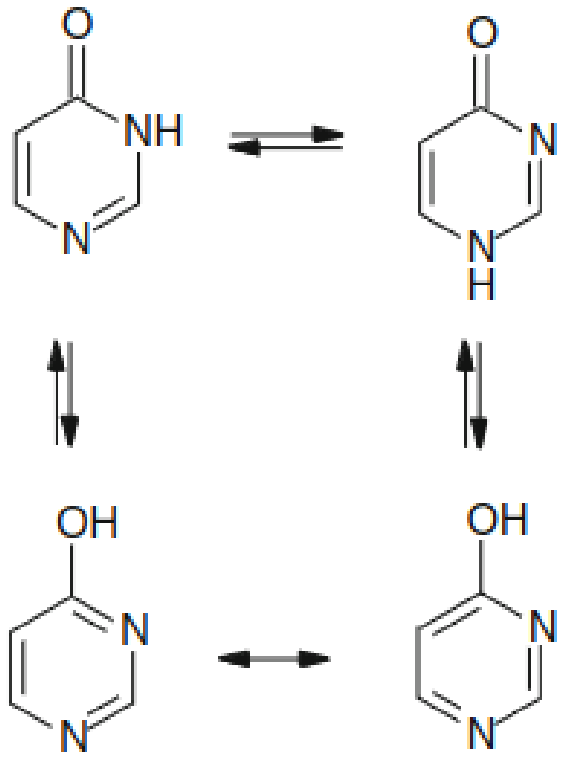
Gregory A. Landrum, Sereina Riniker

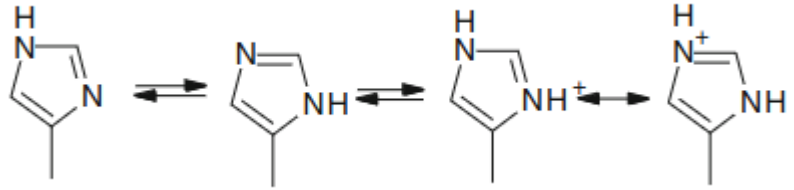
Show author details

Data cleanup

- Molecules:
  - Molecular identity and normalization (tautomers, protonatino, kekulization, charges, salts, …)
  - Solvent, pH, T dependent
  - Parsability in software (some SMILES will load in obabel but not RDKit abd vice versa)
  - Undesired properties (e.g. atoms that forcefield can't deal with)
  - In some case, chemical entities can also be mixtures, reactions, specific conformations, …

Some examples





Formally tautomers, but won't interconvert:

Source: So you think you understand tautomerism?

Featurization of molecules

- Descriptors:
  - Graph invariants for a chemical graph
    - Take into account chemical graph is colored, weighted, non-directed and has extra data for each vertex like charge, hybridization status, …
  - Should be invariant to:
    - Vertex order of the graph
    - But also rotation and translation (if vertices have 3d coordinates included)
  - "There are three keys to the success of any QSAR model building exercise: descriptors, descriptors, and descriptors" - Alexander Tropsha

Descriptors

- Many, many possibilities
- A **fingerprint** is essentially a list of descriptors resulting from a common process
  - Usually binary vector, sometimes integer, occasionally float
- Binary descriptors:
  - Presence (1) or non-presence (0) of one or several structural motifs
  - Topological pharmacophore
  - Bag of fragments (e.g. a bit in Morgan fingerprint)
  - One specific fragment (a bit in a structural key such as Klekota-Roth FP)
- Integer descriptors:
  - Count of one or several structural motifs:
  - Same possibilities as above, but count occurence
  - Some Lipinski-like descriptors: HBD, HBA, HAC

Descriptors

- Real descriptors:
  - Continuous range
  - Topological indices
    - Zagreb, Kier hall, Randic, Kappa, Wiener, etc
  - Physicochemical descriptors (MW, TPSA, …)
  - cLogP (this is in a sense calculated by a QSPR model itself)
  - QM calculated properties
  - Empirical (e.g. boiling point, NMR shift, partition coefficient, …)

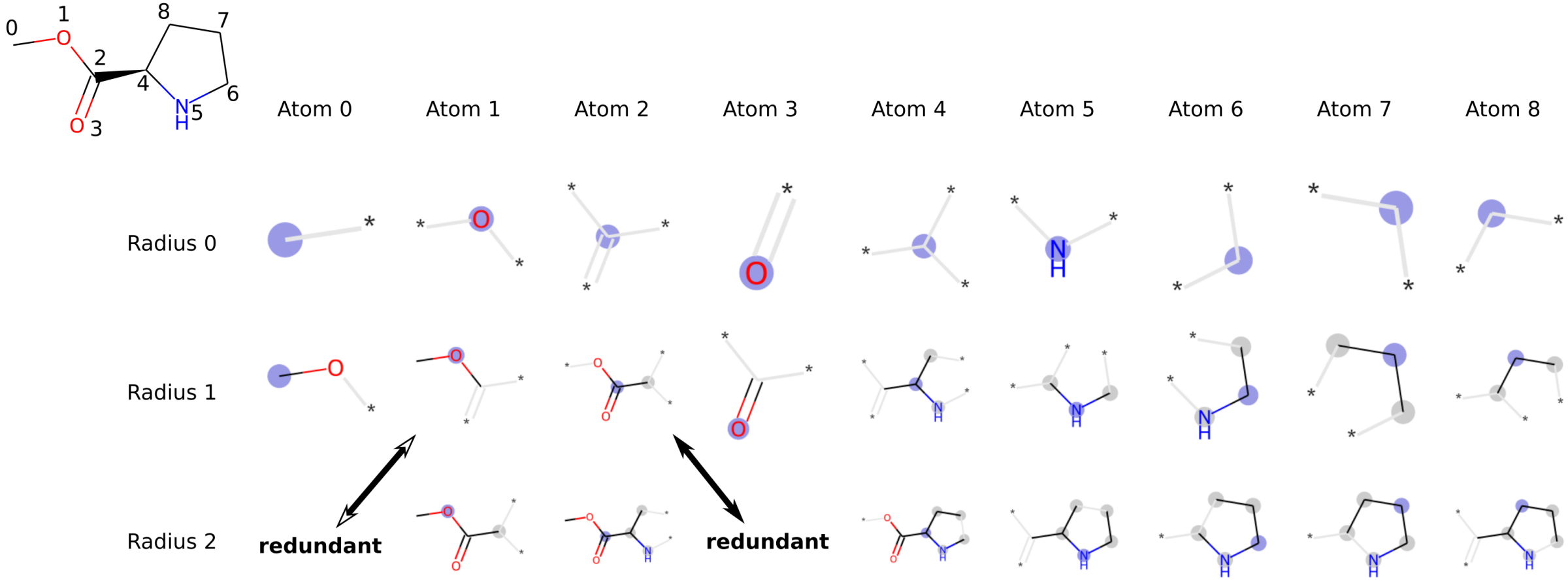
Descriptors

- 1D:
  - Molecular weight, formula, ...
  - "Whenever the underlying biological observables may depend on transport as well as receptor fit, such as passive membrane penetration, “1D’’ descriptors, such as log P, polar surface area, and pKa, should be considered. " - Richard D. Cramer
- 2D:
  - Based on "2D structure" of molecule (actually a dimensionless graph), topological indices, substructure presence, ...
- 3D:
  - Conformation dependent, surface area, volume, shape, QM-calculated descriptors, ...
- 4D:
  - Dynamics, flexibility of bonds, MD-PLIF, ...

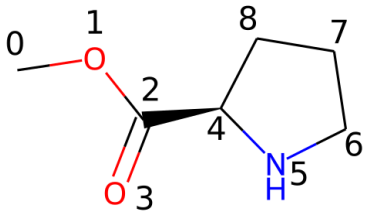
Fingerprint example: ECFP

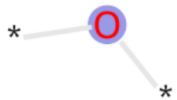
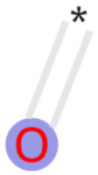
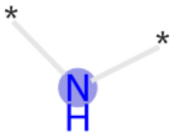
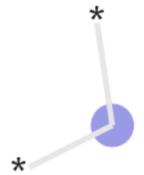
- Extended connectivity fingerprint aka Morgan(like) fingerprint
- Molecular fingerprint based on topological neighborhoods around atoms at a given topological radius threshold
- Commonly used as input in QSAR, highly performant and generic
- Implemented in major cheminformatics packages

ECFP visual introduction

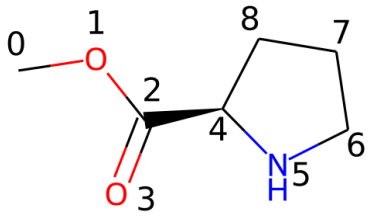


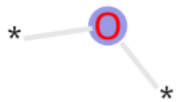
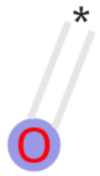
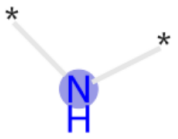
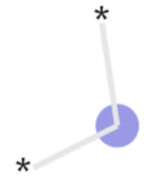
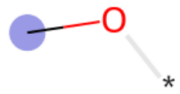
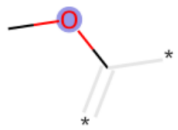
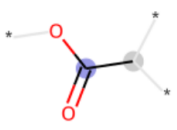
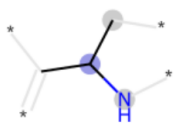
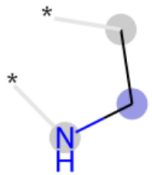
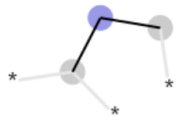
ECFP visual introduction



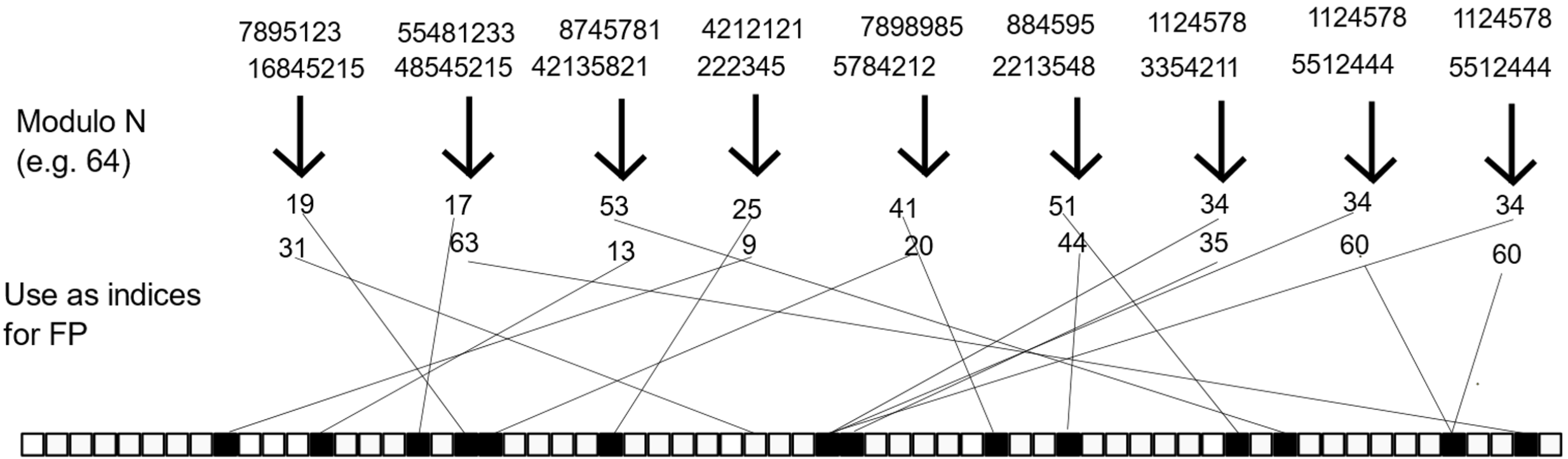
	Atom 0	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6	Atom 7	Atom 8
Radius 0									
Atom properties:	(6,1,3)	(8,2,0)	(6,3,0)	(8,1,0)	(6,3,1)	(7,2,1)	(6,2,2)	(6,2,2)	(6,2,2)
Apply hash function	↓	↓	↓	↓	↓	↓	↓	↓	↓
32-bit integer	7895123	55481233	8745781	4212121	7898985	884595	1124578	1124578	1124578

ECFP visual introduction



	Atom 0	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6	Atom 7	Atom 8
	7895123	55481233	8745781	4212121	7898985	884595	1124578	1124578	1124578
Radius 0									
Radius 1									
Neighbor atoms:	(1,)	(0,2)	(1,3,4)	(2,)	(2,5,8)	(4,5)	(5,7)	(6,8)	(7,4)
hash identifiers e.g. hash(55481233,)	16845215	48545215	42135821	222345	5784212	2213548	3354211	5512444	5512444

ECFP visual introduction



Feature normalization

- Descriptors are often correlated
- Mixing binary, integer and real data
- Some features have no relation with the end point at all
- The fewer features, the faster and more interpretable the model is

Feature normalization

- Scaling descriptors:
  - e.g. using z scores
  - Scale between [0,1] by linearly scaling [min(D),max(D)]
- Collinearity:
  - R^2 between descriptor should be under a given threshold (e.g. 0.8)

Feature selection

- Pruning after model building based on feature importances
  - If method permits this, e.g. RF
- Building models with one or more descriptors excluded
- Principal component analysis

Training a model

- In general, QSAR tends to use "standard machine learning" algorithms and simpler methods like logistic regression
  - As opposed to more complex deep learning based approaches
- Supervised learning (has an endpoint guiding it)
  - Classification:
    - The endpoint is categorical: two or more labels. E.g. Active/Inactive
  - Regression:
    - The endpoint is continuous. E.g. pIC50 between 3.0-9.0
- Unsupervised learning:
  - Clustering:
    - The data set gets divided into two or more clusters
    - Is this really QSAR?

Training a model

- Linear approaches
  - PArtial Least Squares,Multiple Linear Regression, Free-Wilson
- Non-linear approaches
  - "All the rest"
  - See next slide

Typically used approaches

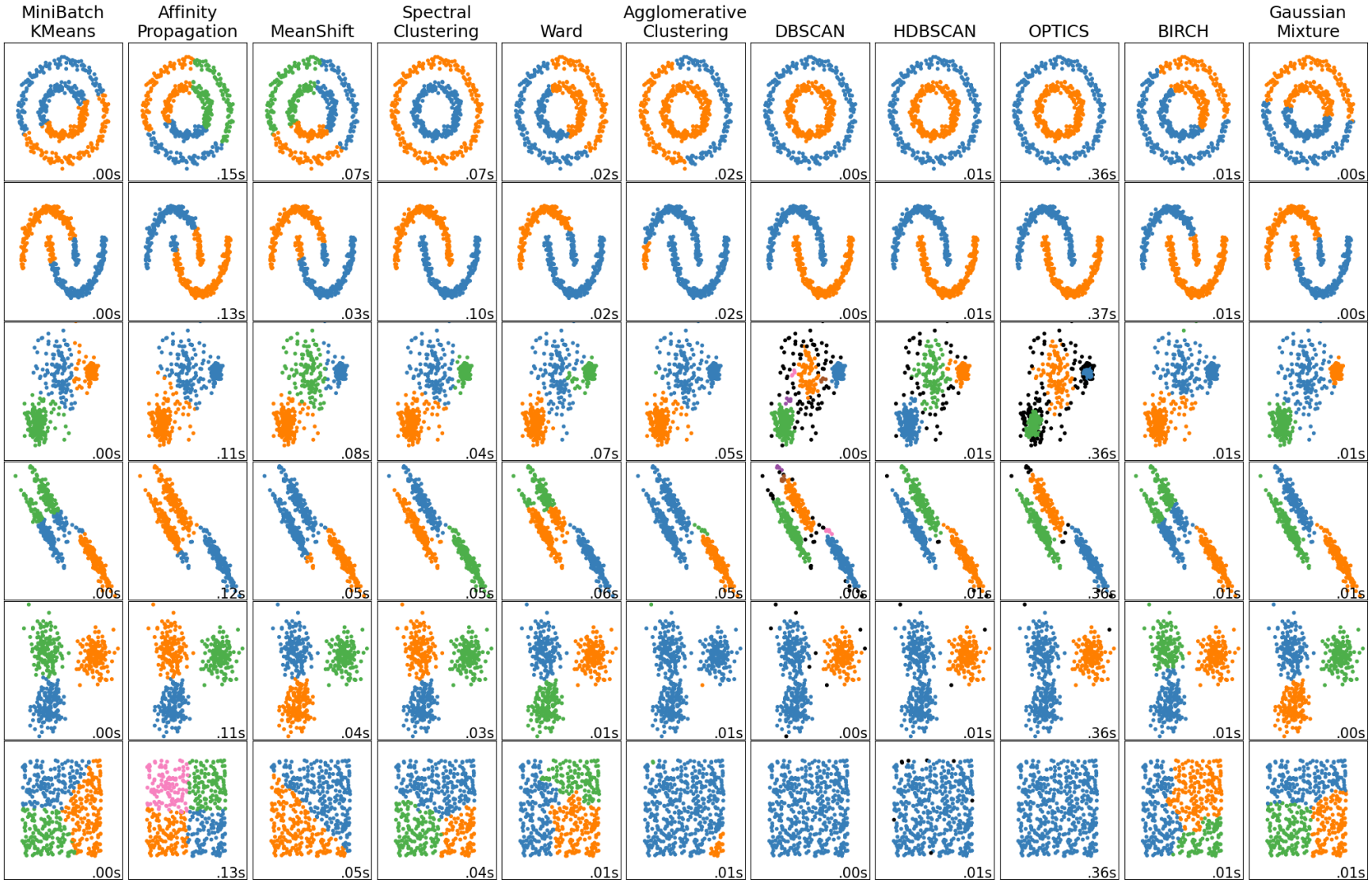
- Regression:
  - Decision tree
  - Support Vector Machine
  - Neural Network
  - Random Forest
  - k Nearest Neighbors
  - Multiple linear regression
  - Partial least squares
  - Gaussian process

Typically used approaches

- Classification:
  - Decision tree
  - Support Vector Machine
  - Neural Network
  - Random Forest
  - k Nearest Neighbors
  - Logistic regression
  - Naive Bayes

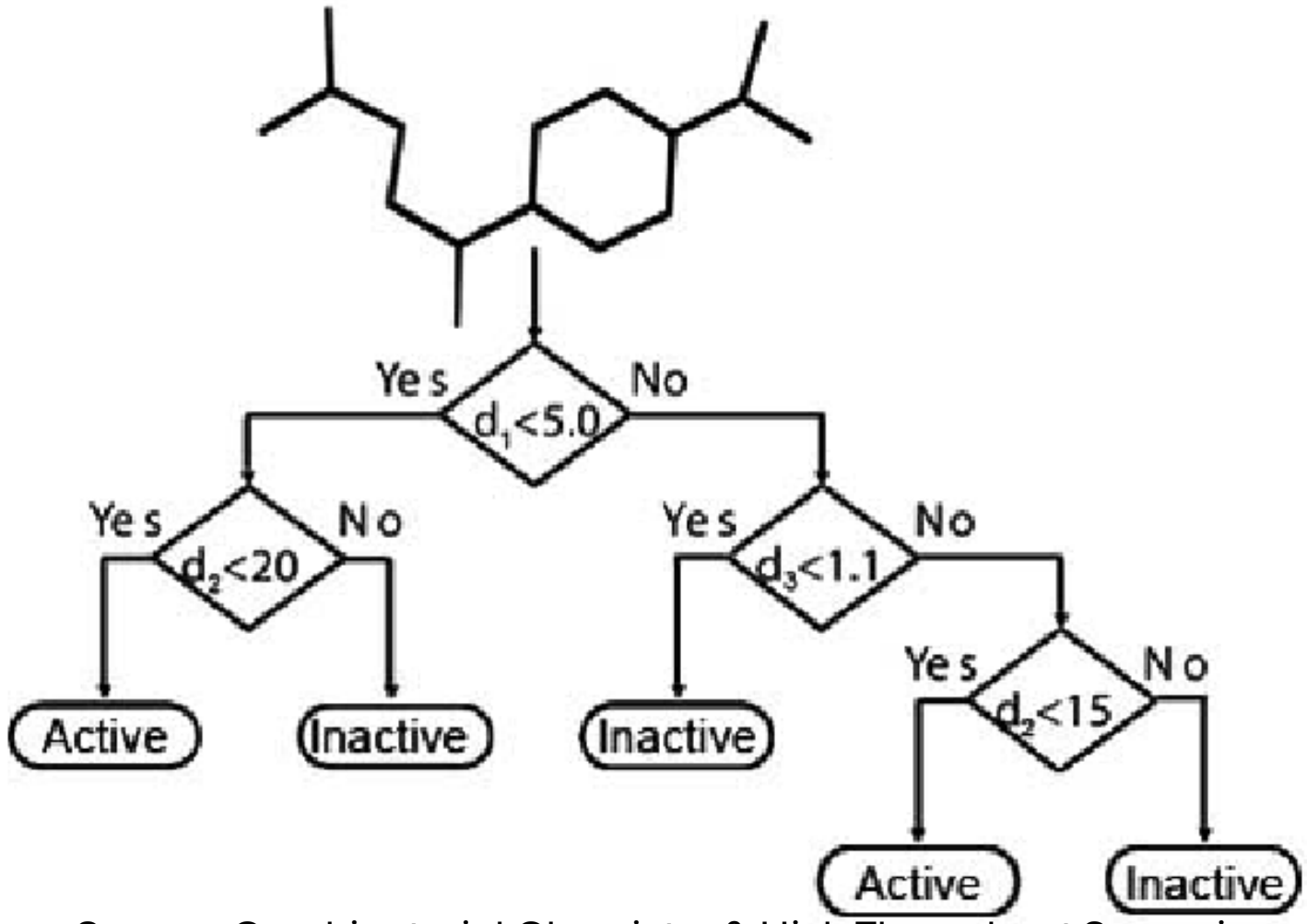
Typically used approaches

- Clustering
- Not really QSAR



Example: Random forest

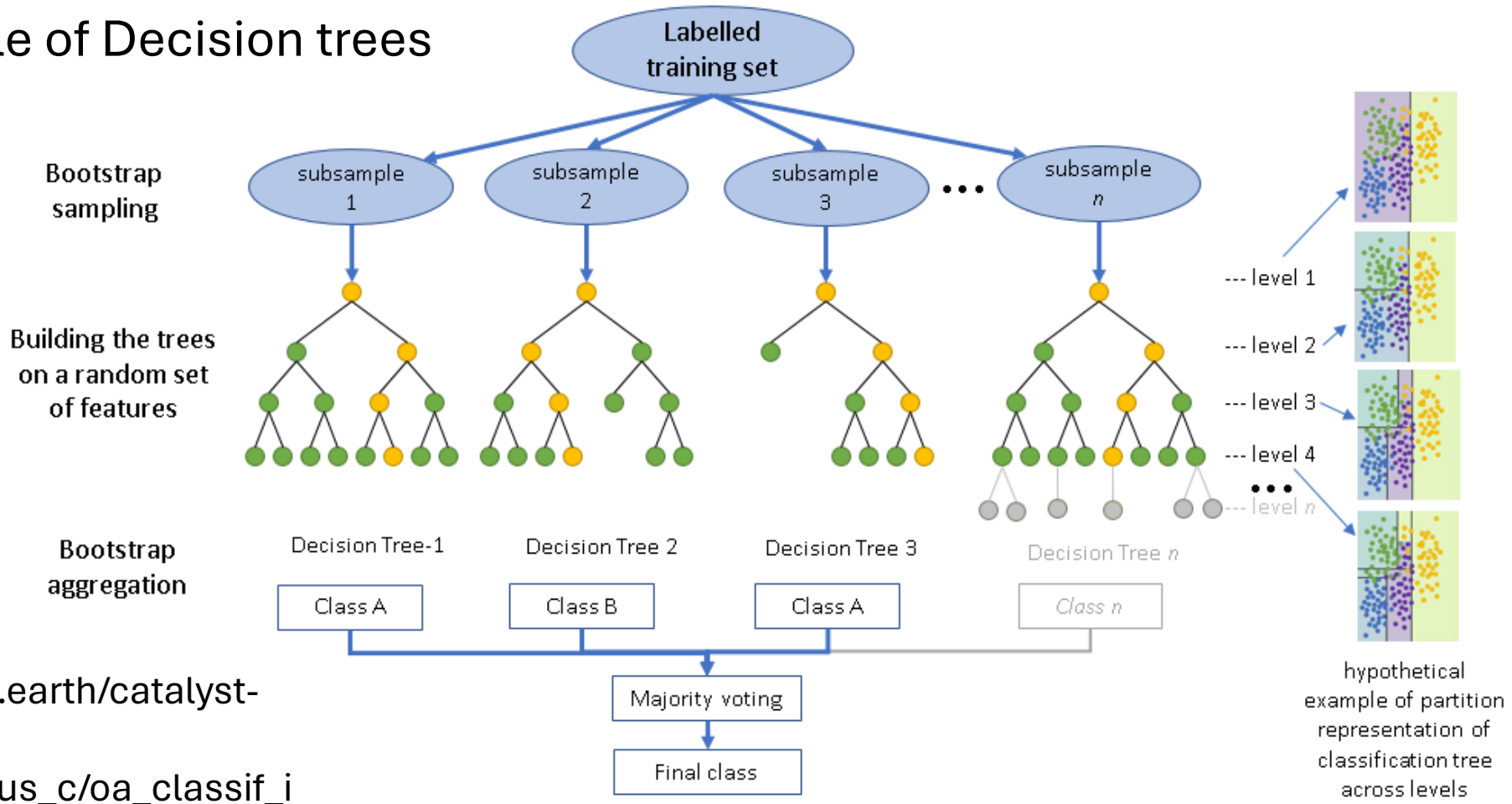
- Decision tree:



Source: Combinatorial Chemistry & High Throughput Screening 9(3):213-28

Example: Random forest

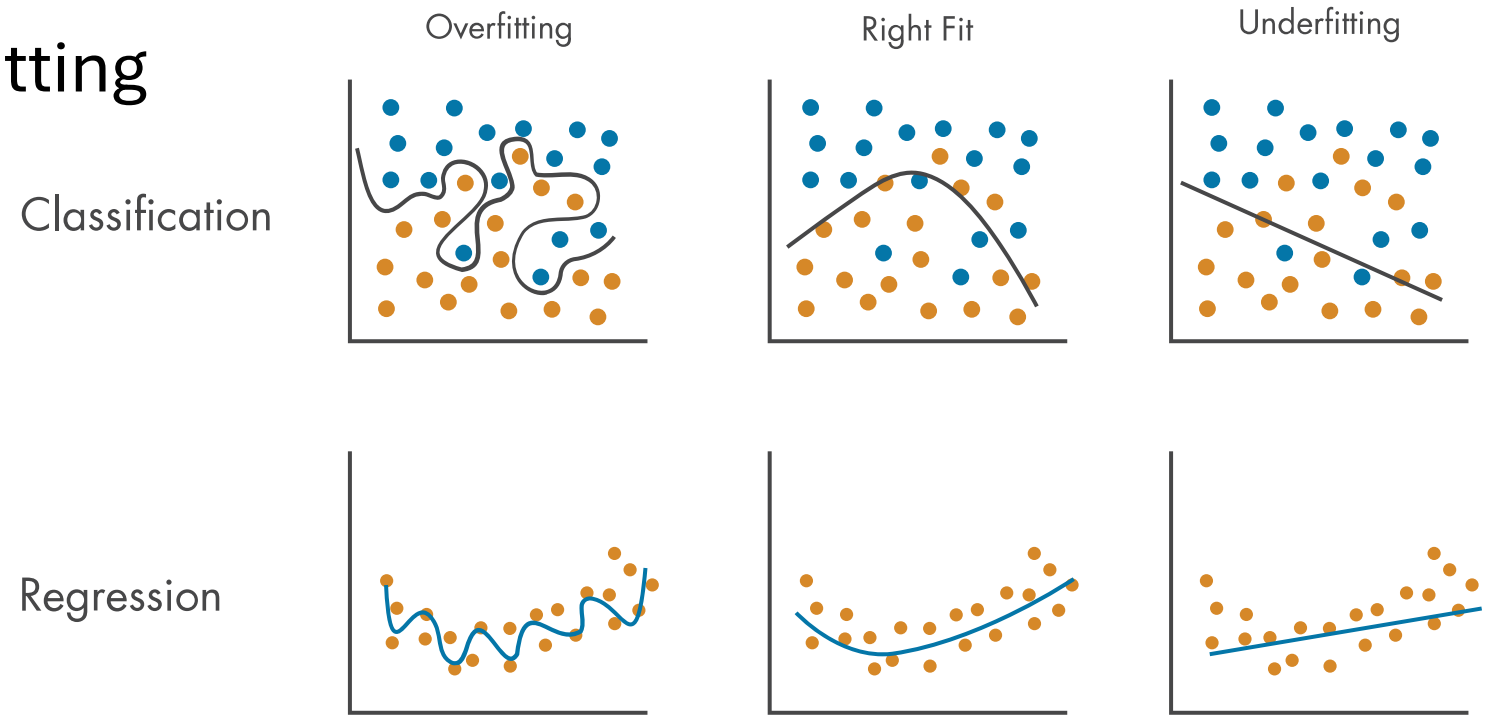
- Random Forest
  - Ensemble of Decision trees



Source: https://catalyst.earth/catalyst-system-files/help/concepts/focus_c/oa_classif_intro_rt.html

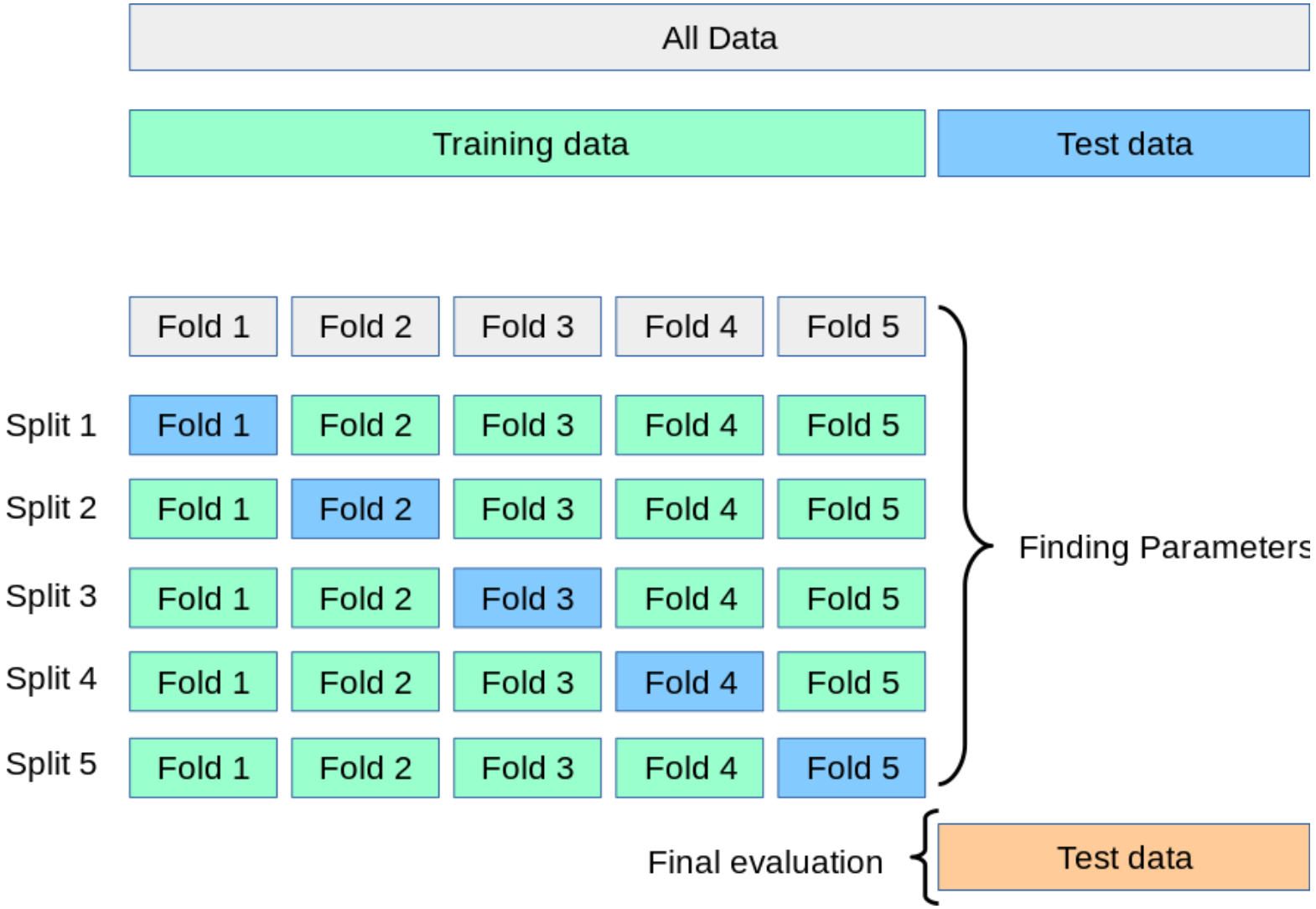
Data splitting

- Why? Over- and underfitting



Source: https://www.mathworks.com/discovery/overfitting.html

Data Splitting



https://scikit-learn.org/stable/modules/cross_validation.html

Data splitting

- Random split
  - This can make tasks too easy: similar scaffolds in test and train
- Scaffold split
  - Shows if model is able to hop between scaffolds (to some extent)
- Time split
  - To mimic actual discovery process

Metrics

- Binary classification metrics:
  - Often calculated from confusion matrix (see next slide)
  - Source of slide: wikipedia

Sources: [11][12][13][14][15][16][17][18][19] view • talk • edit

		Predicted condition			
	Total population $= P + N$	**Predicted Positive (PP)**	**Predicted Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm, overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Metrics

- Regression metrics:
- Q^2 Squared leave-one-out cross-validation correlation coefficient
- R^2 Coefficient of determination
- MAE Mean absolute error
- …

$$Q^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{LOO})^2 / \sum_Y (Y_{exp} - <Y>_{exp})^2$$

$$R^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{pred})^2 / \sum_Y (Y_{exp} - <Y>_{exp})^2$$

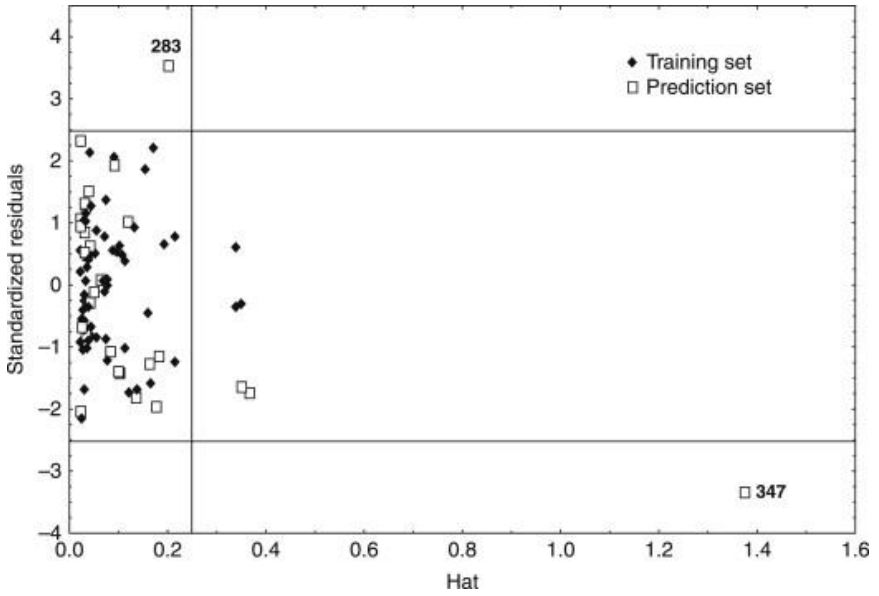
$$MAE = \sum_Y |Y - Y_{pred}| / n$$

Source: Best Practices for QSAR Model Development, Validation, and Exploitation

Interpretability

- Simple models like Hantsch are sometimes naturally interpretable
- Some models are "black boxes" but can be interpreted
- Some models offer feature importances directly
- Some tricks exist for probing explanations (model agnostic):
  - Shapley scores (See the Tutorial) "For the given prediction, how much has each feature contributed compared to the average prediction?"
  - Counterfactuals: “what changes will result in an alternate outcome?”

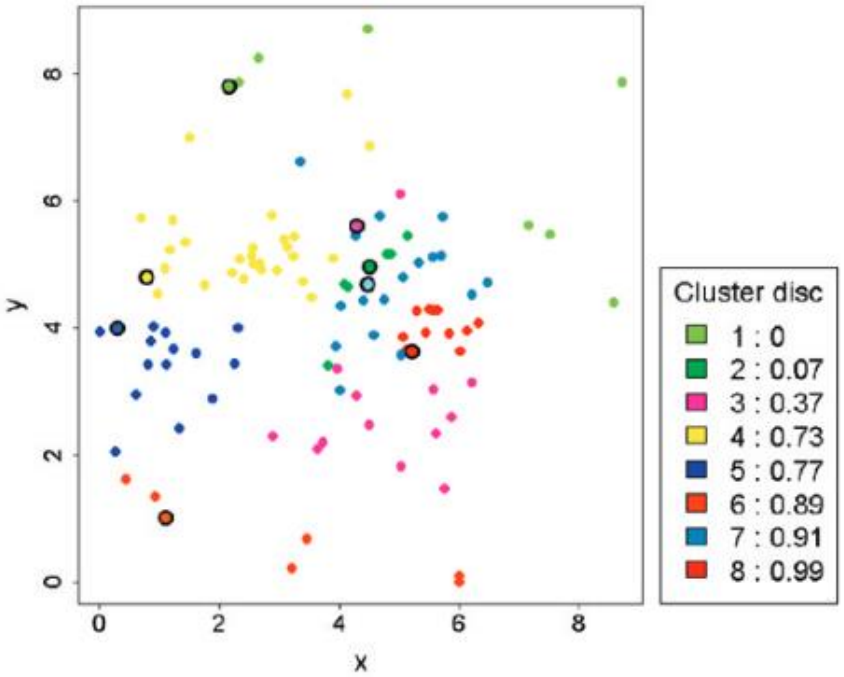
Applicability Domain

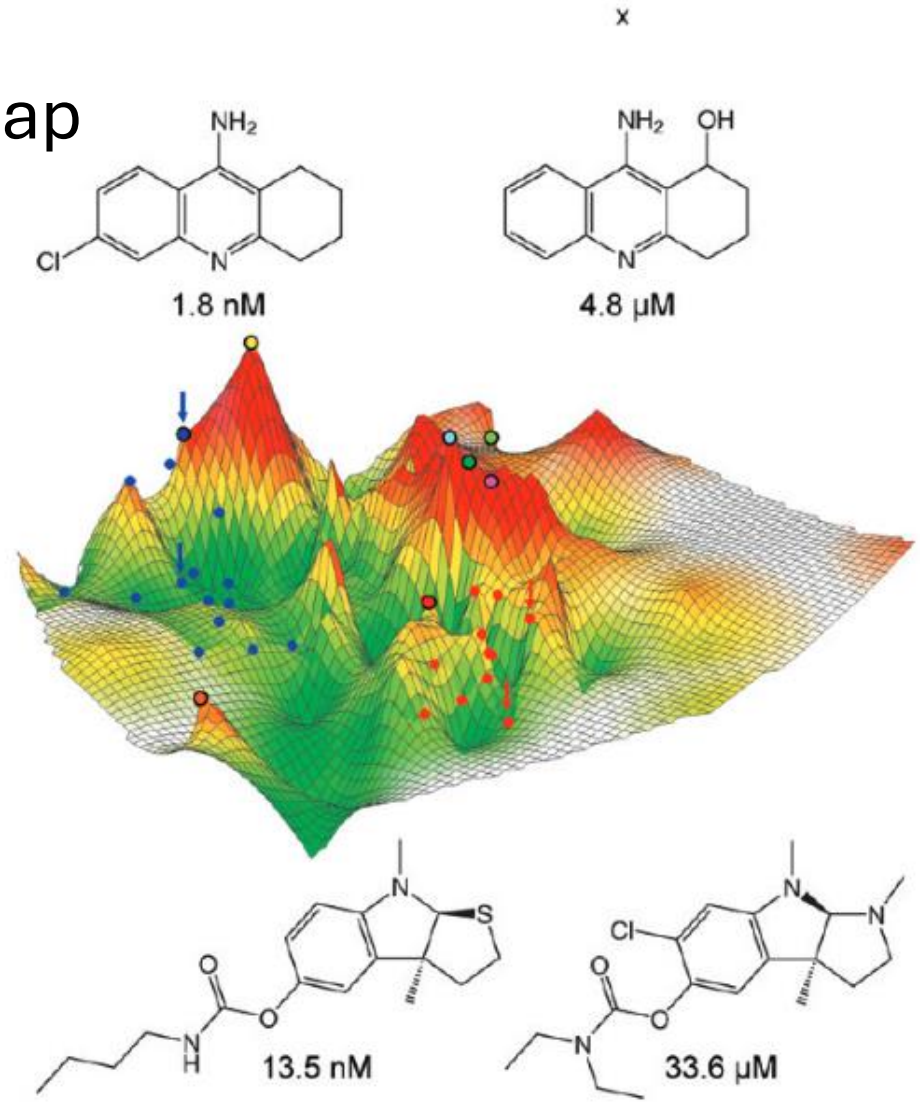


- Several ways of assessing it:
  - Leverage (thresholds on williams plot)
  - Conformal prediction
- In general, a standard QSAR model has a SMALL domain of applicability
- Remember: "*In general classical statistics is far too optimistic when validating a QSAR, because its underlying assumptions about data distributions are contradicted by the extraordinarily heterogeneous nature of chemical structures and mechanisms of biological response. Restricting the structural scope of a QSAR should help, but the distribution of "local" structural variations, within a series undergoing lead optimization, is also unlikely to be uniform.*" Richard A Kramer

Applicability Domain

- Activity landscape: potency vs. Similarity map
- Activity landscape not smooth:
  - Activity cliffs
- SAR paradox
- Similarity-dependent

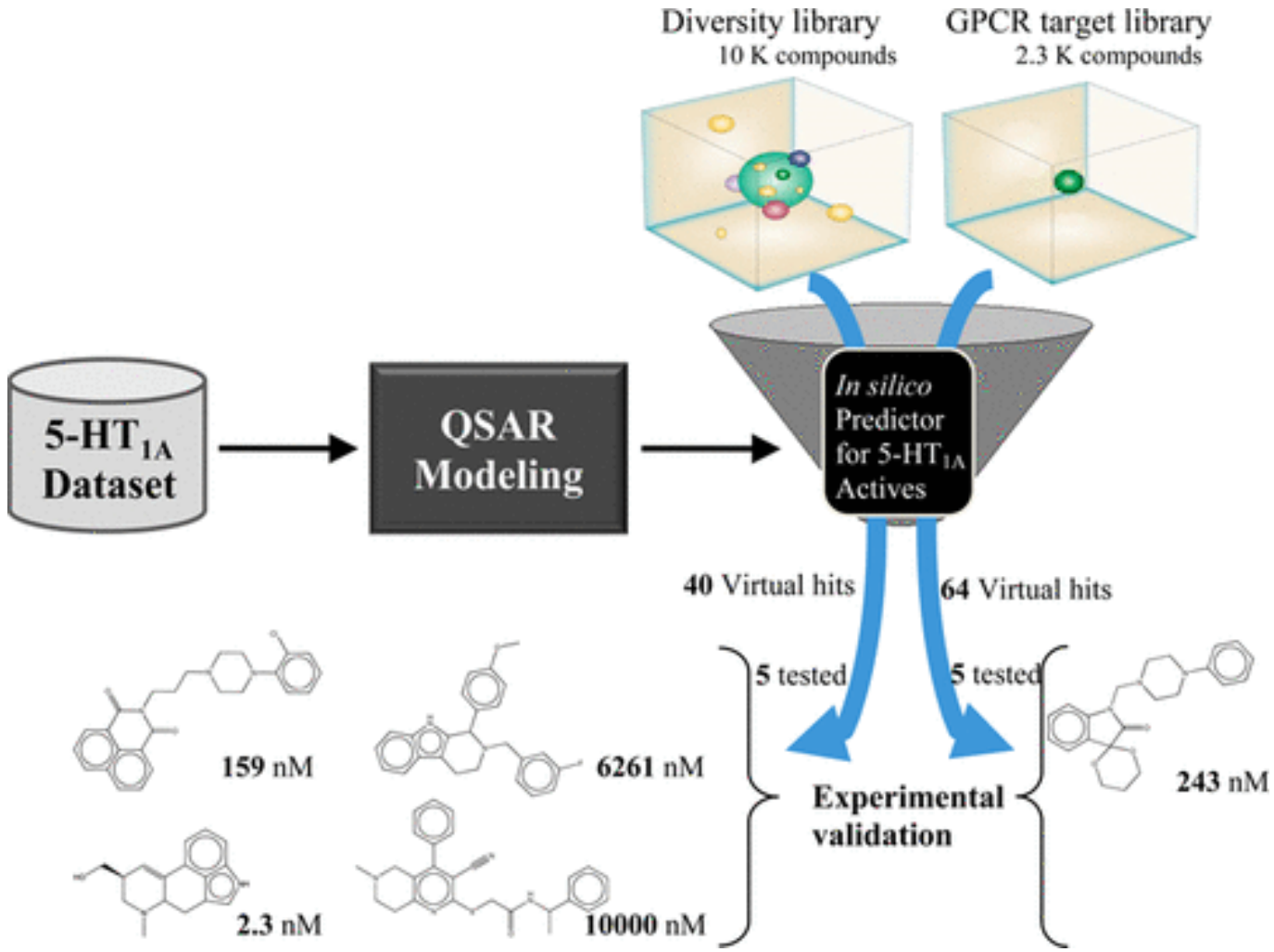




Source: Peltason, L., Iyer, P., & Bajorath, J. (2010). *Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs.*

QSAR applications

- Virtual screening



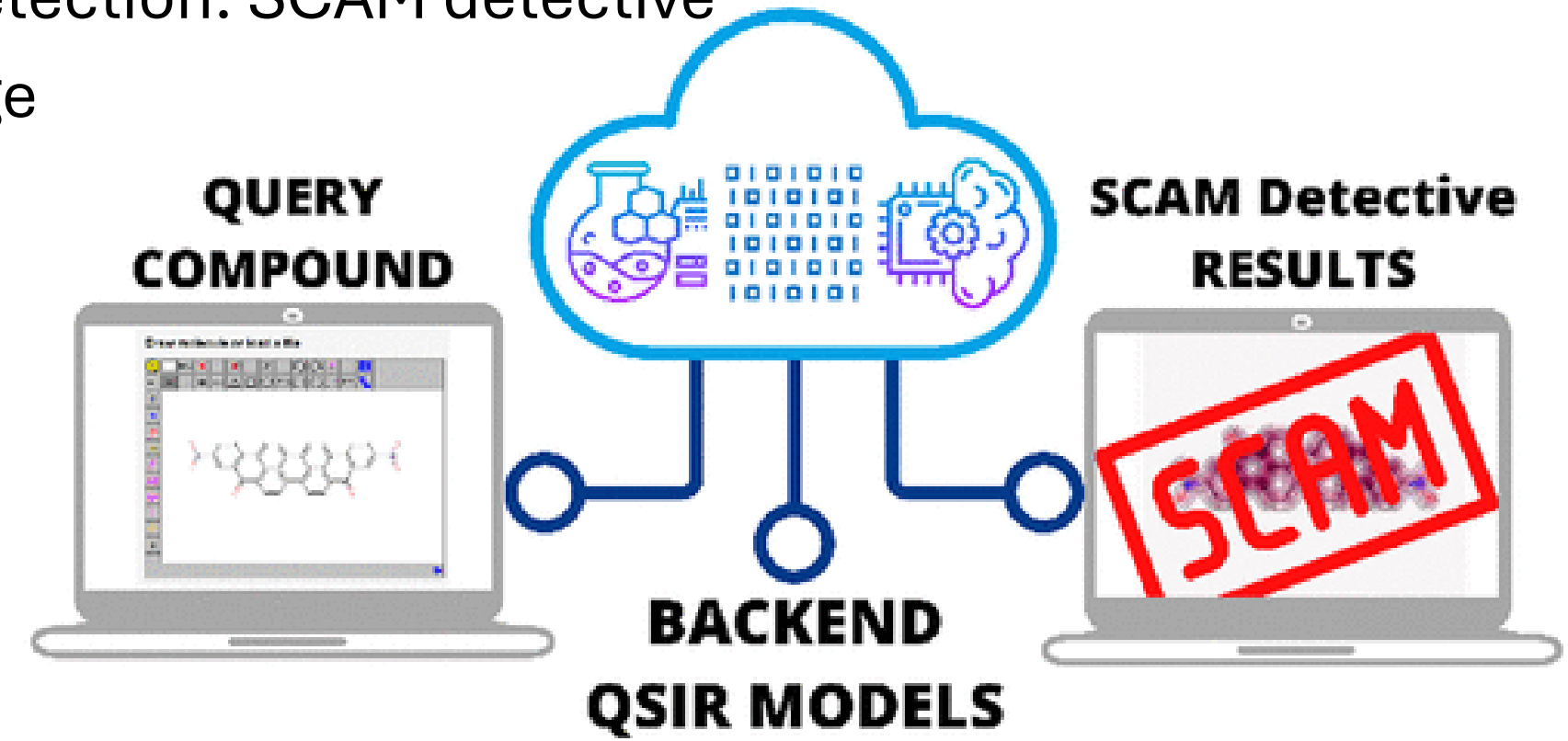
Source: *J. Chem. Inf. Model.* 2014, 54, 2, 634–647

QSAR applications

- Multiparameter optimization: e.g. ADME(T)+antitarget+potency where one or several of the parameters are calculated by discrete QSAR models
- E.g. a reward function in RL can often include "SA" or "QED"

QSAR applications

- ADME(T) filters
- Aggregation detection: SCAM detective
- Tox21 challenge



QSAR tutorial

- Open source software:
  - RDKit (cheminformatics), scikit-learn (machine learning), python data science stack (scipy, numpy, pandas), shap (model explainability), Jupyter notebooks (interactive python environment)
- Some python knowledge is useful, but notebooks are constructed so they can be re-used and adjusted without too much tinkering

Thanks for the attention

- Thank you to my funders; CZ-OPENSCREEN and Czech Science Foundation

- And see you @ the tutorial!