



## Novel Procedures for 3D-QSAR Analysis

Bernd Wendt

October 2007

## Thought-provoking impulses

---

- “The standard approach is to generate a QSAR model from all structures that have already been synthesized and tested, and then to use the model to predict for new molecules.”
- **Molecular Similarity is not working!**
- “3D-QSAR is only applied at the end of a project when every direction has been explored and nothing more could be done.”
- **3D-QSAR is dead!**

# QSAR datasets

---

- Public datasets

- Classical

- Selwood (1990)
    - Steroids (1988)

- JMC-sets

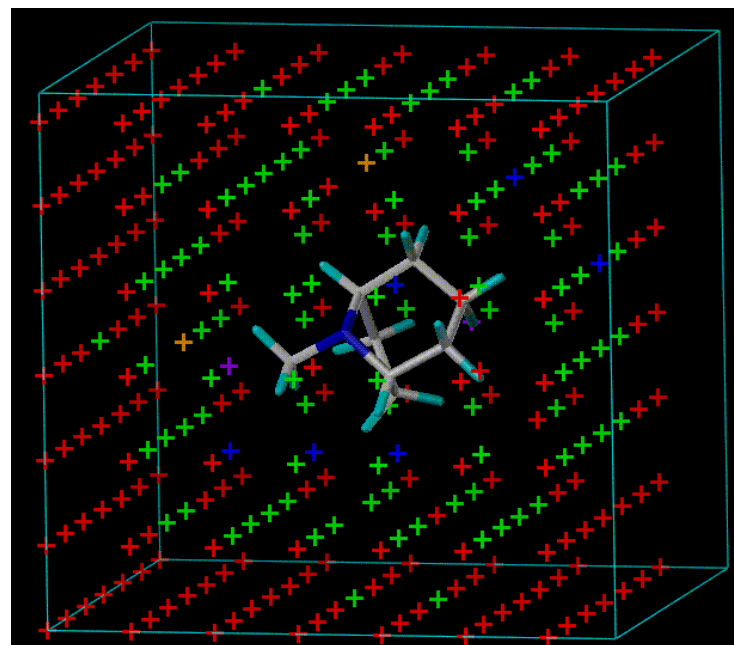
- Sutherland (2004): ACE, ACHE, BZR, COX2, DHFR, GPB, THER, THR
    - Cramer (2003): ice, thr, mao, hiv, a2a, d4, flav, cannab, aces, 5ht3, rvtrans

- Others

- Jurs: Artemisin, PDGFR
    - Tropsha: D1-Antagonists
    - Scozzafava: HCAII

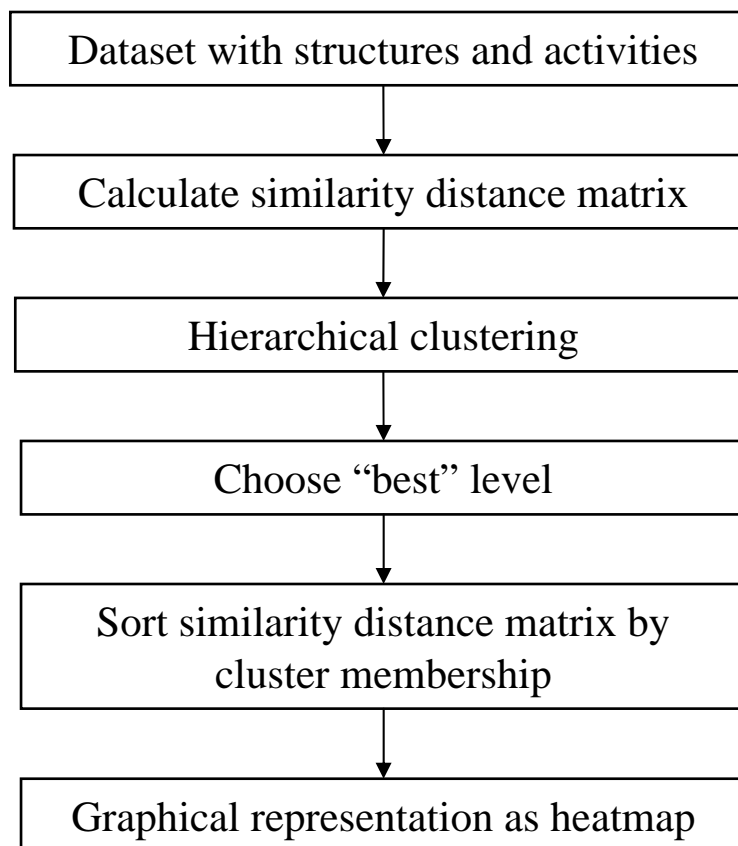
# QSAR & Similarity Descriptor: Topomers

- A topomer is a conformation of a fragment
  - Whole similarity = Euclidian sum of fragment similarities
- How topomers handle 3D
  - Structures oriented
    - Overlay of open valences
  - Single conformer
    - CONCORD 3D structures
    - Side-chain & chiral via rules, not energy
  - Automatic 3D Alignments!
- Topomer similarity involves:
  - Steric fields (as in CoMFA)
  - Rot.bond-attenuated atomic fields
- Feature matching (a bit like conventional 3D searching)



# Procedures Part 1: Construction of heatmap

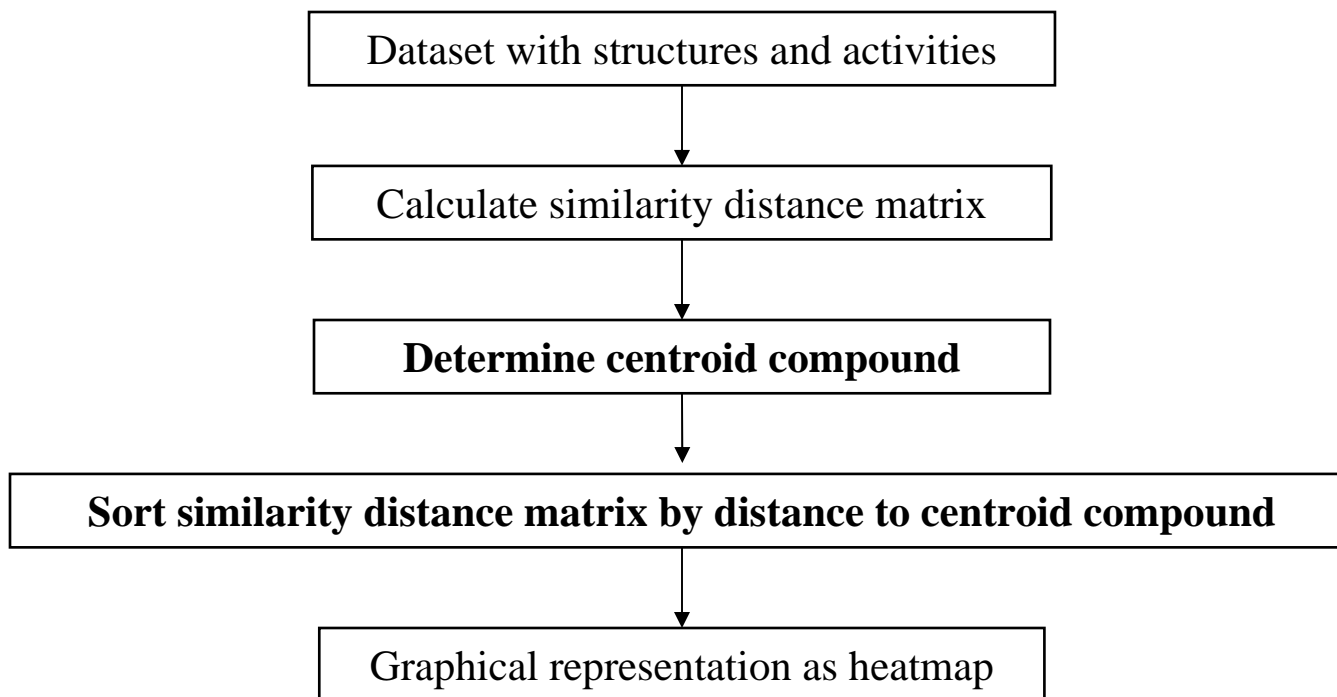
---



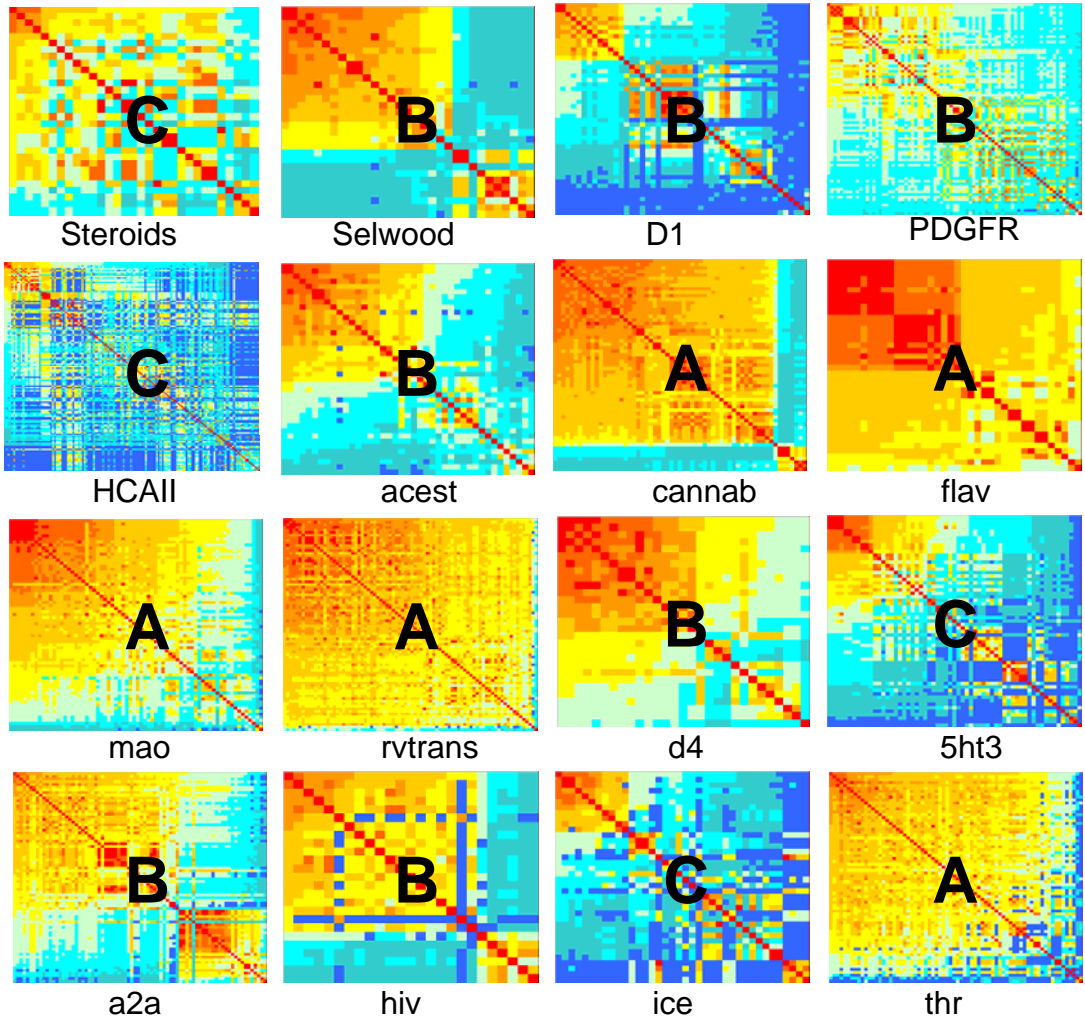


# Procedures Part 1 (modified): Construction of heatmap

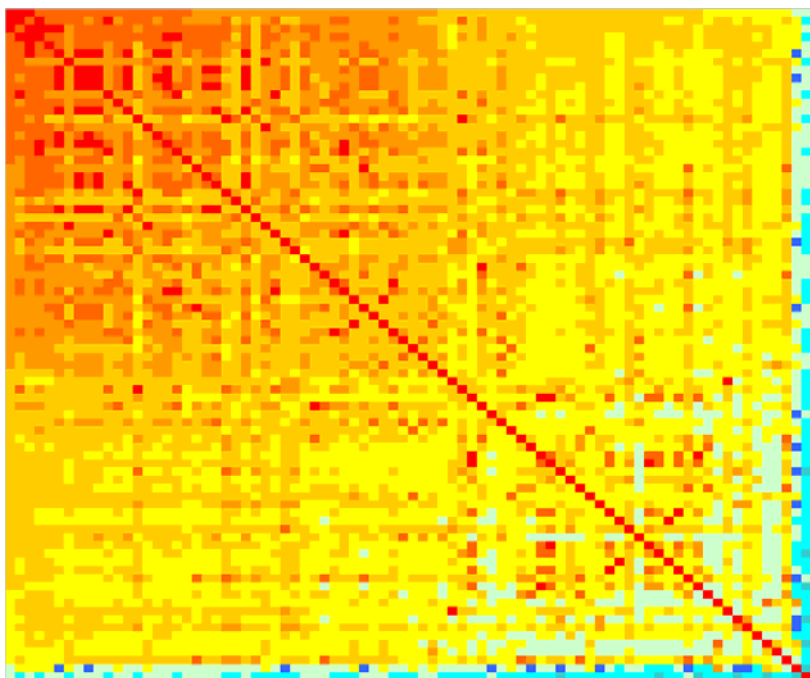
---



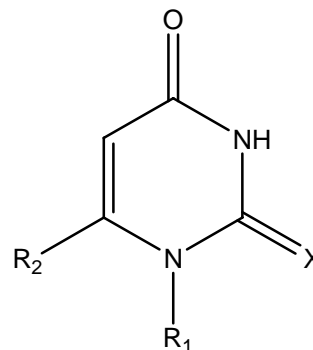
# Heatmap Examples: 16 datasets



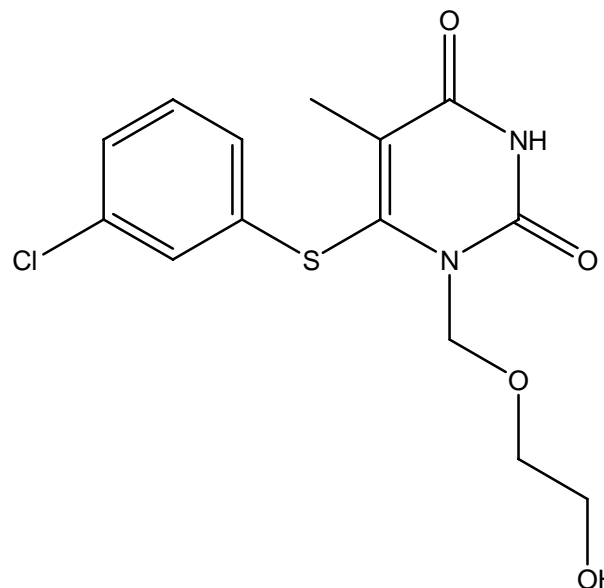
## Heatmap: Category A – one big set: rvtrans



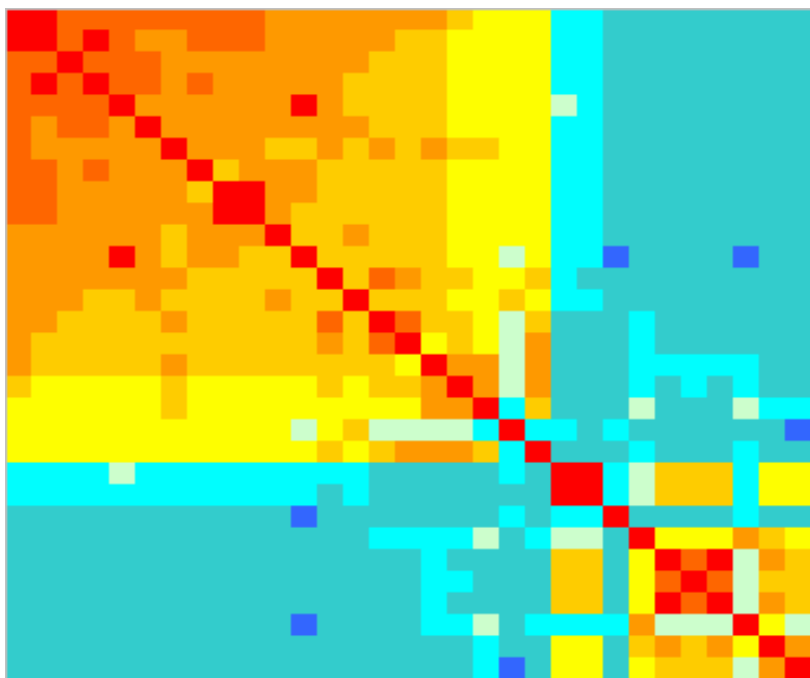
- Common Core



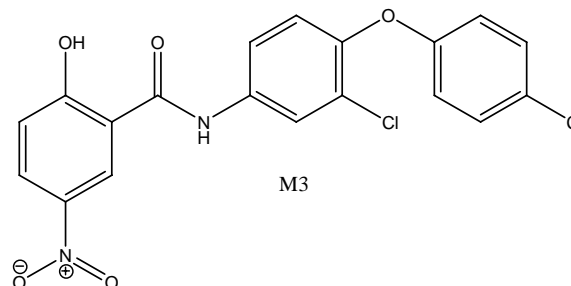
- Centroid compound



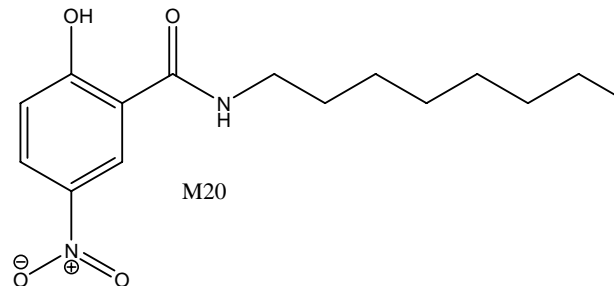
## Heatmap: Category B- a few subsets: Selwood



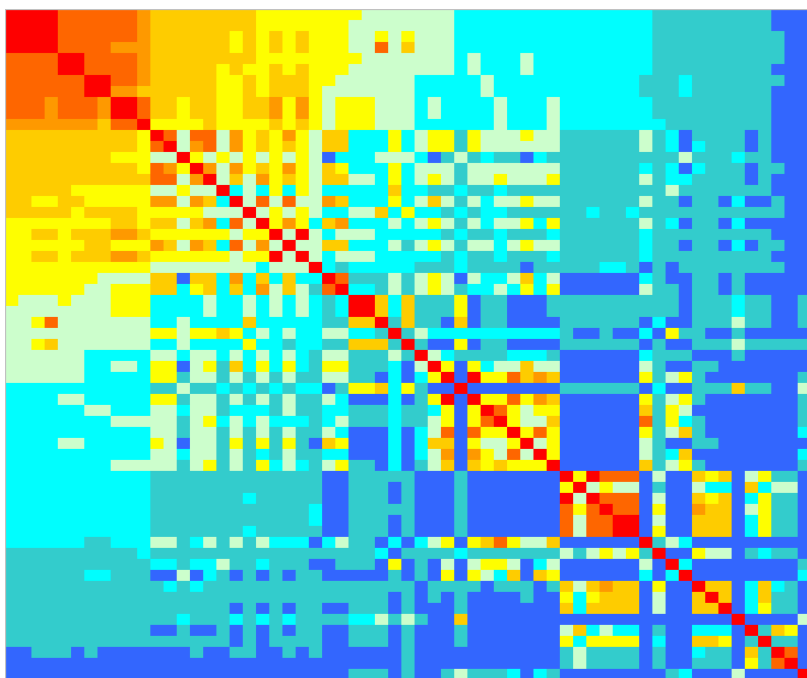
- Centroid compound (1<sup>st</sup> cluster)



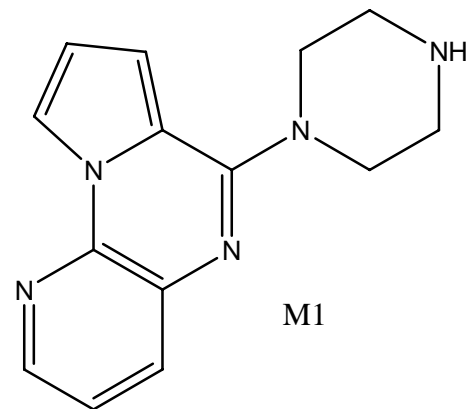
- Structure from 2<sup>nd</sup> cluster



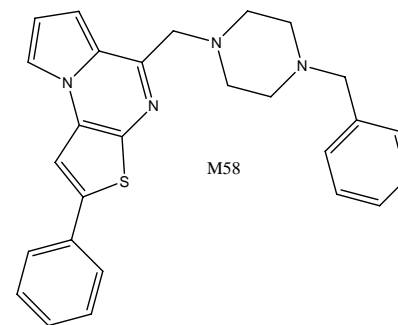
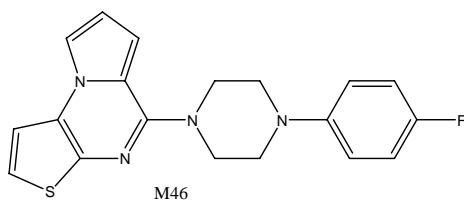
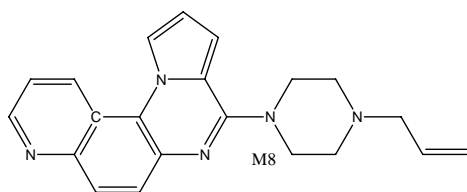
## Heatmap: Category C – multiple subsets: 5ht3



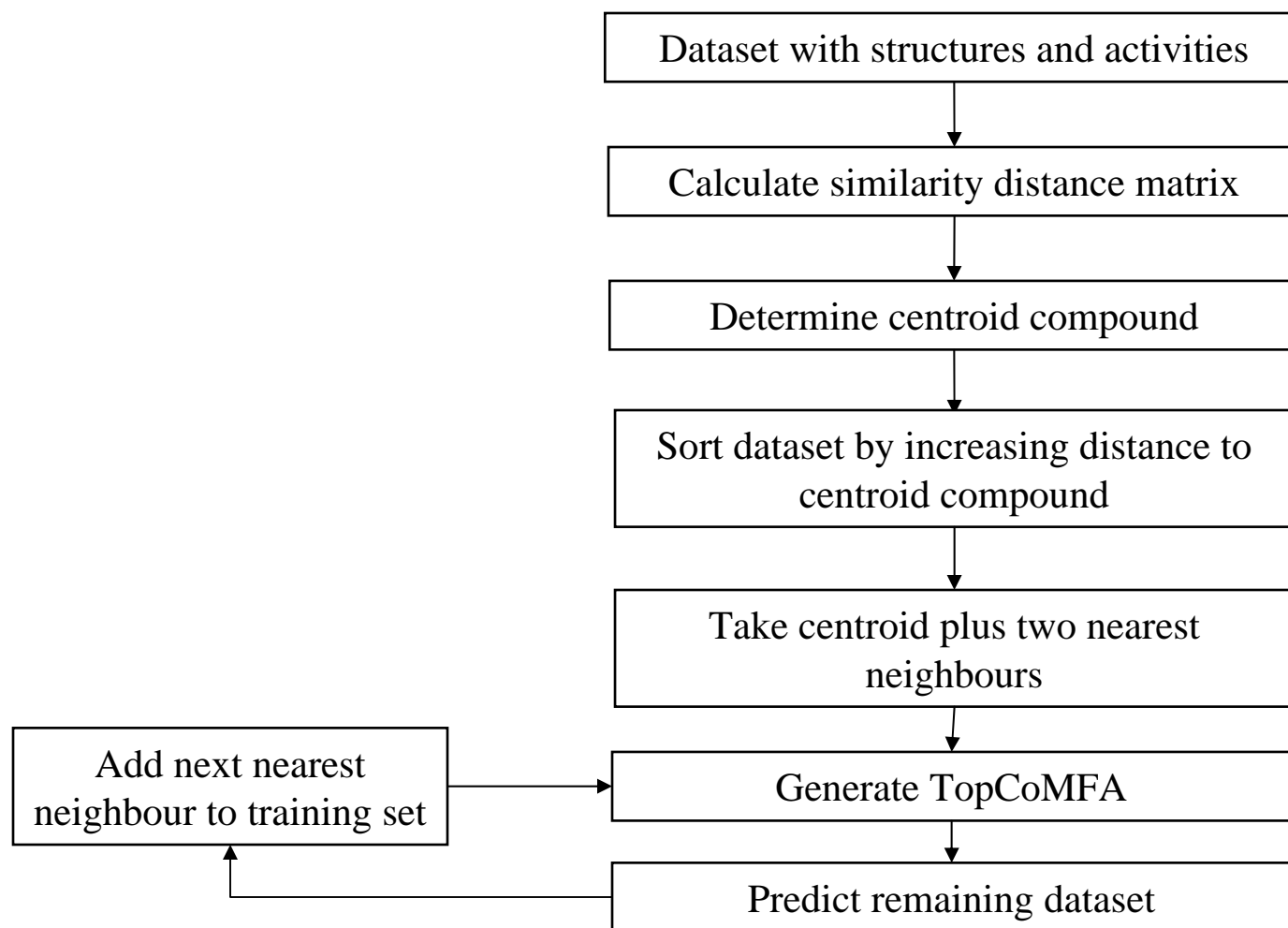
- Centroid compound



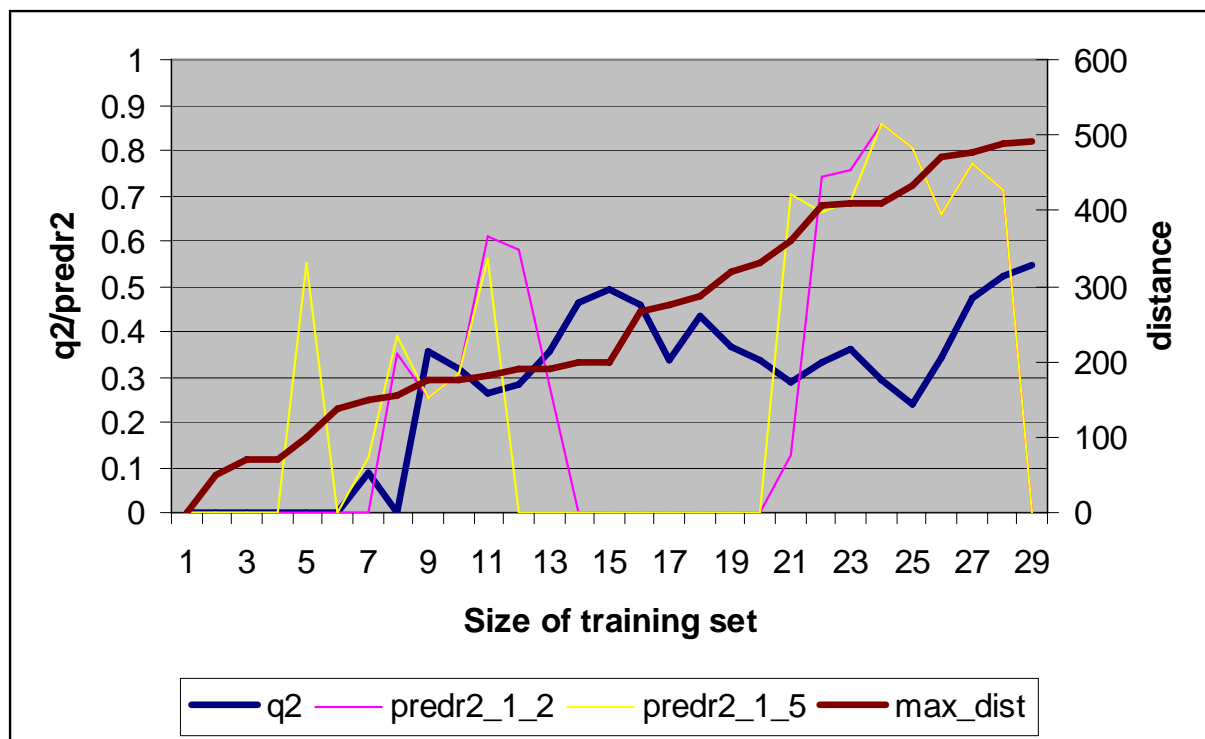
- Structures from other clusters



## Procedures Part 2 : Quantitative Series Enrichment Analysis (QSEA)

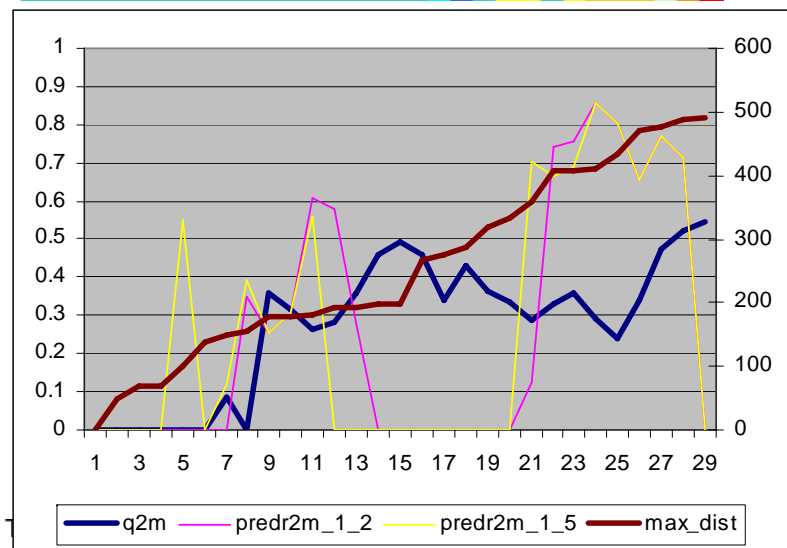
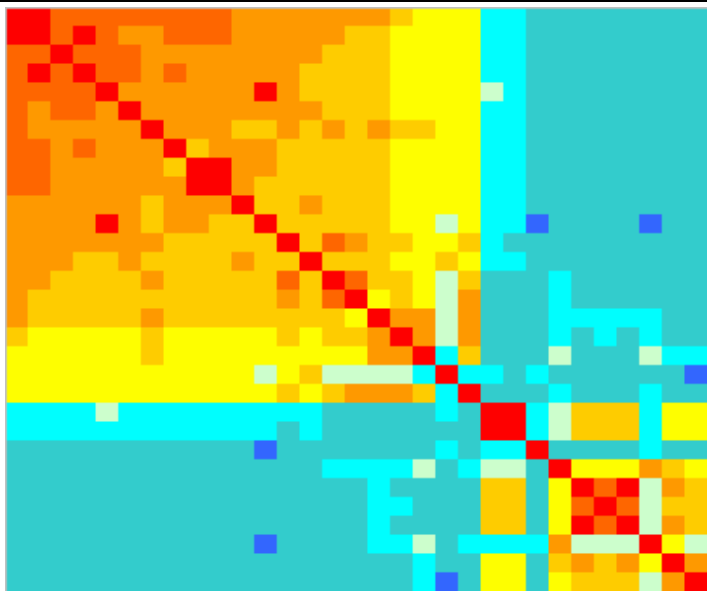


## QSEA: Series Trajectory: Selwood dataset



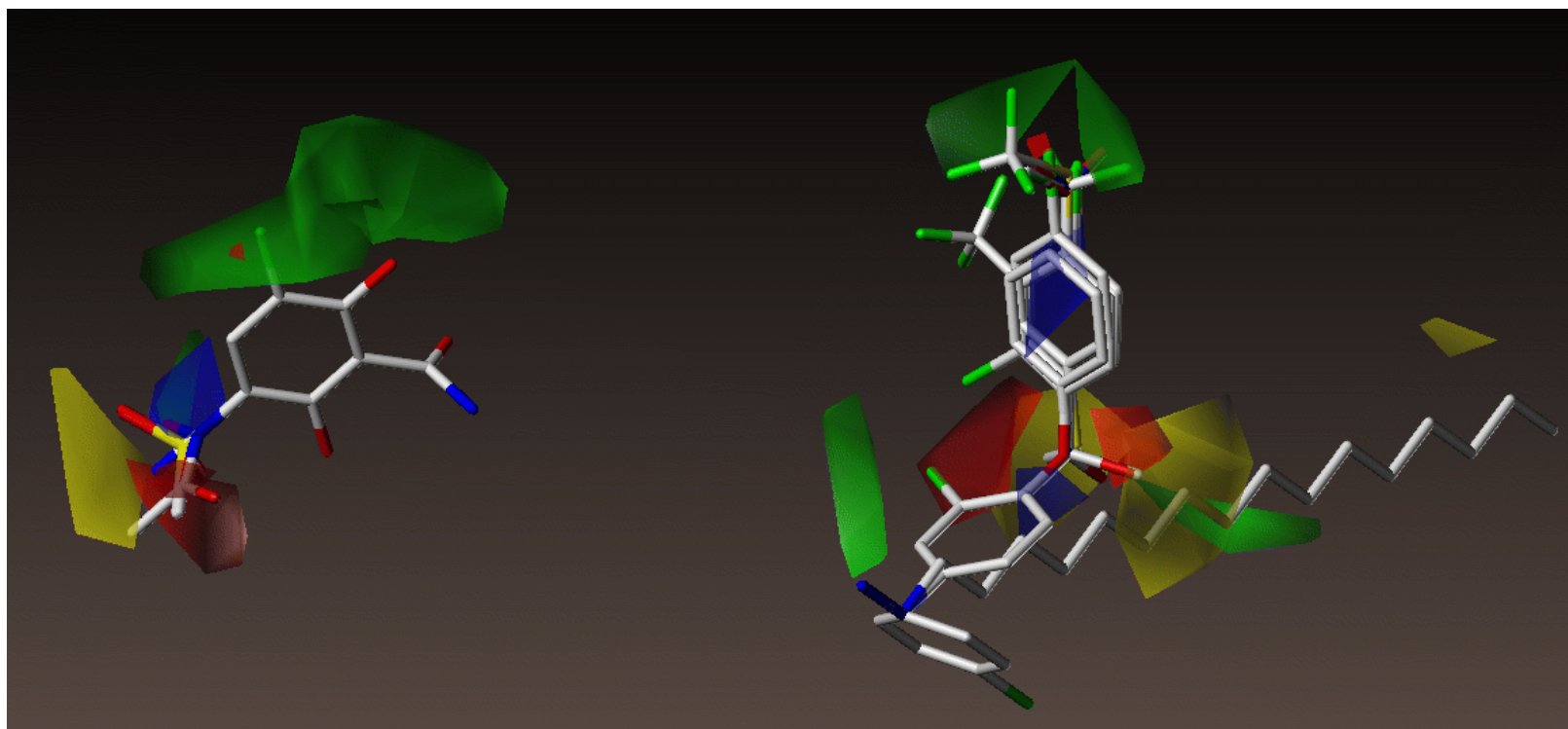
- Observables in series trajectory:
  - Similarity (maximum distance within training set)
  - Redundancy (crossvalidated  $r^2$  ( $q^2$ ))
  - Predictivity (pred- $r^2$  within distance-controlled set of test compounds)

## QSEA: Selwood



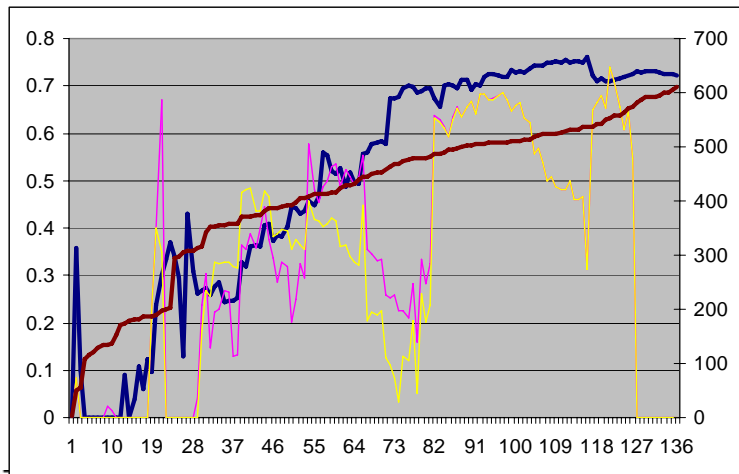
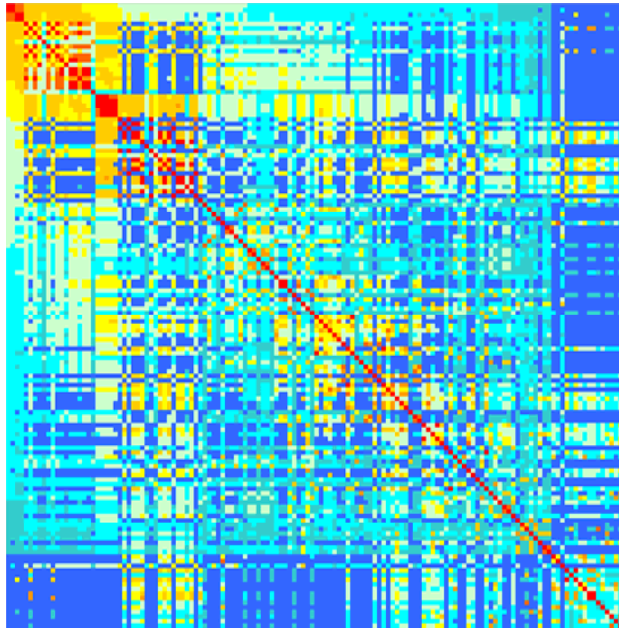
- Q2 is going up and stays at  $\sim 0.4$
- Predictions are controlled by relative distance to the training set
  - Predr2m\_1\_2: all compounds with 1.2 times maximum distance of the training set
- Predictive  $r^2$ 
  - builds up during first cluster
  - Then disappears at transition to 2<sup>nd</sup> cluster
  - Finally rebounds
- “Well-behaved dataset”
  - Continuous and independent variation

## TopCoMFA Contours: Selwood



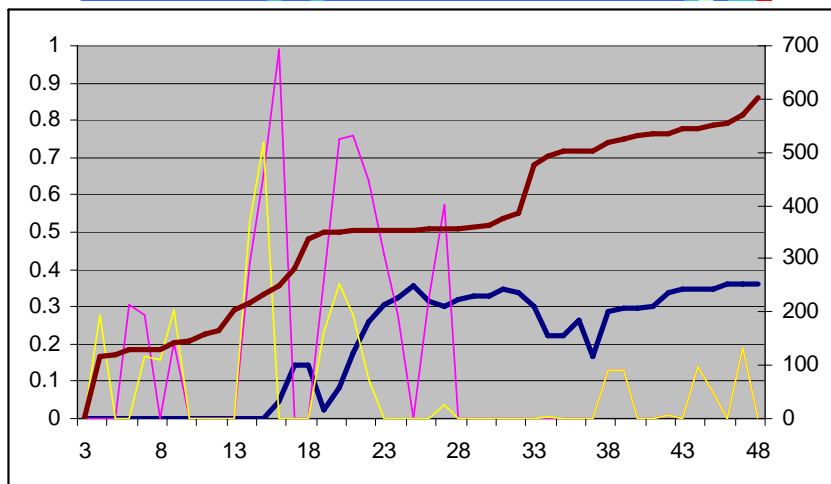
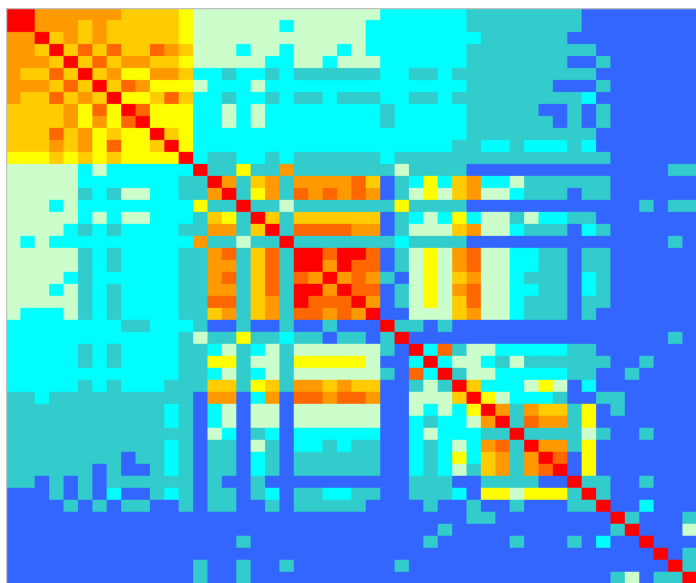
- All 31 compounds aligned
  - Carboxylic-R-group on left – Amine-R-group on right
- Structures from two clusters readily identifiable

# QSEA: HCAII



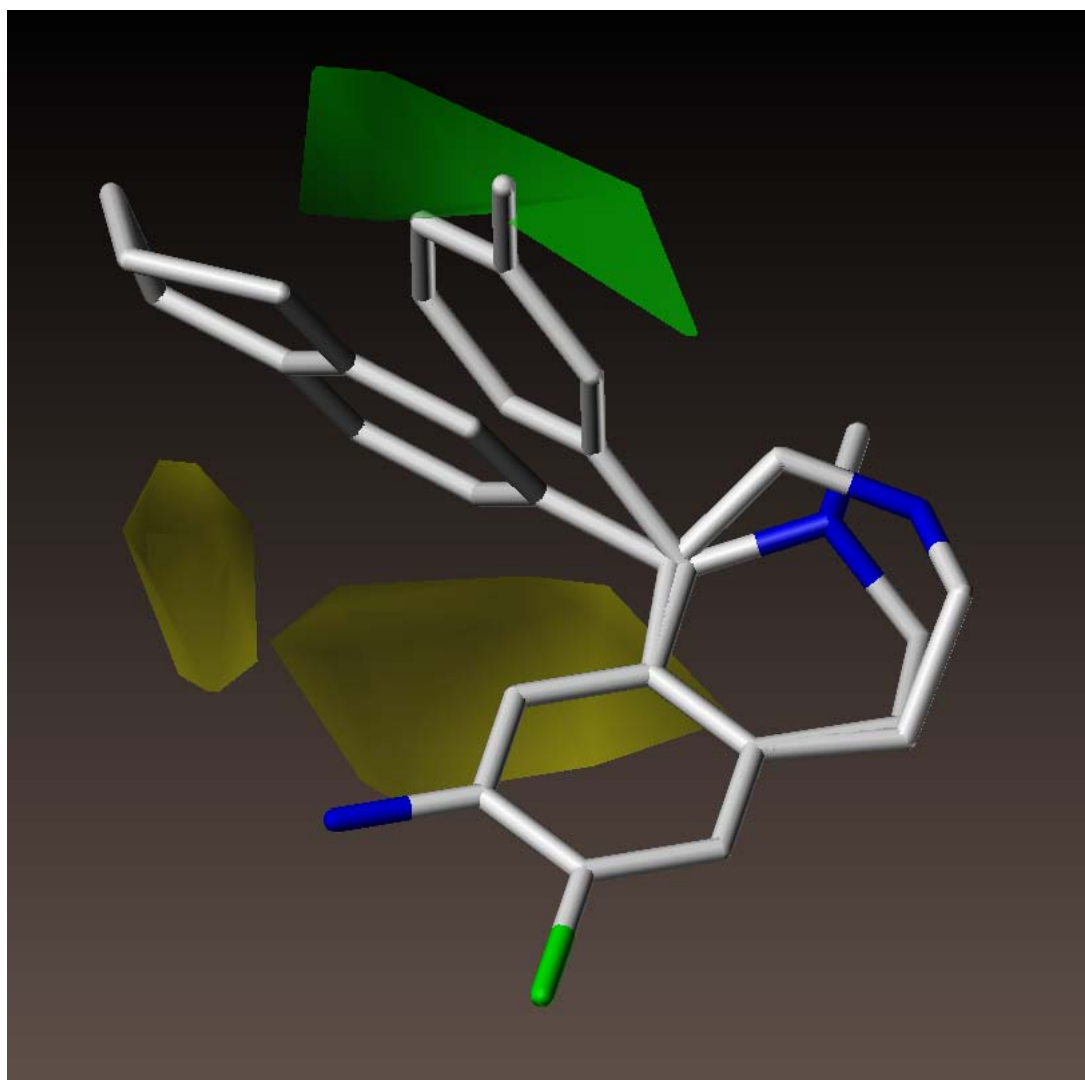
- Combinatorial library
  - One global series
  - many local series
- Continuous increase of:
  - Maximum distance within training set (brown curve)
  - Redundancy (q2 – blue curve)
  - Predictivity (pred\_r2 – yellow/magenta curves)
- “Well-behaved dataset”
  - Continuous and independent variation

## QSEA: D1



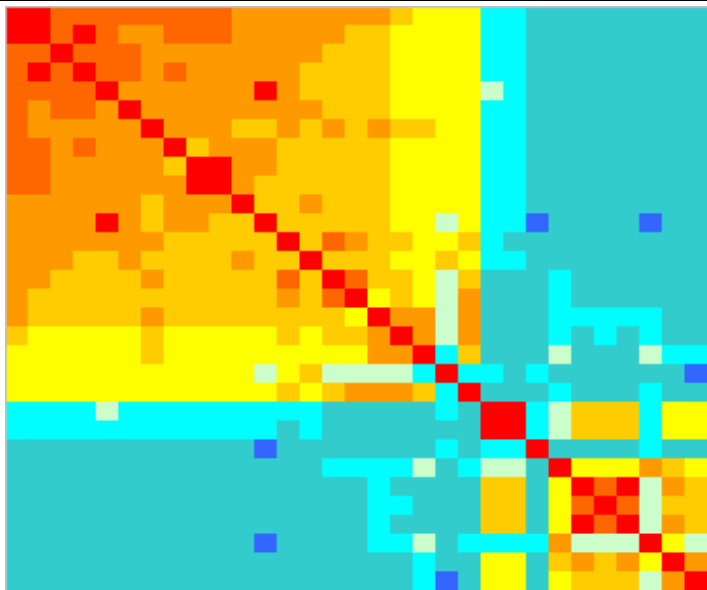
- First two clusters with consistent behaviour
  - Tetrahydro-isoquinolines with two different substitution patterns
- Third cluster destroys predictivity
  - Benzodiazepines
- Some outliers
  - Thioxanthenes
- Problematic dataset:
  - Better divide dataset into subsets

## TopCoMFA Contours: D1

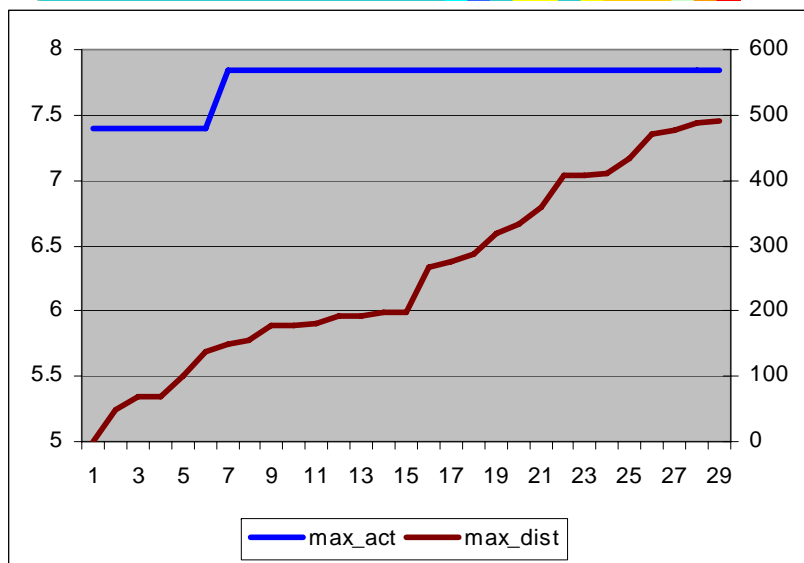


- Tetrahydro-isoquinoline
  - Robust model
    - $Q^2=0.81$ ;
    - $r^2=0.9$ ,
    - $\text{stderr}=0.1$
  - Green contour around pendent phenyl indicates favoured region
- Benzodiazepine
  - Weakest compound penetrates favoured region of model

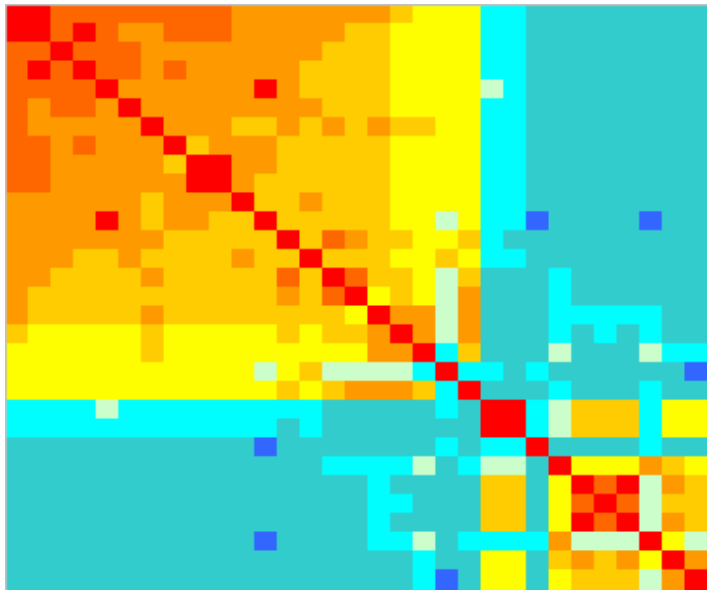
## QSEA: Guiding Synthesis: Selwood



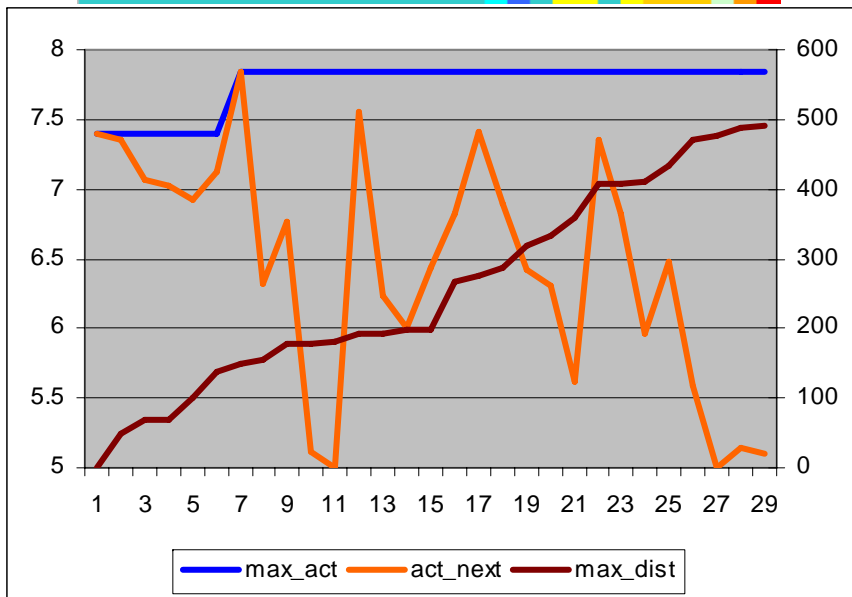
- Highest activity (max\_act) associated with the center of the dataset
- No incentive to move away from center
  - A new direction might be desired to direct synthesis



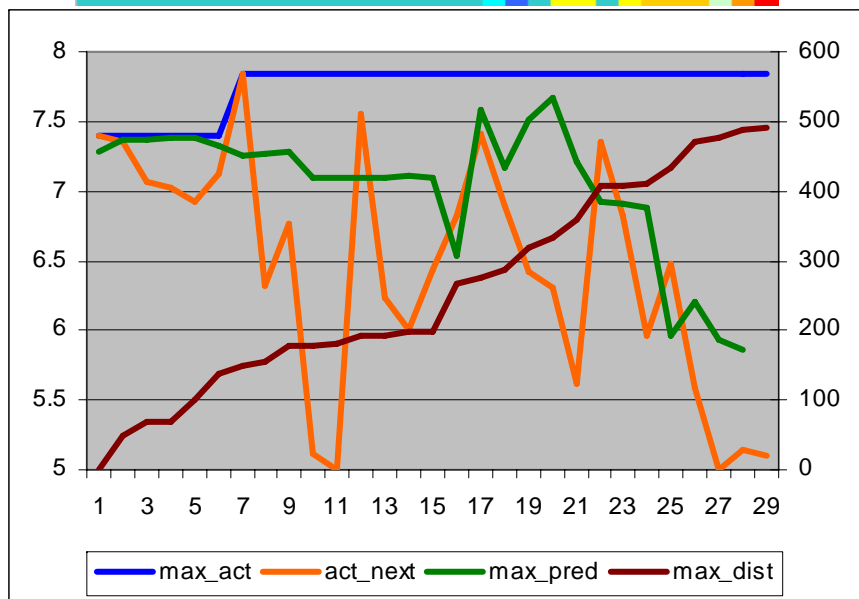
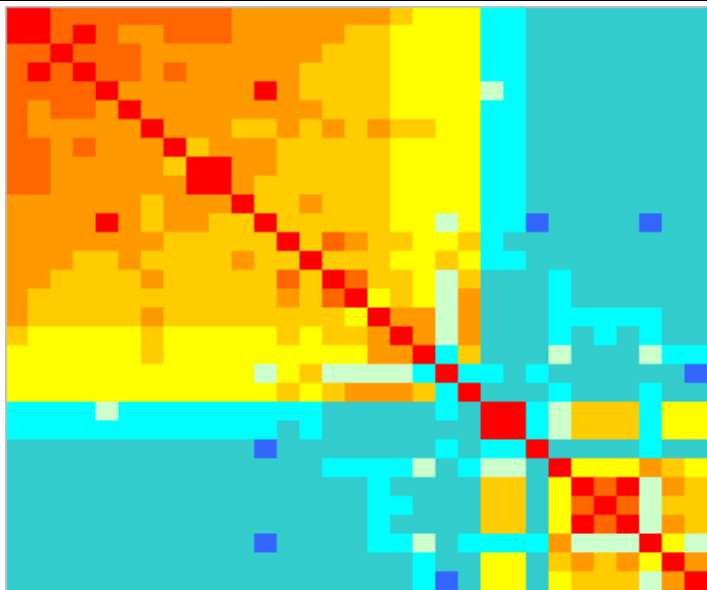
# QSEA: Guiding Synthesis: Selwood



- Highest activity (max\_act) associated with the center of the dataset
- No incentive to move away from center
  - A new direction might be desired to direct synthesis
- Rollercoaster ride of activity (act\_next)
  - Important points for SAR



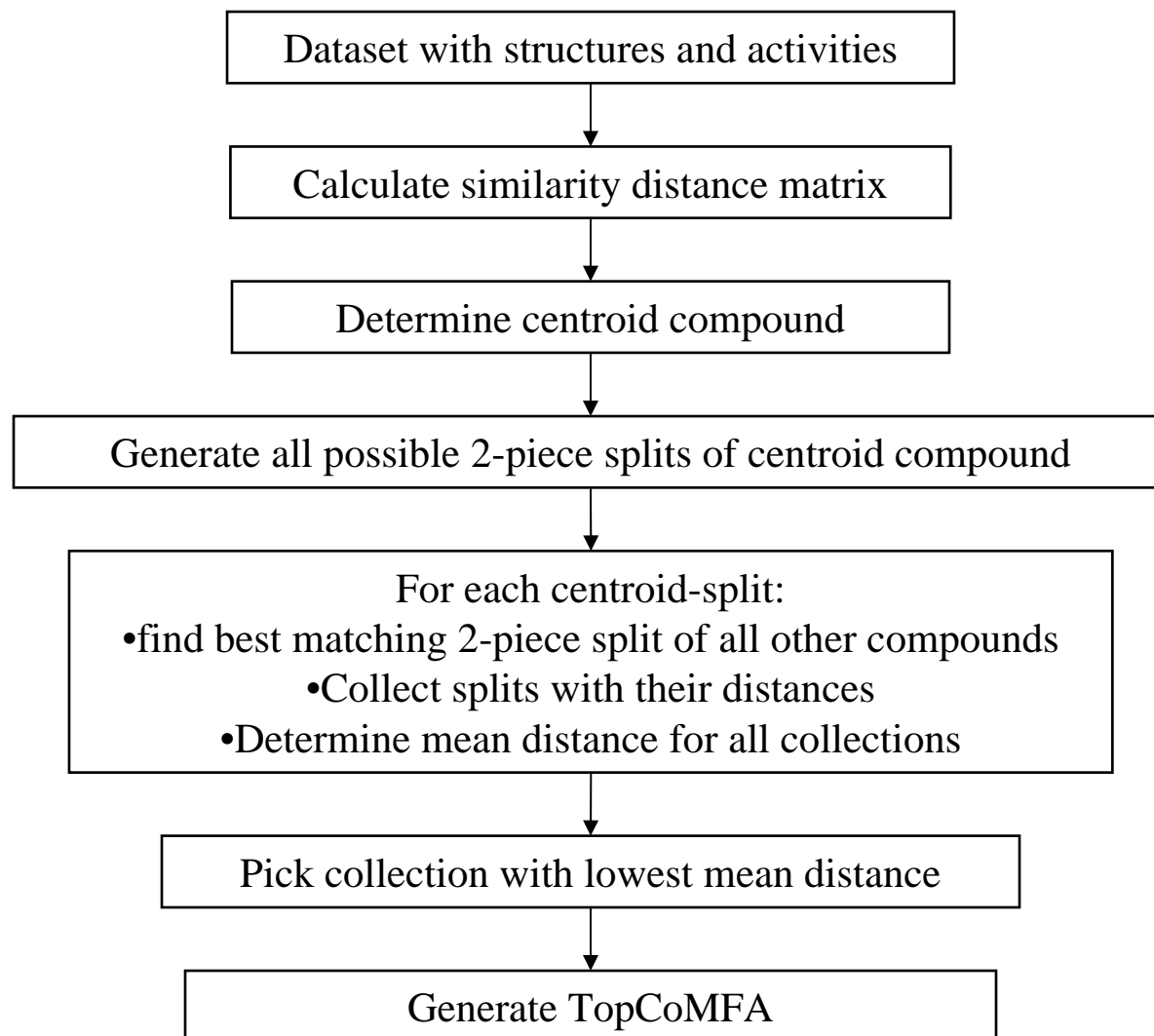
## QSEA: Guiding Synthesis: Selwood



- Highest activity (max\_act) associated with the center of the dataset
- No incentive to move away from center
  - A new direction might be desired to direct synthesis
- Rollercoaster ride of activity (act\_next)
  - Important points for SAR
- Using QSAR aggressively:
  - Report highest predicted activity (max\_pred) at each step
  - QSAR would not have suggested these compounds!
  - Search large Fragment-DB (Allchem) for better structures

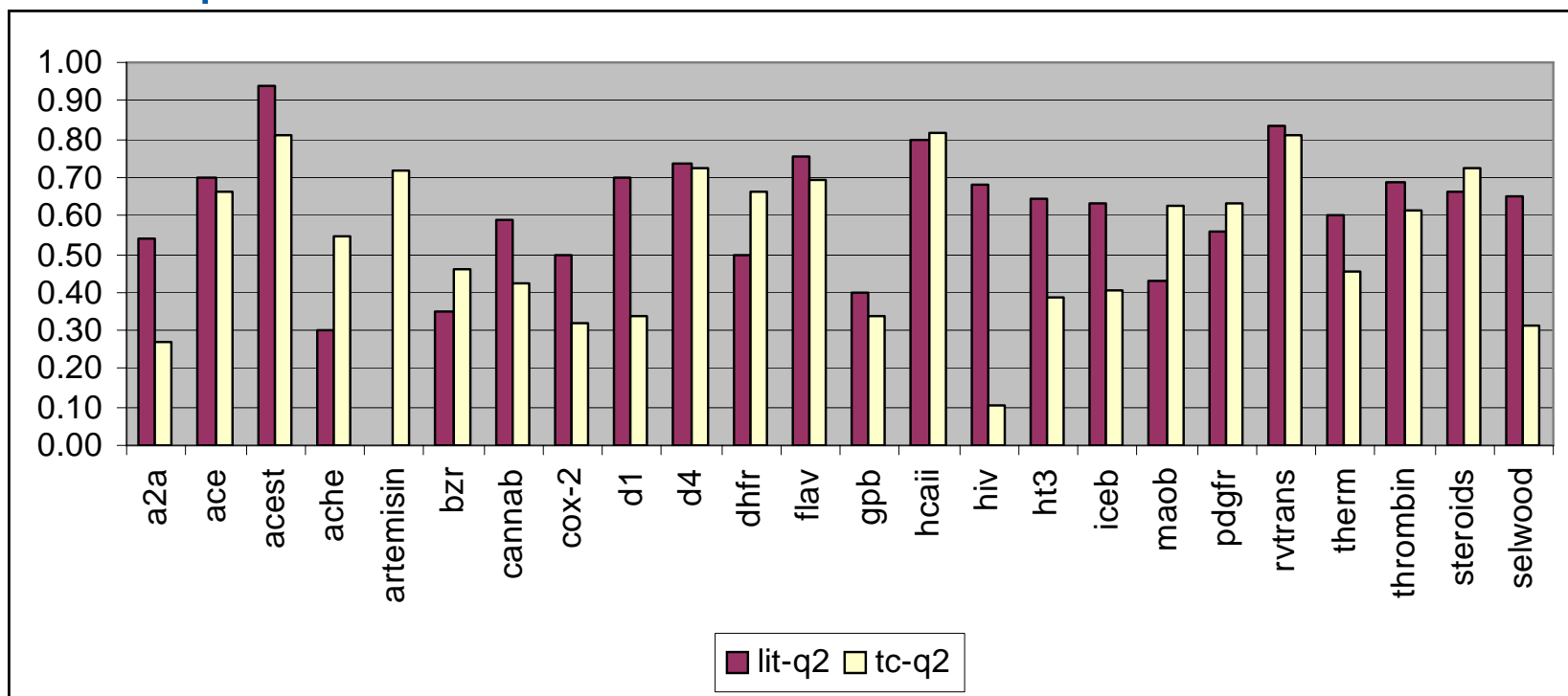
## Procedures Part 3 : Fully automated TopCoMFA

---



# Fully automated TopCoMFA: Applied to 24 Datasets

- Comparison of model statistics



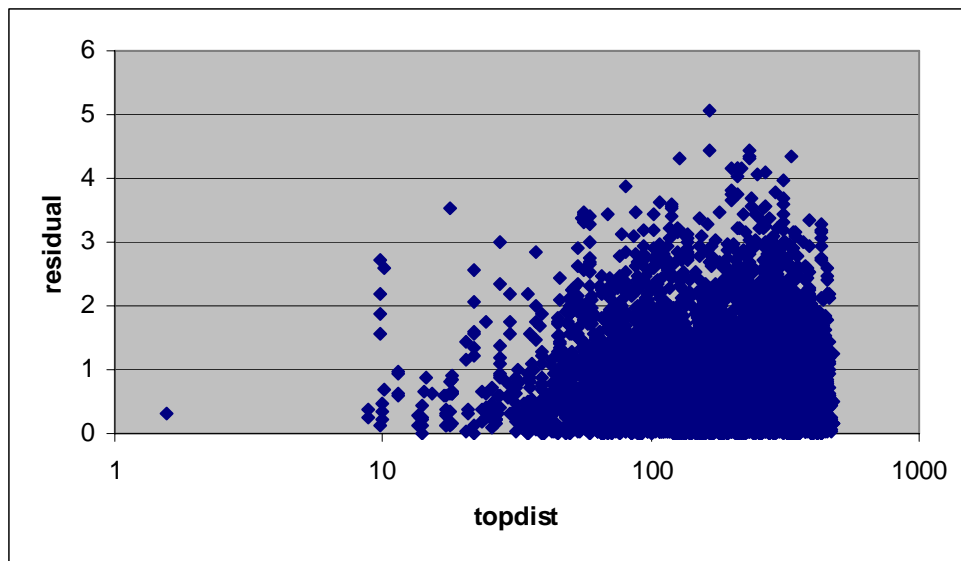
- Average q2 from literature: 0.61
- Average q2 from Auto-TopCoMFA: 0.54

# Fully automated TopCoMFA: Leave-some-out validation

---

- Split up datasets into training and test
  - 24 datasets from literature
  - 67% train / 33% test
  - Repeat each one 10 times
- Use training set
  - To determine centroid compound
  - To generate TopCoMFA
  - To predict test set

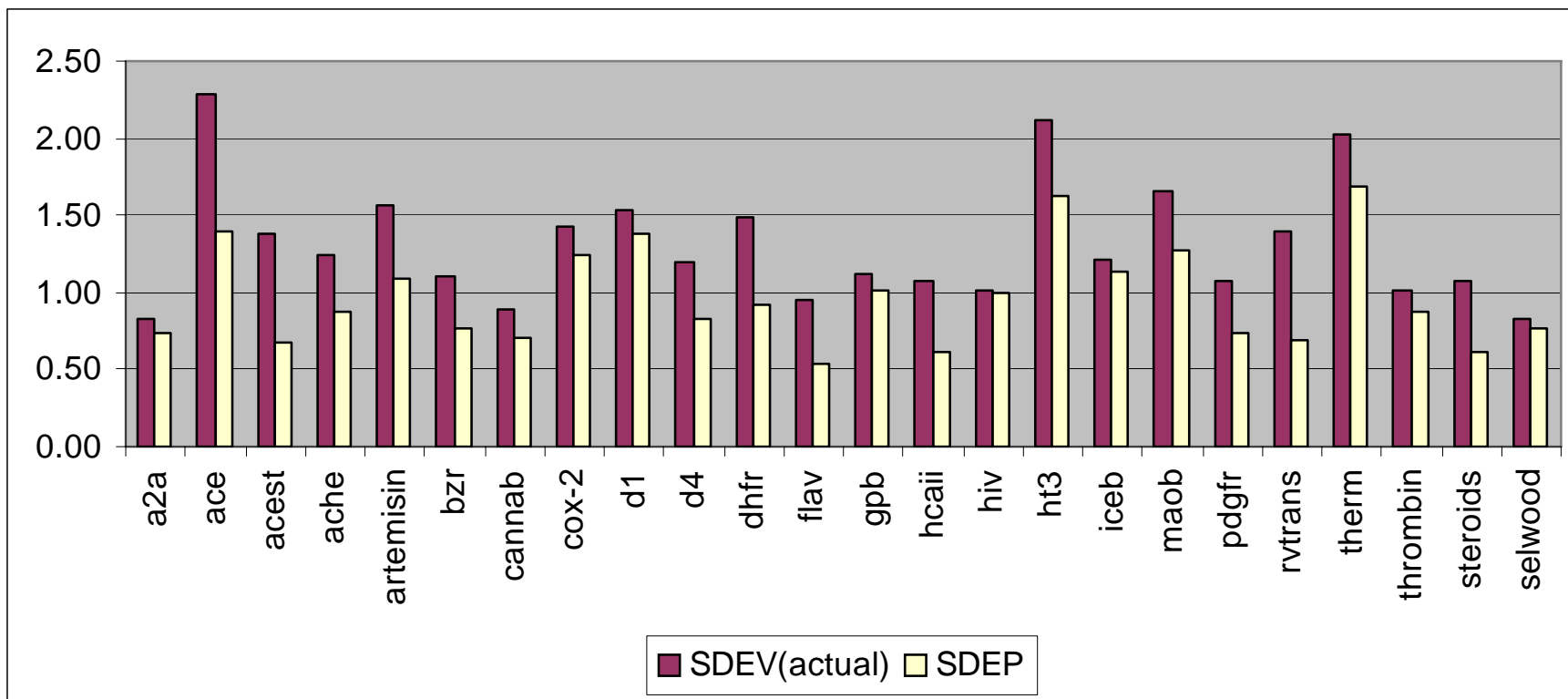
# Fully automated TopCoMFA: LSO-Validation



- >7,500 predictions
- Trapezoidal distribution of residuals versus topomeric distance reflects good neighbourhood behaviour of descriptor

# Fully automated TopCoMFA: LSO-Validation

- Comparison of prediction statistics



- Average SDEP is <1 with average SDEV of activities >1.3
- To be improved by QSEA

## Conclusions

---

- “The standard approach is to generate a QSAR model from all structures that have already been synthesized and tested, and then to use the model to predict for new molecules.”
- **Molecular Similarity works more often than not!**
- “3D-QSAR is only applied at the end of a project, when every direction has been explored and nothing more could be done.”
- **3D-QSAR can also be applied in early stages of the project to identify turning points or to confirm known trends.**