The Role of the Neighborhood in QSAR Modeling and Cheminformatics

Rajarshi Guha

School of Informatics
Indiana University

15th January, 2007

Department of Pharmaceutical Technology
Jadavpur University, Calcutta

Outline

- Local regression models
- Outlier detection in compound libraries and counting the *natural* number of clusters
- Ensemble feature selection

Outline

QSAR Modeling

Global models

- Traditional models are global since they encompass the whole dataset
- They can be influenced by molecules at the extremes of the activity range
- A complex SAR may not be characterized well by a single model

QSAR Modeling

Local models

- Global models capture *major* trends
- These might overshadow minor features which are relevant to a small group of molecules
- To take into account local trends we can build models on clusters of compounds
  - Generally requires us to define the number of clusters
  - Valid if there are distinct clusters

QSAR & Near Neighbors

- Rather than cluster *a priori*, detect near neighbors on the fly
- Fit a linear model to the neighborhood
  - Can be extended to higher order models
  - Dependent on the nature of the neighborhood
- This implies that there is no single model for the dataset
- Previously used for time series analysis

Birattari, M. et al., *A*dv. Neural. Inf. Proc. Sys., **1999**, 375–381

Prior Work

- Genetic algorithm to search for subsets exhibiting a linear trend
- Clustering
  - Spanning trees
  - *k*-means
- Kriging

Barakat, N. et al., *Chemom. Intell. Lab. Systems*, **2004**, *72*, 73–82

Shen, Q. et al., *J. Chem. Inf. Model.*, **2004**, *44*, 2027–2031

Fang, K.T. et al., *J. Chem. Inf. Model.*, **2004**, *72*, 2106–2113

Local Linear Regression

- Traditional OLS

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\beta = (X^T X)^{-1} X^T Y$$

- Local regression

$$\beta_{x'} = (X^T_{NN(x')} X_{NN(x')})^{-1} X^T_{NN(x')}$$

- Models complex relationships using simple approximations
- Saves training time by deferring model building

Tools & Datasets

Tools and techniques

- R 2.2.0
- The `lazy` package
- k was automatically determined using LOO cross-validation
- Since the neighborhood is usually small, ridge regression is used

Datasets

- Artemisinin analogs
  - 179 molecules
  - Reduced pool of 65 descriptors
- DHFR inhibitors
  - 672 molecules
  - Reduced pool of 36 descriptors

Summary of the Results

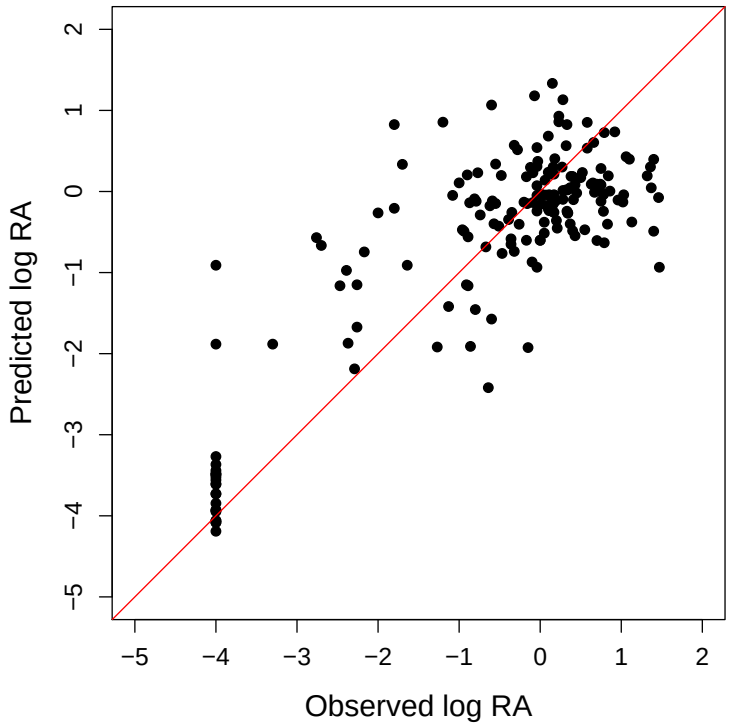
Dataset		RMSE	
		Global Model	Local Model
Artemisinin	Whole Dataset	0.86	0.62
	Prediction Set	0.92	0.94
PDGFR	Whole Dataset	0.50	0.43
	Prediction Set	0.36	0.31
DHFR	Whole Dataset	2.14	1.56
	Prediction Set	2.16	2.01

The range of the dependent variable for the artemisinin dataset was 2.53 log units, for the PDGFR dataset was 2.95 log units and for the DHFR dataset was 17.09 log units.

Artemisinin Analogs

Global Regression Model

- Built a 4-descriptor OLS model using the whole dataset
- Statistically valid
- RMSE = 0.86
- Significant variation for actives
- 23 *inactives* are not predicted well



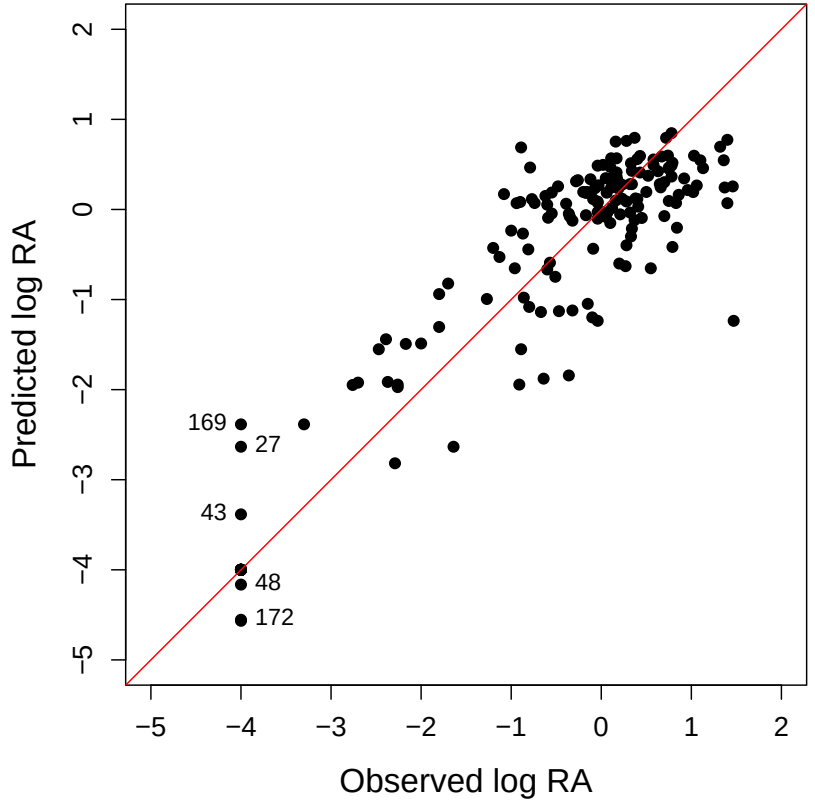
	Estimate	Std. Error	t value
(Intercept)	−60.17	5.03	−11.96
N7CH	−0.21	0.01	−17.30
NSB	0.22	0.02	10.30
WTPT-2	27.73	2.48	11.18
MDE-14	0.11	0.02	4.93

F-value $= 100.2\ (4, 174)$
$F_{crit} = 2.42\ (\alpha = 0.05)$

Artemisinin Analogs

Local Regression Model(s)

- Since we see two groups of molecules, local regression should work better
- Used descriptors from the global model
- Overall RMSE = 0.62
- Of the 23 inactives
  - 5 are mispredicted
  - exhibit increased variance

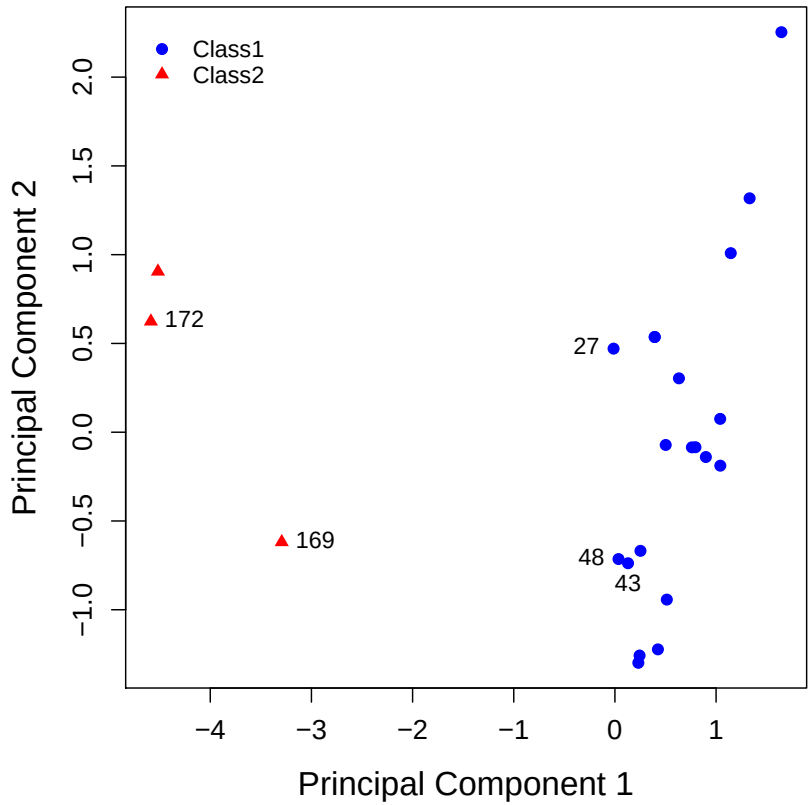


Artemisinin Analogs

Fuzzy Clustering of Inactives

- Evaluate principal components
- Colored based on cluster membership
- Number of cluster defined *a priori*

- **169** and **27** had the largest error
- Though **27** is close to a cluster, its silhouette value is very low compared to that of cluster average

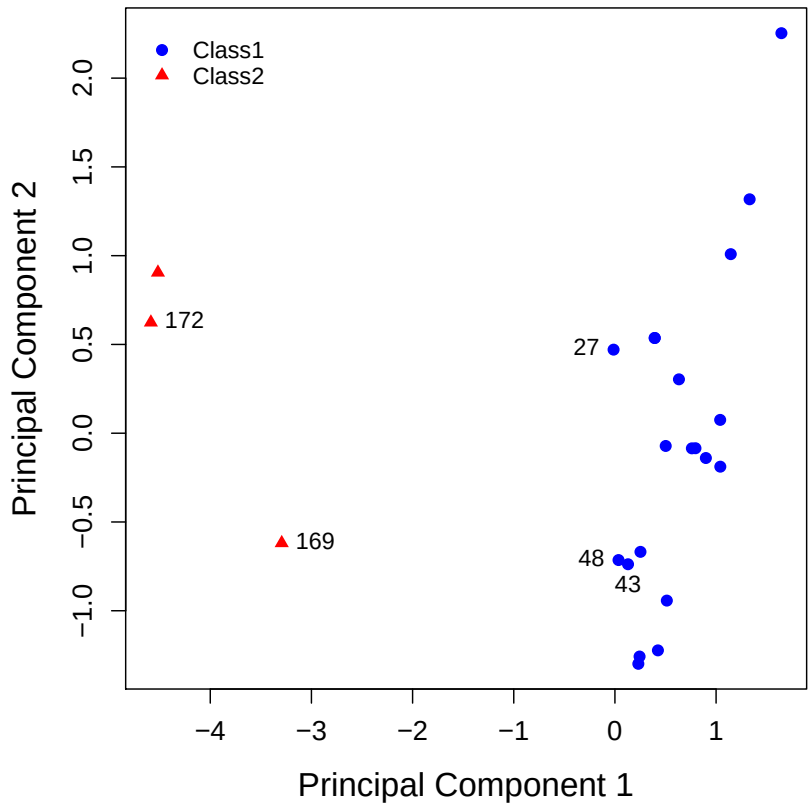


Artemisinin Analogs

Fuzzy Clustering of Inactives

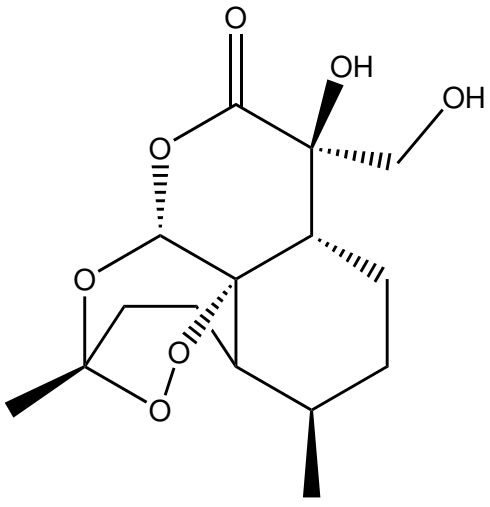
- Evaluate principal components
- Colored based on cluster membership
- Number of cluster defined *a priori*

- **169** and **27** had the largest error
- Though **27** is close to a cluster, its silhouette value is very low compared to that of cluster average

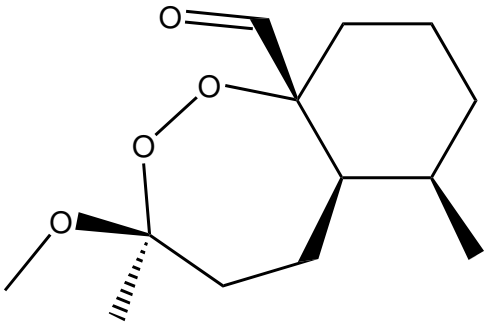


Artemisinin Analogs

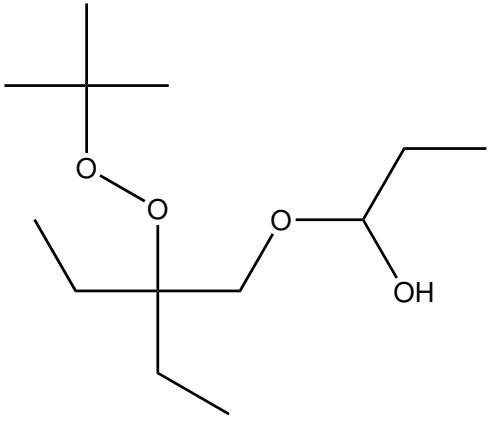
Inactive Outliers



27



169



172

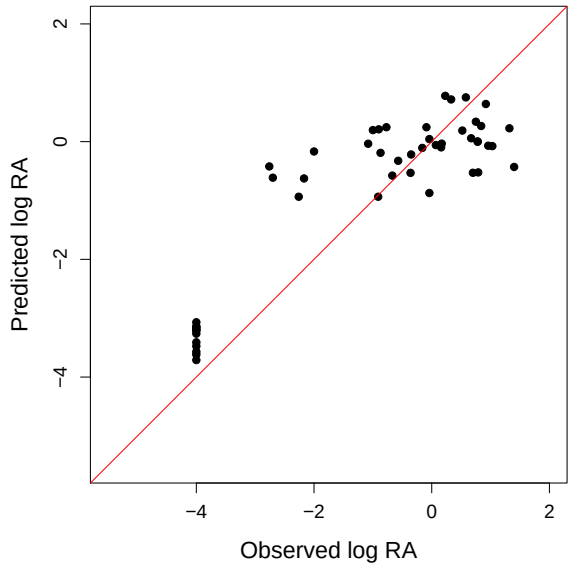
Artemisinin Analogs

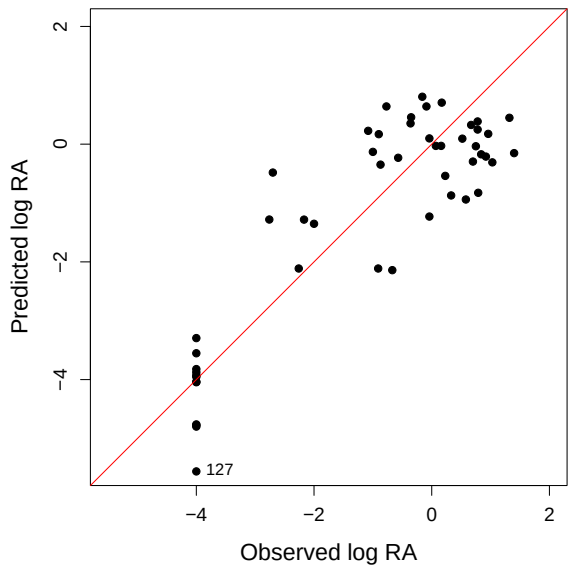
TSET/PSET Split

- Split the dataset into TSET (129) and PSET (50)
- Built a global model using the TSET, predict the PSET
- Used local regression to get predictions for the PSET

Results

- Global, RMSE = 0.92
- Local, RMSE = 0.94
- Removing *127* results in RMSE = 0.91

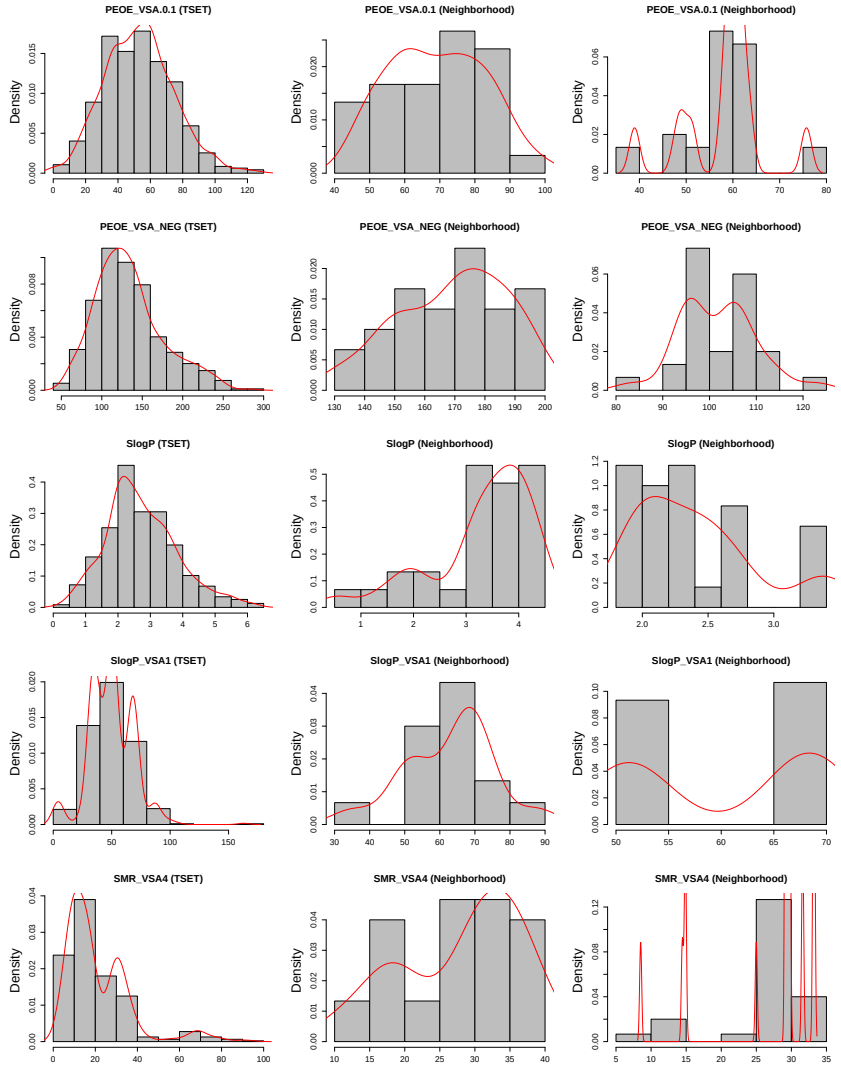




Local Descriptor Distribution

DHFR Inhibitors

- If the local distribution is similar to the population distribution one may expect reliable predictions
- Skewed distributions will not lead to significant models
- If the underlying relationship between descriptors and observed activity is not linear in the neighborhood, the local model will be poor

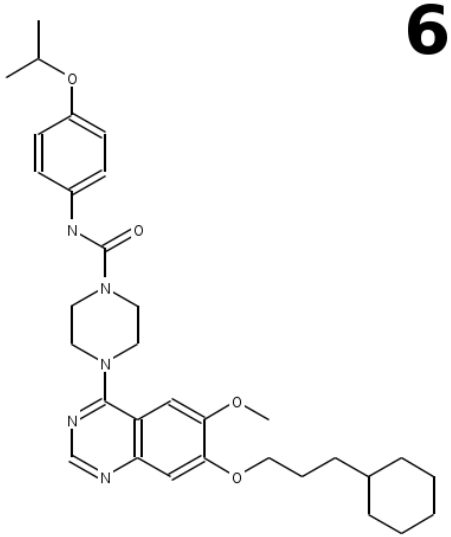


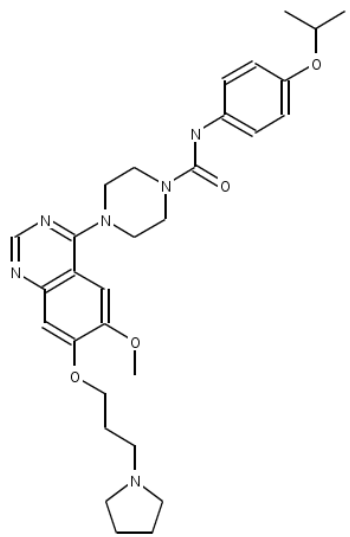
Neighborhood Similarity

PDGFR Inhibitors

- The prediction for **64** is not improved by local regression
- Neighborhood contained 19 molecules, average Tanimoto similarity was 0.94
- **58** has a Tanimoto similarity of 1.0 to **64**
- Range of activities was 1.41 log units
- Given significant structural similarity but differing activity values, local regression fails

64





58

Some caveats . . .

- Sparse training data can lead to larger prediction errors
- Requires that there be structural differences between groups of molecules
- Due to its focus on individual query molecules and the absence of an *a priori* model, structure-activity trends cannot be extracted

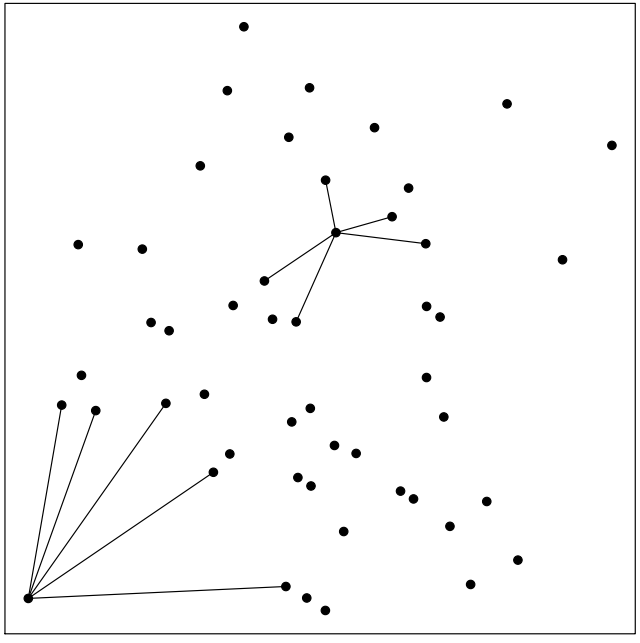
Summary

- Local regression generally shows improved accuracy
- Dependent on the nature of the neighborhood around a query point
- Can be performed very rapidly making it good for large datasets that are increasing in size (HTS data)
- No need to perform *apriori* clustering
- The correlation between neighborhood similarity and predictive ability is not significant

Outline

Nearest Neighbor Methods

- Traditional *k*NN methods are simple, fast, intuitive
- Applications in
  - regression & classification
  - diversity analysis
- Can be misleading if the *nearest* neighbor is far away
- *R*-NN methods may be more suitable



Diversity Analysis

Why is it Important?

- Compound acquisition
- Lead hopping
- Knowledge of the distribution of compounds in a descriptor space may improve predictive models

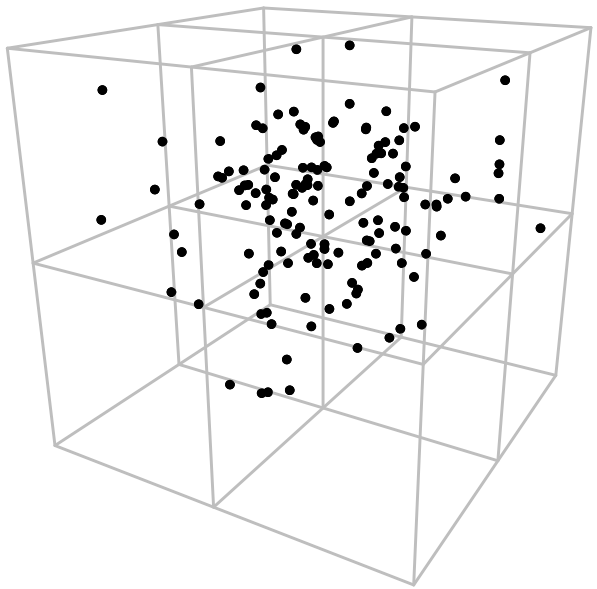
Approaches to Diversity Analysis

Cell based

- Divide space into bins
- Compounds are mapped to bins

Disadvantages

- Not useful for high dimensional data
- Choosing the bin size can be tricky



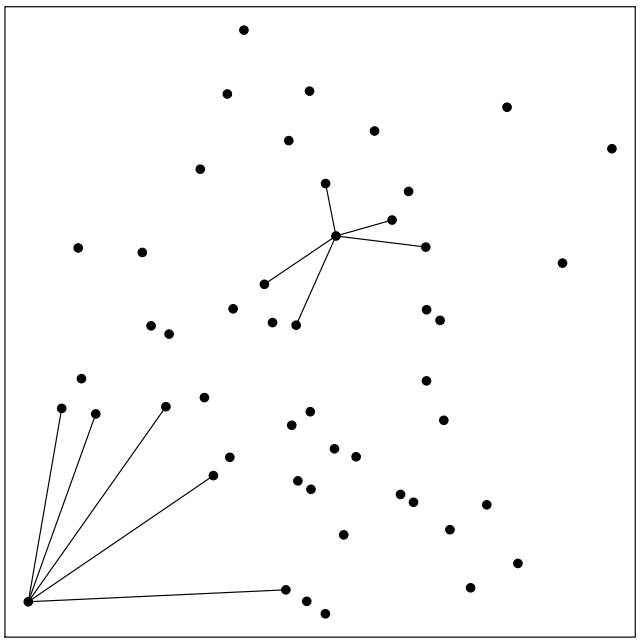
Schnur, D.; *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 36–45

Agrafiotis, D.; Rassokhin, D.; *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 117–12

Approaches to Diversity Analysis

Distance based

- Considers distance between compounds in a space
- Generally requires pairwise distance calculation
- Can be sped up by *k*D trees, MVP trees etc.



Agrafiotis, D. K. and Lobanov, V. S.; *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 51–58

Generating an *R*-NN Curve

Observations

- Consider a query point with a hypersphere, of radius R, centered on it
- For small R, the hypersphere will contain very few or no neighbors
- For larger R, the number of neighbors will increase
- When $R \geq D_{max}$, the neighbor set is the whole dataset

The question is . . .

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

Generating an *R*-NN Curve

Observations

- Consider a query point with a hypersphere, of radius R, centered on it
- For small R, the hypersphere will contain very few or no neighbors
- For larger R, the number of neighbors will increase
- When $R \geq D_{max}$, the neighbor set is the whole dataset

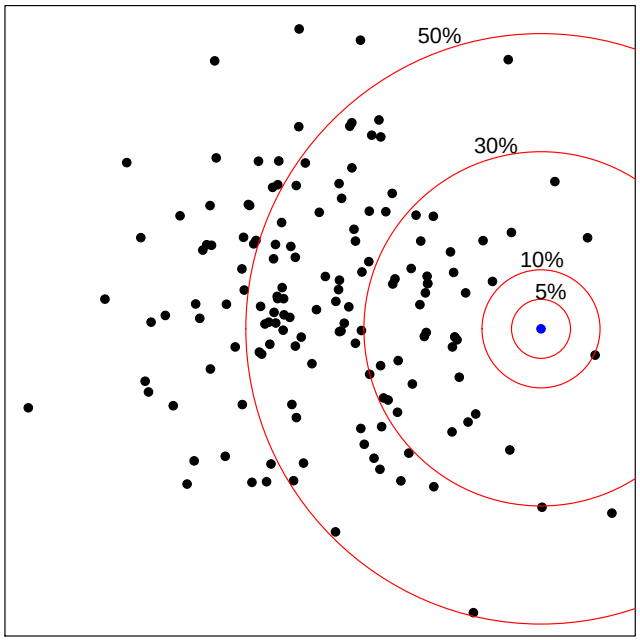
The question is ...

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

Generating an *R*-NN Curve

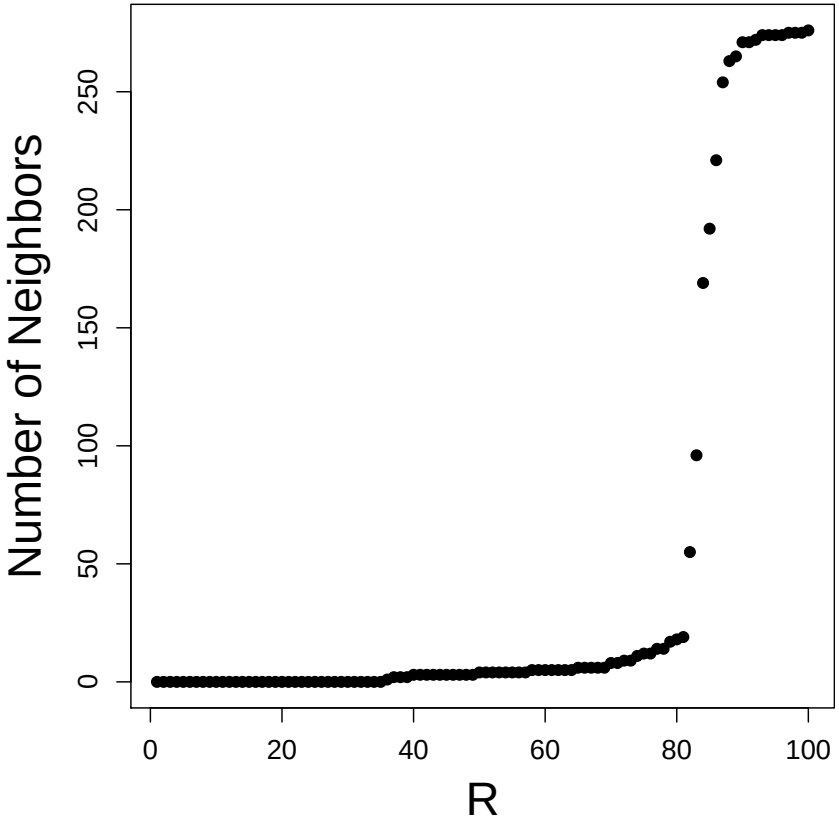
Algorithm

$D_{max} \leftarrow$ max pairwise distance
for molecule *in* dataset **do**
  R $\leftarrow 0.01 \times D_{max}$
  while $R \leq D_{max}$ **do**
    Find NN's within radius R
    Increment R
  end while
end for
Plot NN count vs. R

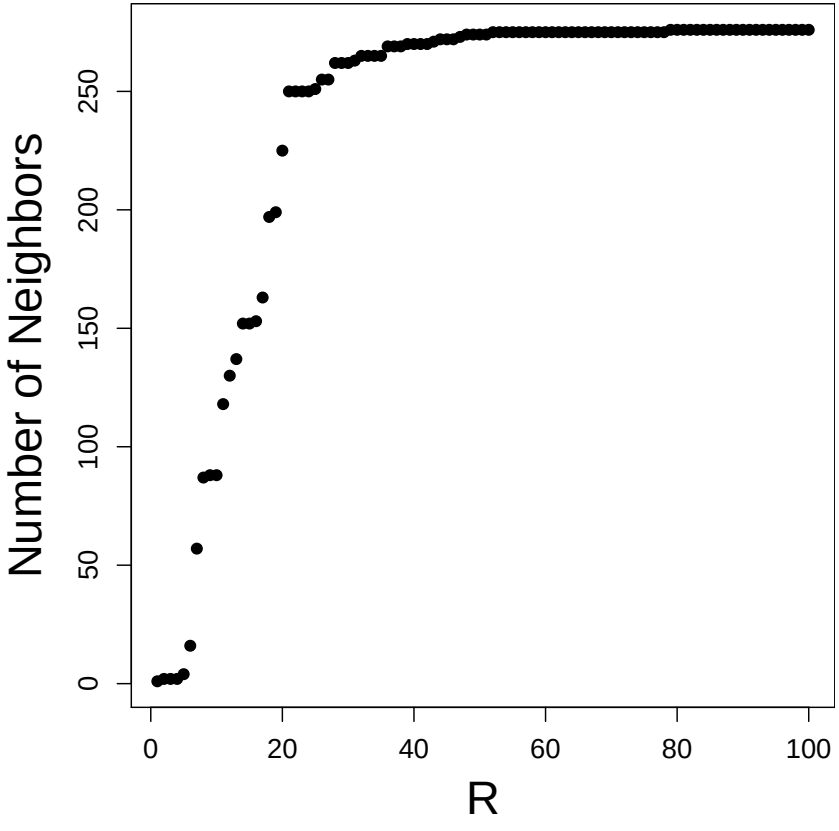


Guha, R. et al.; *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1772

Generating an *R*-NN Curve



Sparse



Dense

Characterizing an *R*-NN Curve

Converting the Plot to a Number

Determine the value of R where the lower tail transitions to the linear portion of the curve

Solution

- Determine the slope at various points on the curve
- Find R for the *first* occurence of the maximal slope ($R_{max(S)}$)
- Can be achieved using a finite difference approach

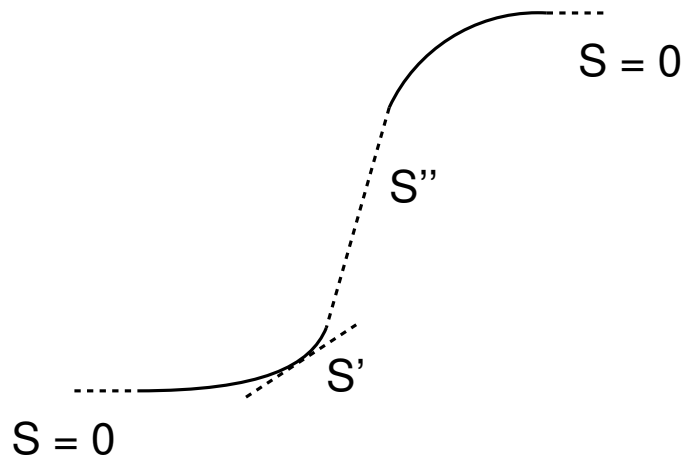
Characterizing an *R*-NN Curve

Converting the Plot to a Number

Determine the value of R where the lower tail transitions to the linear portion of the curve

Solution

- Determine the slope at various points on the curve
- Find R for the *first* occurence of the maximal slope ($R_{max(S)}$)
- Can be achieved using a finite difference approach



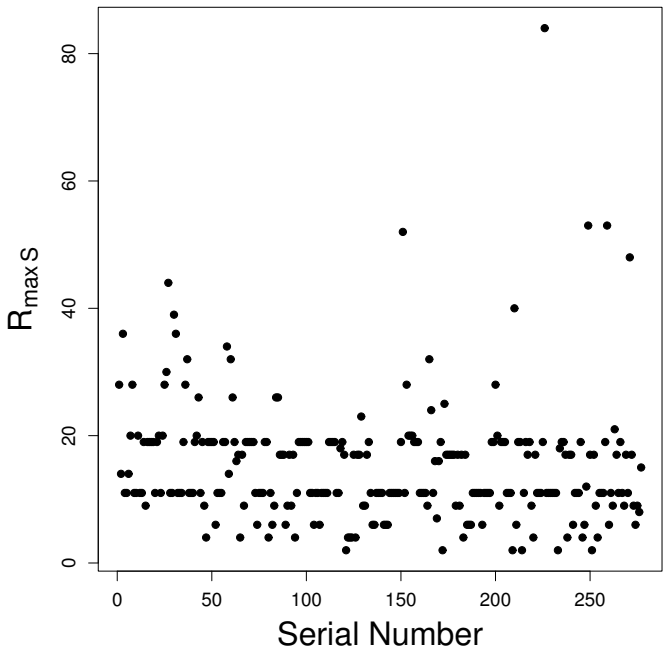
Characterizing Multiple *R*-NN Curves

Problem

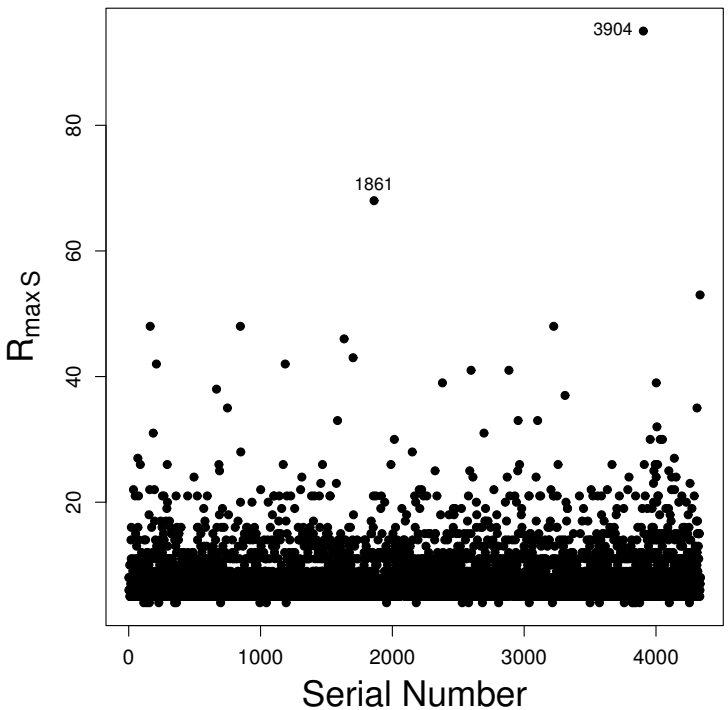
- Visual inspection of curves is useful for a few molecules
- For larger datasets we need to summarize *R*-NN curves

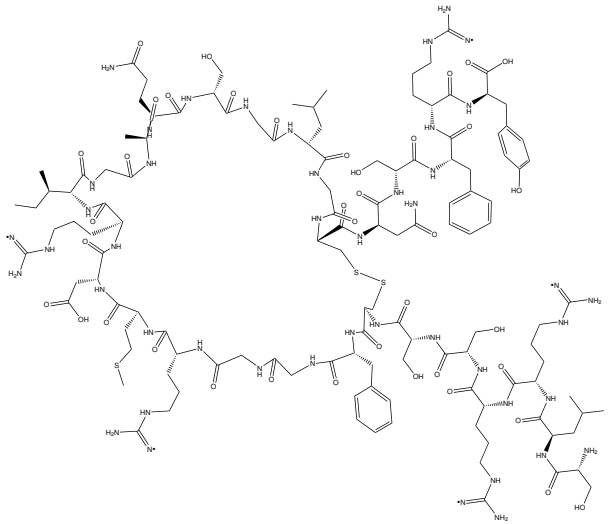
Solution

- Plot $R_{max(S)}$ values for each molecule in the dataset
- Points at the top of the plot are located in the sparsest regions
- Points at the bottom are located in the densest regions

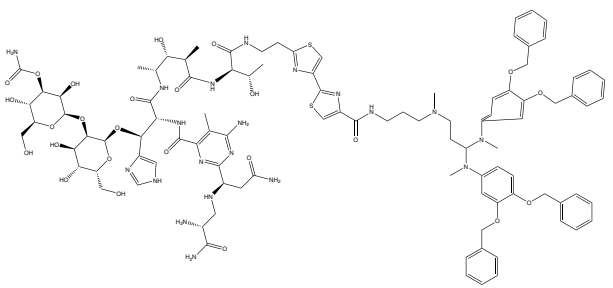


R-NN Curves and Outliers





3904

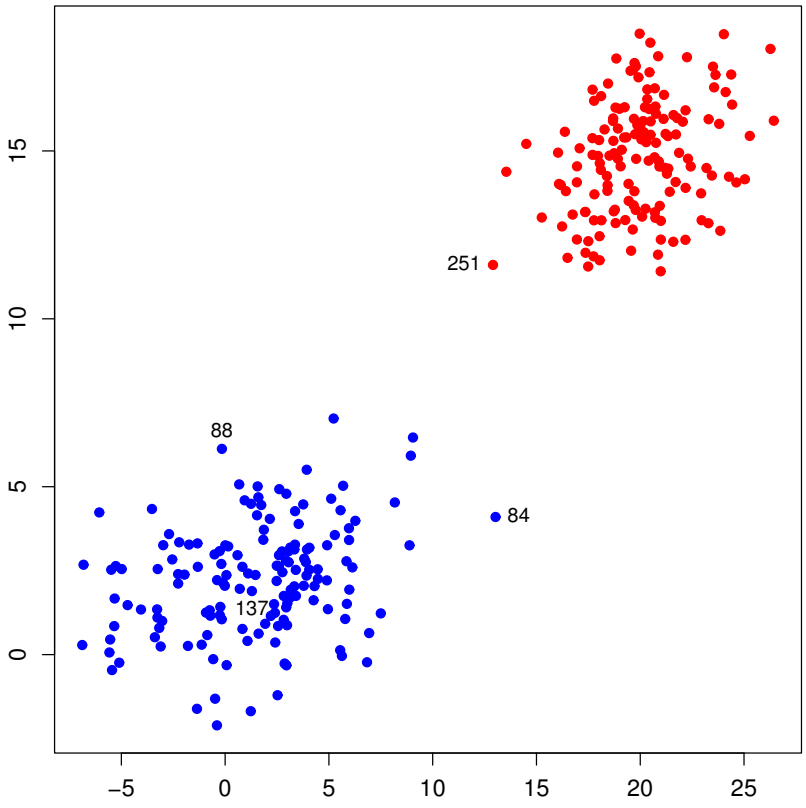


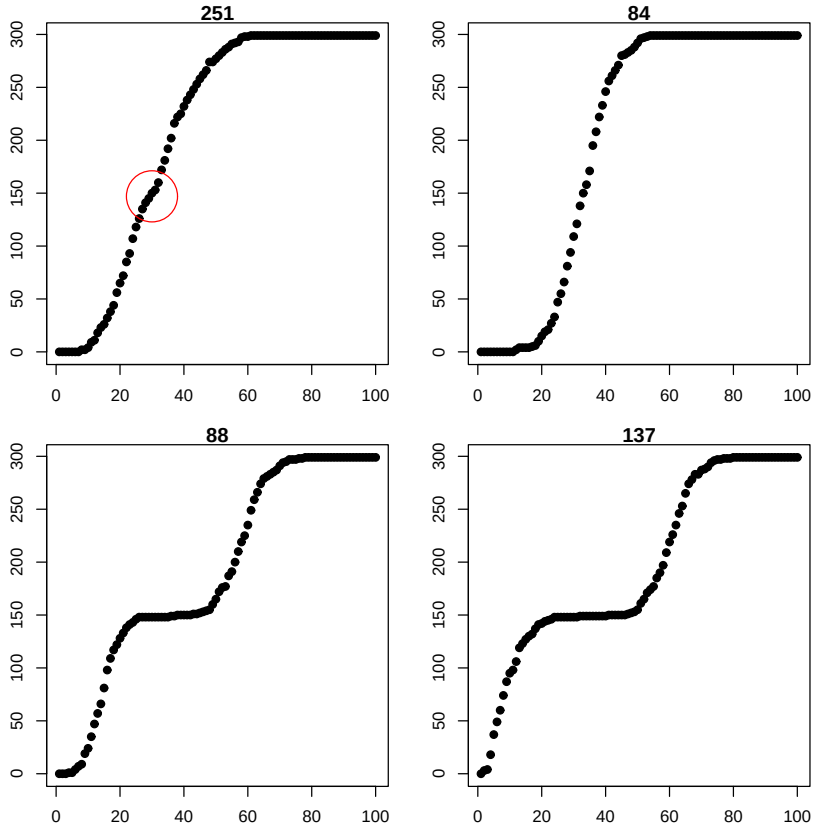
1861

A single plot identifies the location characteristics of *all* the molecules

Kazius, J. et al.; *J. Med. Chem.* **2005,** *48,* 312–320

R-NN Curves and Clusters





Smoothed *R*-NN Curves

R-NN curves are indicative of the number of clusters

R-NN Curves and Clusters

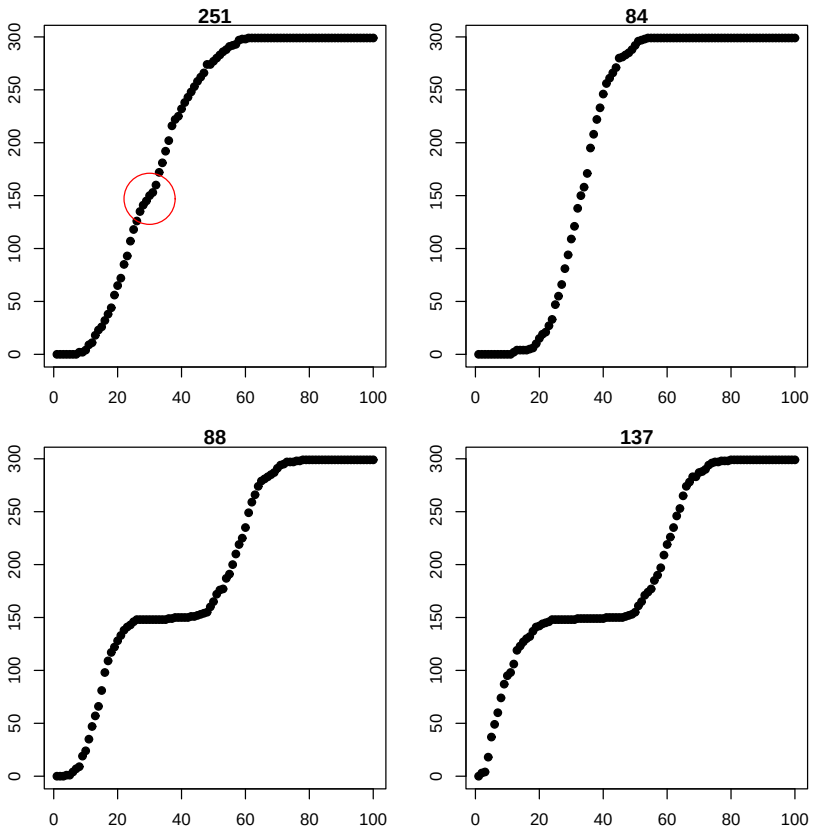
Counting the steps

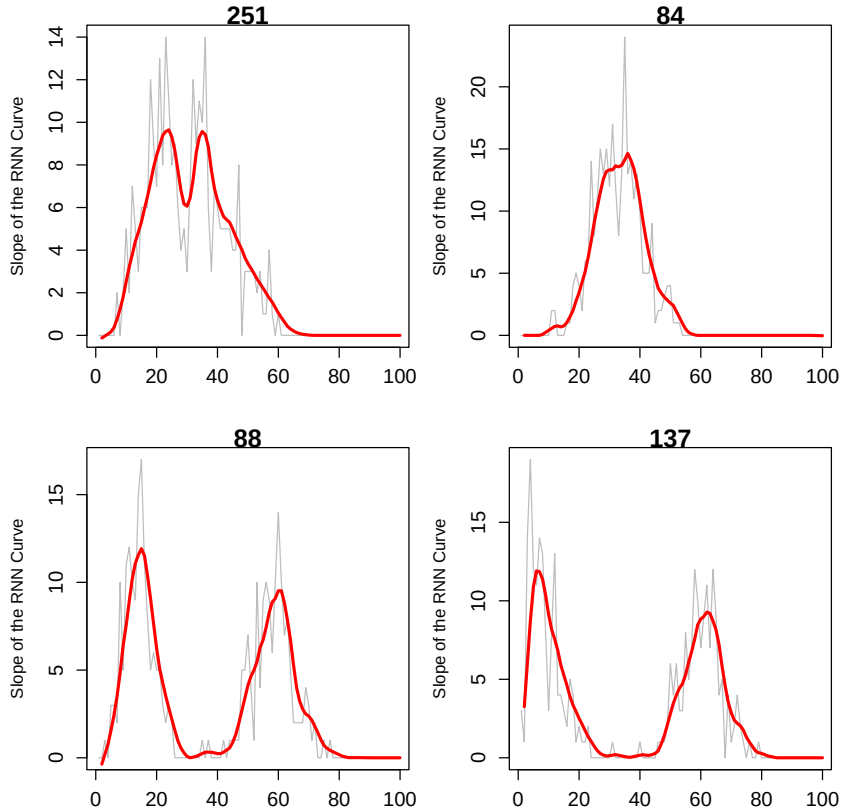
- Essentially a curve matching problem
- All points will not be indicative of the number of clusters
- Not applicable for concentric clusters

Approaches

- Hausdorff / Fréchet distance - requires *canonical* curves
- RMSE from distance matrix
- Slope analysis

R-NN Curves and Their Slopes

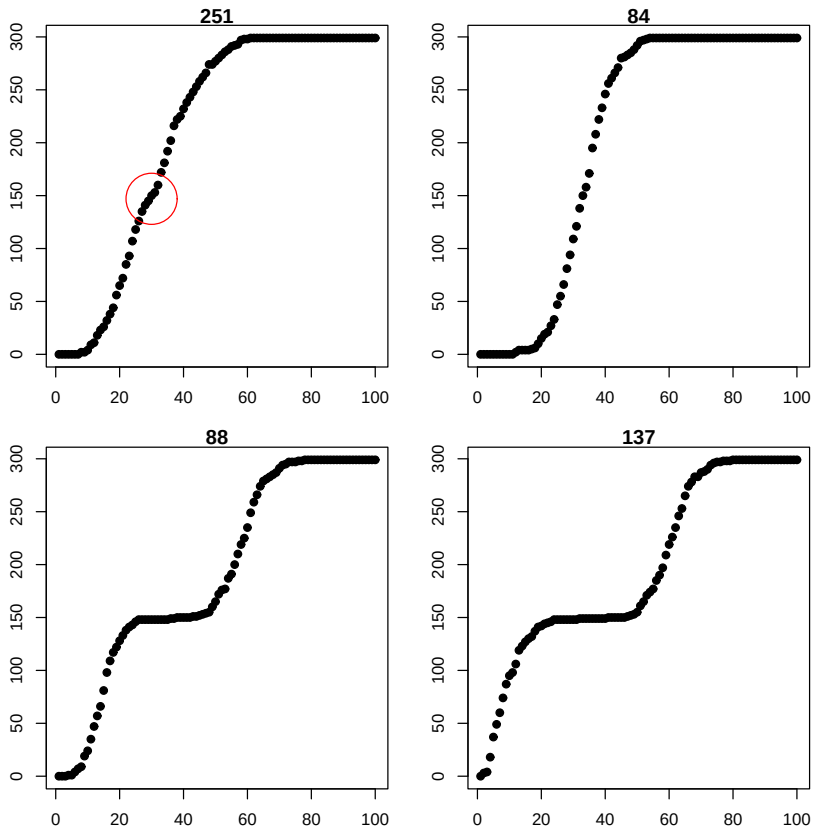


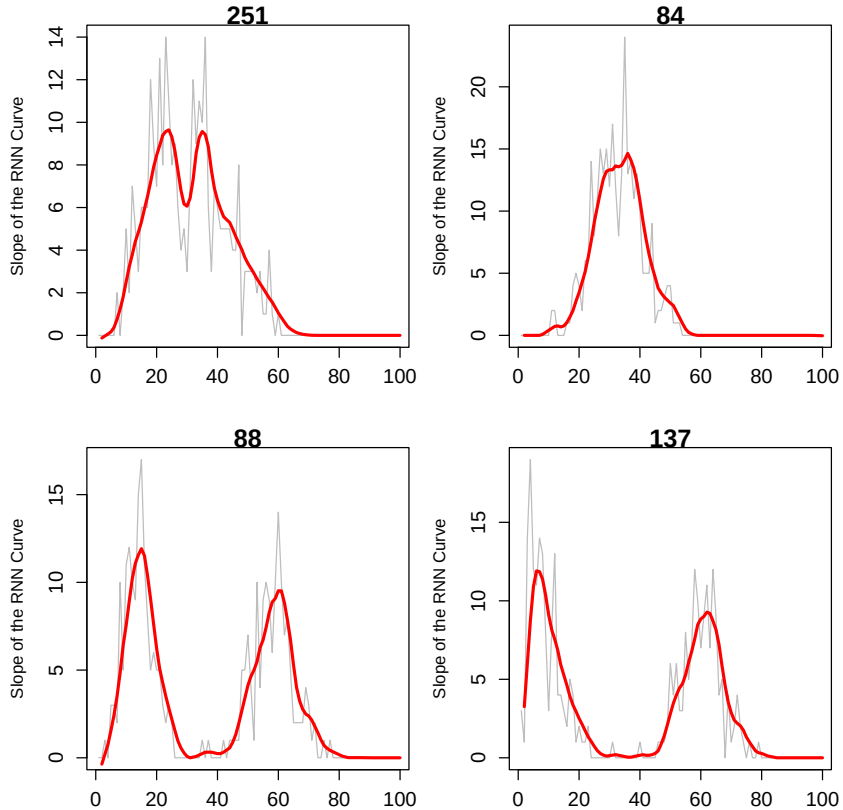


Smoothed first derivative of the *R*-NN Curves

Guha, R. *et al.*, *J. Chem. Inf. Model.*, submitted

R-NN Curves and Their Slopes





Smoothed first derivative of the *R*-NN Curves

- Identifying peaks identifies the number of clusters
- Automated picking can identify spurious peaks

Guha, R. *et al.*, *J. Chem. Inf. Model.*, submitted

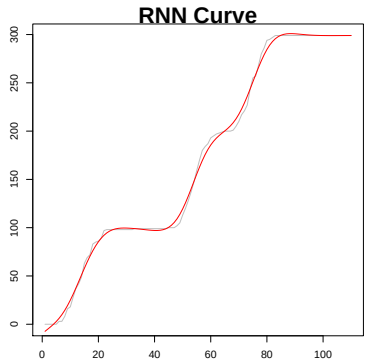
Slope Analysis of *R*-NN Curves

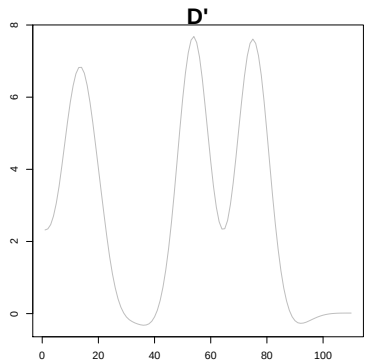
Procedure

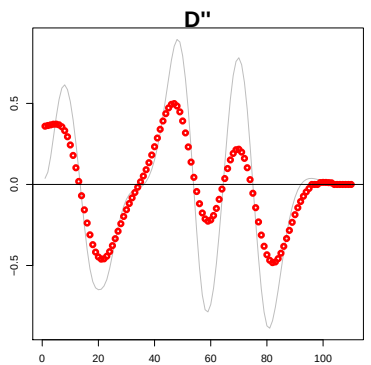
for i *in* molecules **do**
  Evaluate *R*-NN curve
  $F \leftarrow$ smoothed *R*-NN curve
  Evaluate F''
  Smooth F''
  $N_{root,i} \leftarrow$ no. of roots of F''
end for
$N_{cluster} = [\max(N_{root}) + 1]/2$

Possible improvements

- Sample from the collection of *R*-NN curves
- Improve handling of concentric clusters

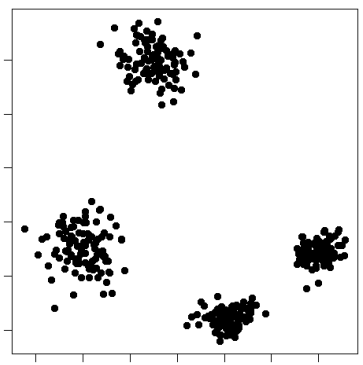




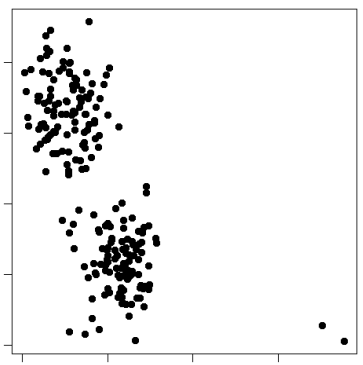


Simulated Data

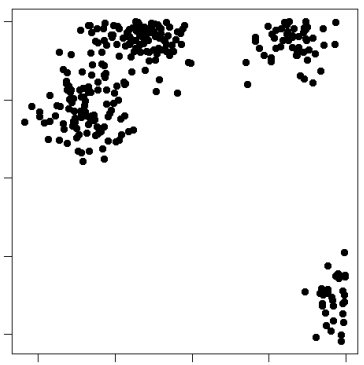
- Simulated 2D data using a Thomas point process
- Predicted k, followed by kmeans clustering using k
- Investigated similar values of k



k	ASW
4	0.61
3	0.74
5	0.70



k	ASW
3	0.44
2	0.65
4	0.47



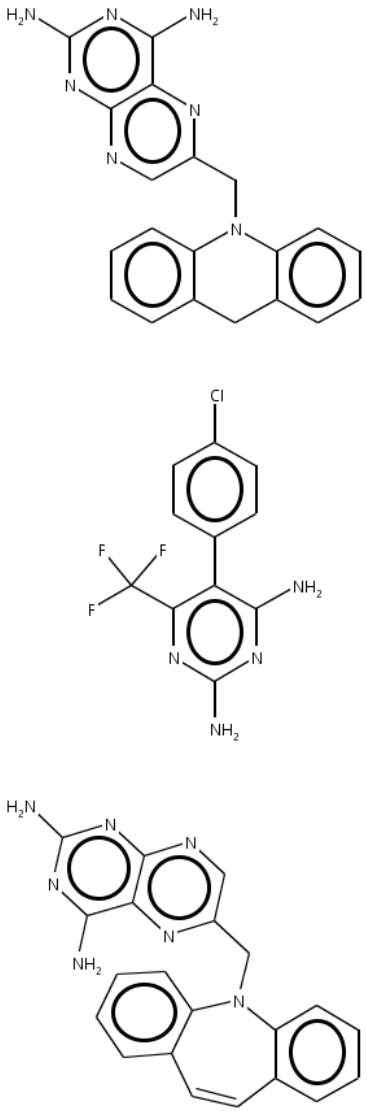
k	ASW
4	0.64
3	0.48
5	0.56

ASW - average silhouette width, higher is better; k - number of clusters

A Mixed Dataset

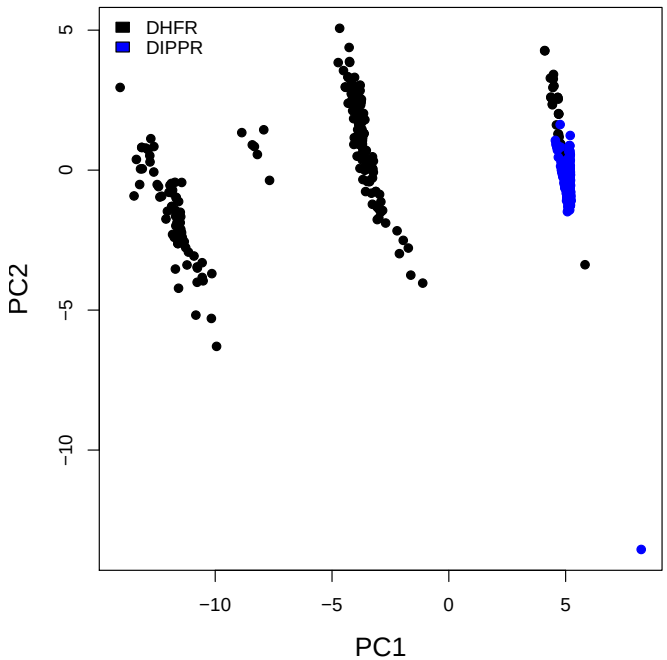
Dataset composition

- 277 DHFR inhibitors based on
  - substituteed pyrimidinediamine and
  - diaminopteridine scaffolds
- 277 molecules from the DIPPR project
  - mainly simple hydrocarbons
  - boiling point was modeled
- Evaluated 147 Molconn-Z descriptors, reduced to 24
- We expect at least 3 clusters



A Mixed Dataset

Desc. Set	k	ASW	Purity
4 descriptors	2	0.71	0.63
	3	0.67	0.89
	4	0.73	0.84
6 descriptors	2	0.67	0.94
	3	0.70	0.97
	4	0.61	0.94
All 24 descriptors	2	0.29	0.96
	3	0.33	0.96
	4	0.23	0.90



- PC plot indicates 3 main clusters
- In all cases, 3 clusters is optimal for both quality measures

Summary

R-NN curves . . .

- Simple way to characterize spatial distributions and identify outliers
- Applicable to datasets of arbitrary dimensions and size, via approximate NN algorithms such as LSH
- Summarizing a dataset does not require user-defined parameters

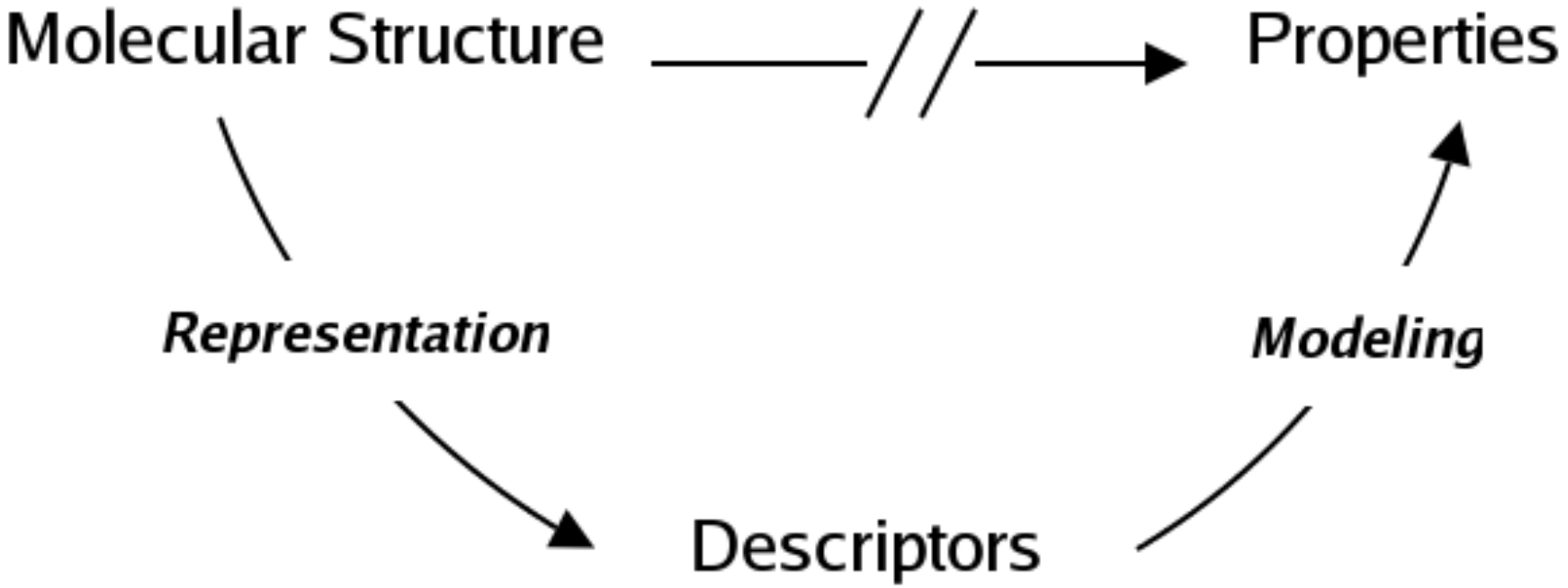
. . . Clustering

- Provides an approach to *a priori* identification of the number of clusters, avoiding trial and error
- Appears to be more reliable than the silhouette width
- Probably not useful for hierarchical clusterings

Guha, R. et al.; *J. Chem. Inf. Model.*, **2006**, *46*, 321–333

Outline

The QSAR Pipeline



Descriptor Selection

Why do we need it?

- We can calculate thousands of descriptors
- Many are correlated
- Many are very abstract and do not have a simple physical interpretation
- By using too many descriptors we can overfit a model

We need to select a small subset of uncorrelated and (hopefully) interpretable descriptors - a.k.a, Occams Razor

Descriptor Selection

Why do we need it?

- We can calculate thousands of descriptors
- Many are correlated
- Many are very abstract and do not have a simple physical interpretation
- By using too many descriptors we can overfit a model

We need to select a small subset of uncorrelated and (hopefully) interpretable descriptors - a.k.a, Occams Razor

Descriptor Selection

What methods are available?

- Stepwise regression
- Stochastic methods
  - Genetic algorithms
  - Simulated annealing
  - Tabu search
- The common feature is that these methods search for a descriptor subset that is optimal for a *single* model

QSAR Model Development

What type of models?

- For a given problem we can consider different model types
- For classification we can have
  - linear discriminants
  - neural networks
  - random forests
- For regression we can have
  - linear regression, partial least squares
  - neural networks
  - support vector machines

Ensemble Models

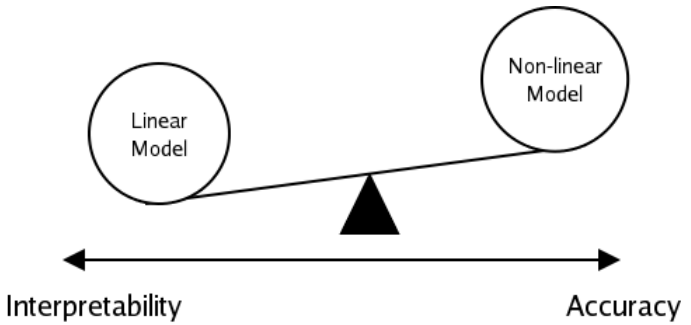
What are they?

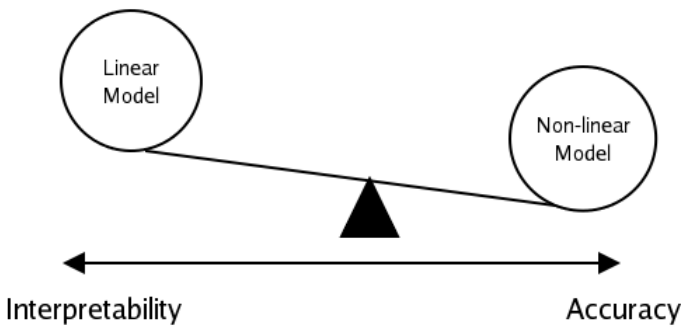
- Traditionally we use a single model of a single type to obtain predictions
- It is now common to *pool* predictions from multiple models
  - of the same type
  - of different types
- *Pooling* can either be majority vote (classification) or the arithmetic mean (regression)
- Essentially we combine several models to get a stronger one
- Examples include
  - Random forest
  - ADABoost

Ensemble Models

How can they be used?

- Combining predictions from multiple models is statistically robust
- We can use multiple model types for different purposes
  - Use one for predictive ability
  - Use another for interpretability





Dietterich, T., *The Handbook of Brain Theory and Neural Networks*, MIT Press; Cambridge, MA; 2002.

Descriptor Selection

Descriptors & models

- Traditionally, ensemble models will use different (but possibly) similar descriptors for each of the models in the ensemble
- This means that the individual models might be encoding slightly different structure-actvitity trends
- It is difficult to derive a single consistent interpretation of the encoded structure-activity trends
- What if we want to use multiple model types but with the same set of descriptors?
  - We get a consistent encoding of the SAR's
  - How much will we loose in terms of accuracy?

Descriptor Selection

Descriptors & models

- Traditionally, ensemble models will use different (but possibly) similar descriptors for each of the models in the ensemble
- This means that the individual models might be encoding slightly different structure-actvitity trends
- It is difficult to derive a single consistent interpretation of the encoded structure-activity trends
- What if we want to use multiple model types but with the same set of descriptors?
  - We get a consistent encoding of the SAR's
  - How much will we loose in terms of accuracy?

What is Consensus Descriptor Selection?

Select a set of descriptors that are *simultaneously* optimal for multiple model types

Examples

- Maximize the % true actives from a LDA and a CNN model
- Minimize the RMSE from an OLS and SVM model

Consensus Descriptor Selection

- Different model types now encode the same structure-activity trends
- The problem being discussed here does not involve *contradictory* objectives
  - No need to consider Pareto optimality
- Note that the optimal descriptor set is for the combined set of models, not for individual models
- We can easily use the individual models for different purposes

Osborne, M. and Rubenstein, A.; *A Course in Game Theory*, MIT Press: Cambridge, MA, 1994

The Formulation

- We used a genetic algorithm to perform descriptor selection
- Consensus selection is performed by using a composite score function
  - For classification we minimize

$$f(S_i) = \frac{1.0}{atan\left(\left[\mathrm{TC_{CNN}} + \mathrm{TC_{LDA}} \right] / 2 \right)}$$

  - For regression we minimize

$$f(S_i) = atan\left(\mathrm{RMSE_{CNN}} + \mathrm{RMSE_{OLS}} \right)$$

Dutta, D. et al.; *J. Chem. Inf. Model.*, submitted

Datasets

PDGFR Inhibitors

- 79 molecules
- Calculated 321 descriptors using ADAPT, reduced to 40 descriptors
- Previous work had developed OLS and CNN models

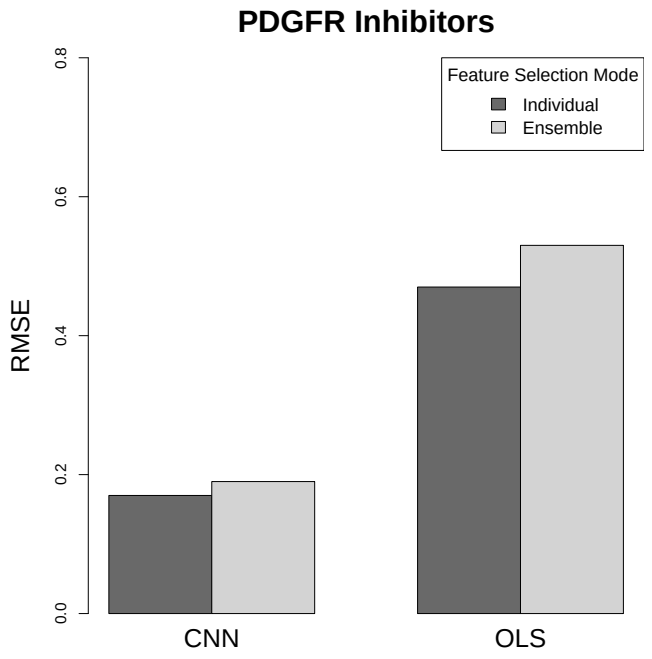
COX-2 Inhibitors

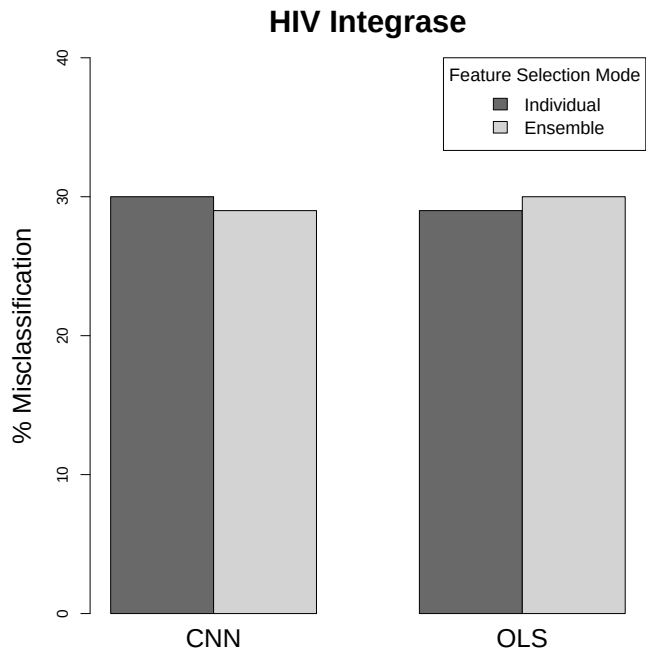
- 273 molecules, inhibited COX-2
- Calculated 321 descriptors, reduced to 54
- Original work presented OLS, CNN and k-NN models.

Guha, R. and Jurs, P.C.; *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2179–2189

Kauffman, G. and Jurs, P.C.; *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1553–1560

Summary





The key features is that the degradation in performace when using a single descriptor set for different model types is minimal

PDGFR Inhibitors

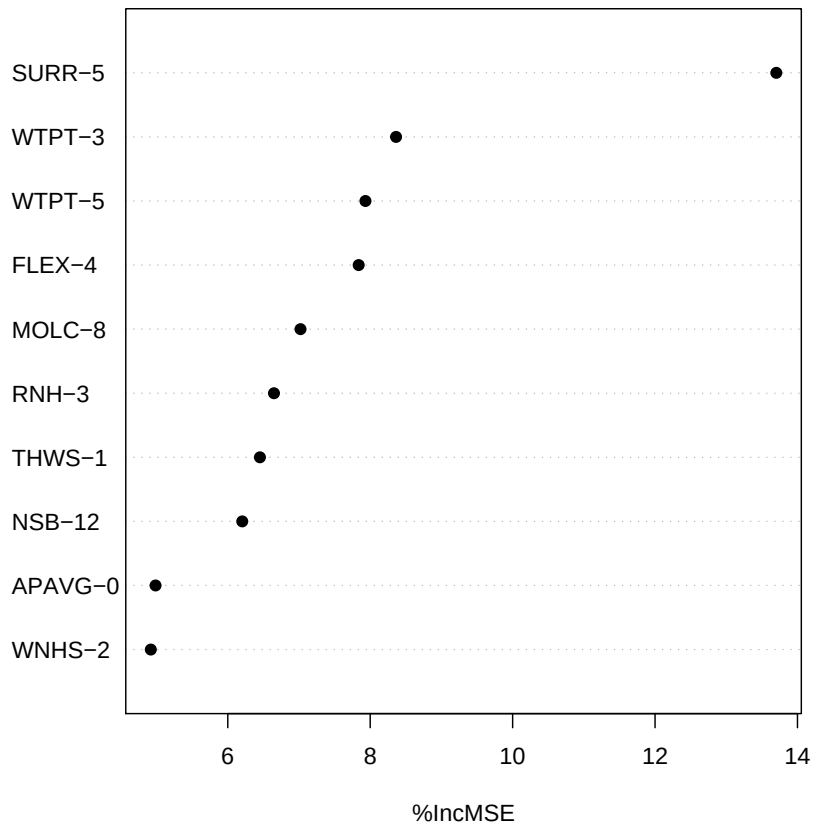
Selection type	Model Type	Descriptors	RMSE
Individual	Neural network	WTPT-3, SURR-5, FLEX-4	0.33 (0.52)
	Linear regression	RNHS-3, SURR-5, NSB	0.54 (0.32)
Ensemble	Neural network	ACHG, NSB, SURR-5	0.37 (0.38)
	Linear regression		0.58 (0.53)

- RMSE for PSET is lower than TSET due to set composition
- RMSE for random PSETS ranged between 0.32 and 0.79 log units
- Linear models were statistically significant - F-value = 8.67 (3,59) and 10.66 (3,59)

PDGFR Inhibitors

The descriptors . . .

- Both individual models included SURR-5
- PLS analysis of the linear model indicates that it is the most important descriptor
- Ensemble selection also chooses SURR-5
- Interpretation of the CNN and OLS model from ensemble selection lead to similar trends
    - Smaller, less polar molecules are predicted to be more active



PDGFR Inhibitors - PLS Analysis

Descriptor	Component 1	Component 2	Component 3
ACHG	-2.619	-7.304	4.218
SURR-5	-7.577	-0.542	-4.506
NSB	6.211	-5.060	-3.718

Loadings obtained from a partial least squares analysis using the three descriptors from the linear regression model obtained using the ensemble descriptor selection method.

PDGFR Inhibitors - CNN Interpretation

Descriptor	H1 (0.52)	H3 (0.48)	H2 (0.00)
ACHG	245.115	-236.285	-0.597
NSB	50.903	-52.638	4.442
SURR-5	211.904	-199.019	-1.235

Effective weight matrix for the neural network model obtained using the ensemble descriptor selection method

COX-2 Inhibitors

Selection type	Model Type	Descriptors	RMSE
Individual	Neural network	NDB, PND-6, WTPT-5, V6C, V4PC, MDE-11, MDE-34,MDEO-12	0.65 (0.85)
	Linear regression	NCI, V7CH, PND-3, MDEO-22, MDEO-12, EMIN, EMAX, WTPT-3	0.88 (0.97)
Ensemble	Neural network	WTPT-5, WTPT-4, WTPT-3, NC, MREF, PND-5, PND-3, MDEO-22	0.65 (0.76)
	Linear regression		0.87 (0.81)

- Ensemble optimization leads to similar descriptors
- Resulting models compare well with individually optimized models

COX-2 Inhibitors

Structure-activity trends . . .

- Original work did not provide any interpretations
- Performing an interpretation of the OLS model's indicate higher activity is correlated to
  - higher polarity
  - larger size
- An interpretation of the CNN model reveals the same trends

These conclusions correspond well to the fact that the design of selective inhibitors is focused on the difference in size between the central channels of COX-2 and COX-1

Kauffman, G. and Jurs, P.C.; *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1553–1560

COX-2 Inhibitors

Structure-activity trends . . .

- Original work did not provide any interpretations
- Performing an interpretation of the OLS model's indicate higher activity is correlated to
  - higher polarity
  - larger size
- An interpretation of the CNN model reveals the same trends

These conclusions correspond well to the fact that the design of selective inhibitors is focused on the difference in size between the central channels of COX-2 and COX-1

Kauffman, G. and Jurs, P.C.; *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1553–1560

COX-2 Inhibitors - PLS Analysis

Descriptor	Comp 1	Comp 2	Comp 3
WTPT-3	11.390	-6.402	-4.677
WTPT-4	1.799	0.545	-12.161
WTPT-5	11.553	-7.174	5.132
MREF	-6.284	2.987	-7.697
NC	-9.361	7.445	-5.090
PND-3	1.667	-14.431	1.170
PND-5	8.351	4.959	-9.145
MDEO-22	0.533	-1.563	-10.760

Loadings obtained from a partial least squares analysis of the using the eight descriptors from the linear regression model obtained using the ensemble descriptors selection method

COX-2 Inhibitors - CNN Interpretation

Descriptor	H1 (0.71)	H3 (0.29)	H2 (0.00)
WTPT-5	17.148	-89.547	6.743
NC	68.857	257.125	-2.686
MREF-1	-22.425	-257.601	-0.050
WTPT-4	19.637	55.858	5.751
PND-5	-515.789	152.250	0.981
WTPT-3	30.742	145.523	-3.999
PND-3	8.220	14.565	-6.383
MDEO-22	-8.866	-18.464	-6.091

Effective weight matrix for the neural network model obtained using the ensemble descriptor selection method on the COX-2 dataset

Summary

- Ensemble descriptor selection provides an easy way to obtain a consistent set of descriptors for multiple model types
- Minimal degradation of predictive performance
- Currently multiple model types contribute equally to the objective function - can be easily changed
- By virtue of using the same descriptors, multiple model types encode the same structure-activity trends
- In comparison to traditional ensemble models, we do not loose interpretability