Characterizing the Density of Chemical Spaces and its Use in Outlier Analysis and Clustering

Rajarshi Guha

School of Informatics
Indiana University

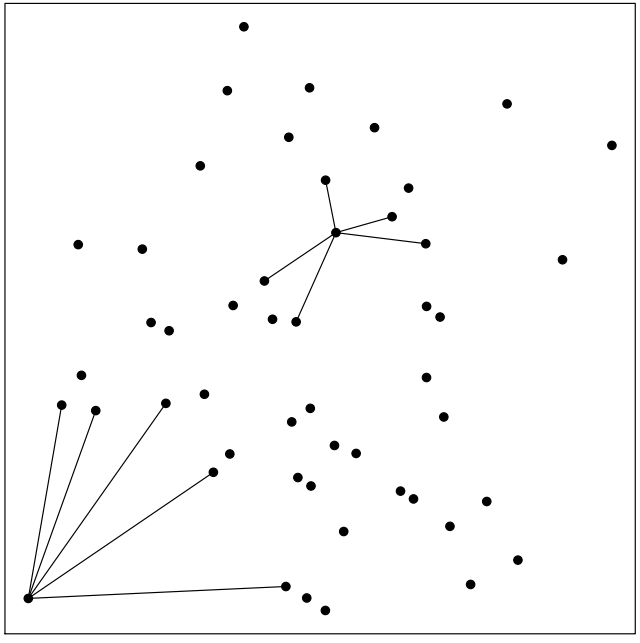
17th August, 2007
Cambridge, MA

Outline

- R-NN curves for diversity analysis
- Counting clusters with R-NN curves
- Density and domain applicability

Nearest Neighbor Methods

- Traditional kNN 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

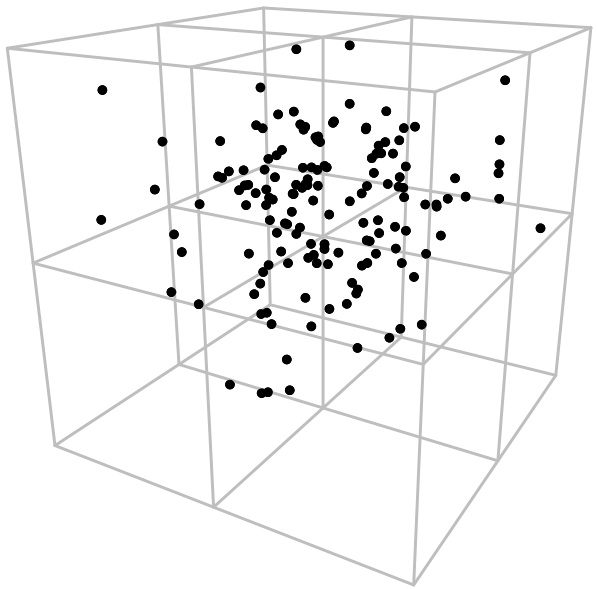
Approaches to Diversity Analysis

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



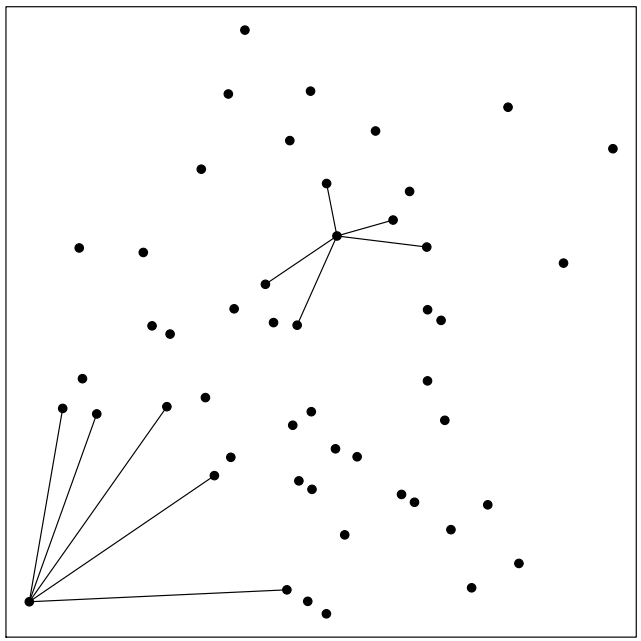
Schnur, D.; *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 36–45

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

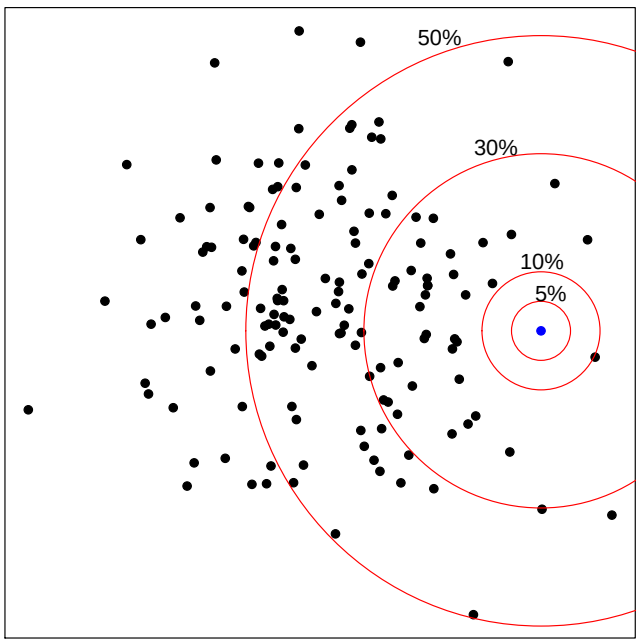
The question is ...

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

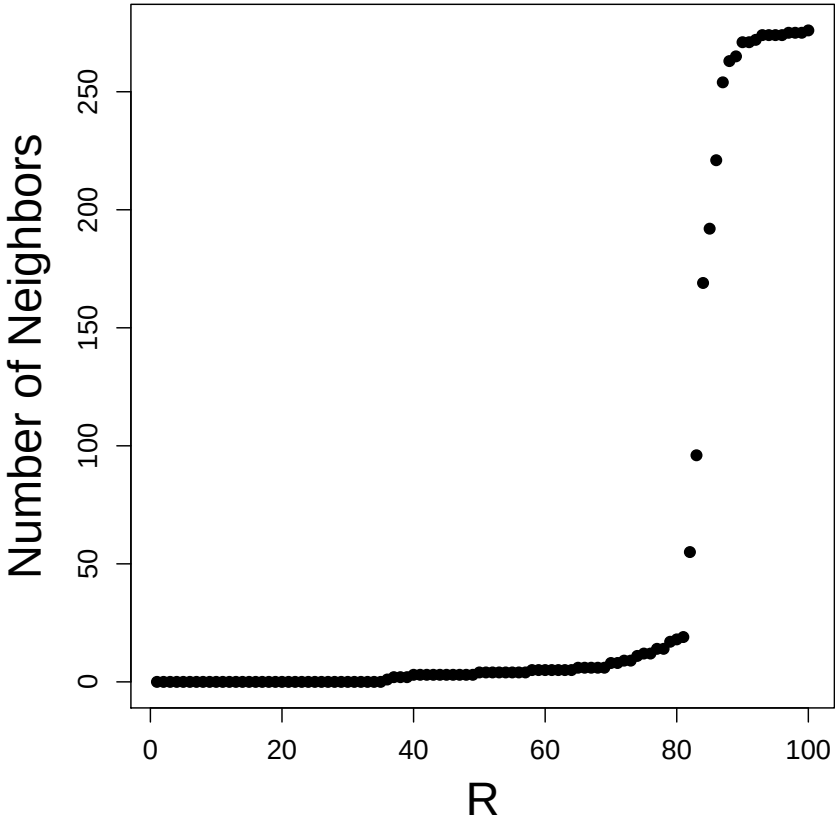
Algorithm

$D_{max} \leftarrow$ 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

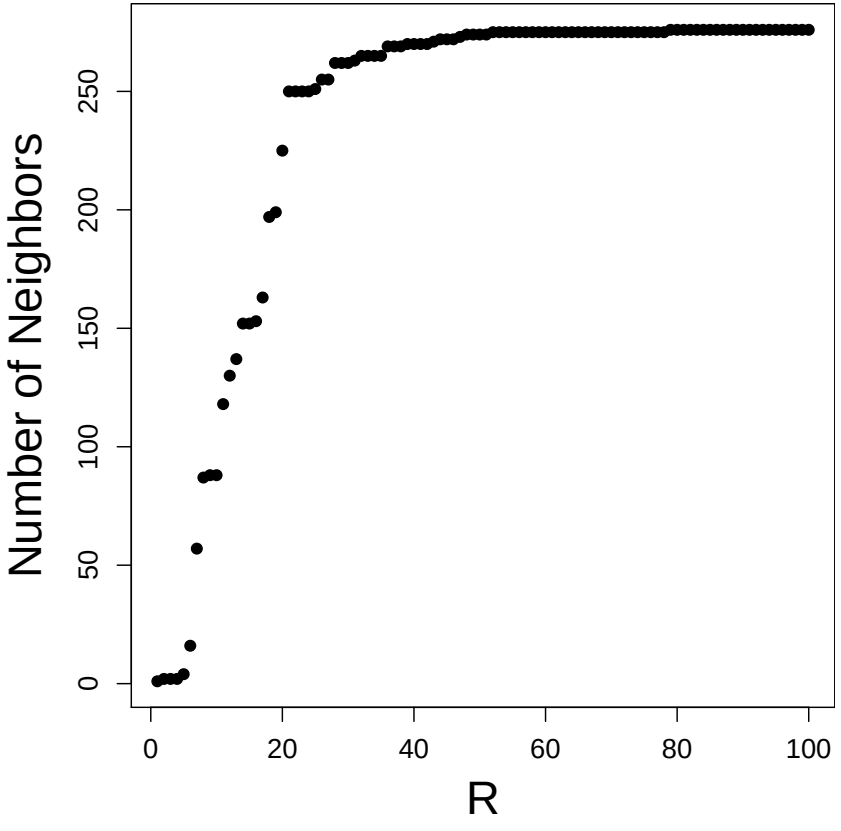


Guha, R. et al.; *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1772

Generating an *R*-NN Curve



Sparse



Dense

Characterizing an *R*-NN Curve

Converting the Plot to Numbers

- Since *R*-NN curves are sigmoidal, fit them to the logistic equation

$$N_N = a \cdot \frac{1 + me^{-R/\tau}}{1 + ne^{-R/\tau}}$$

- m, n should characterize the curve
- Problems
  - Two parameters
  - Non-linear fitting is dependent on the starting point
  - For some starting points, the fir does not converge and requires repetition

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

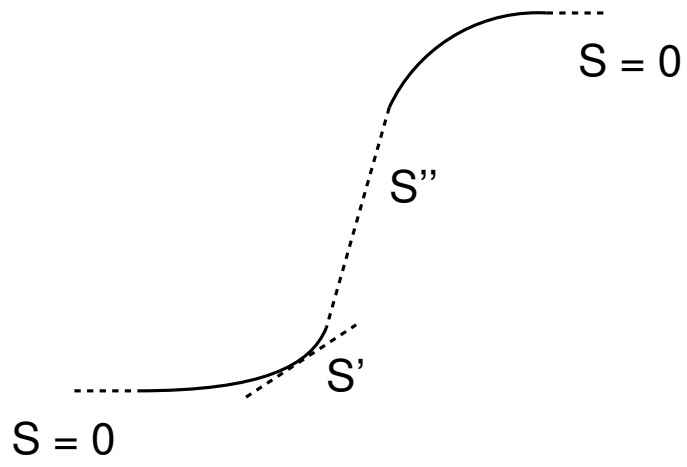
Characterizing an *R*-NN Curve

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



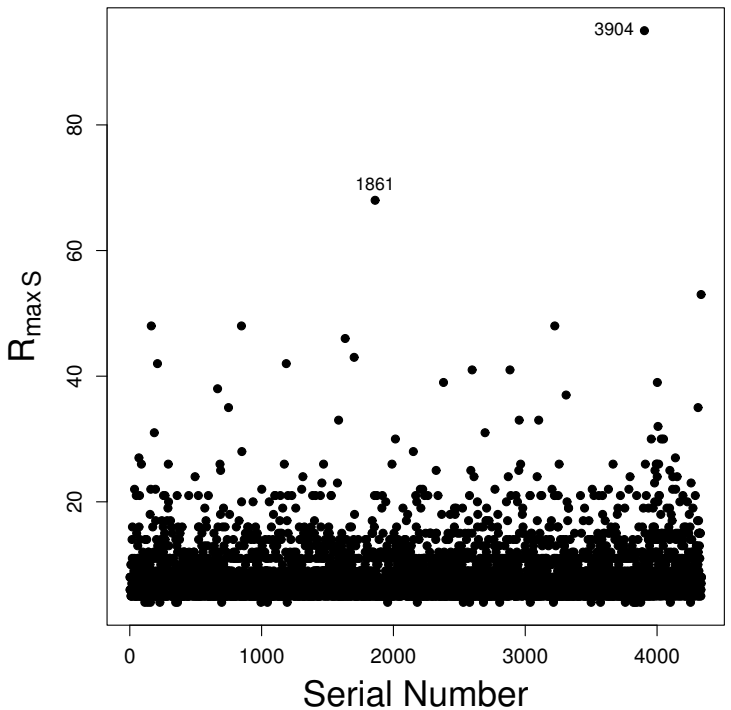
Characterizing Multiple *R*-NN Curves

Problem

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

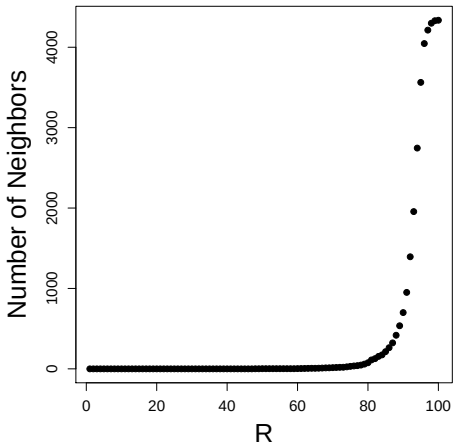
Solution

- 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

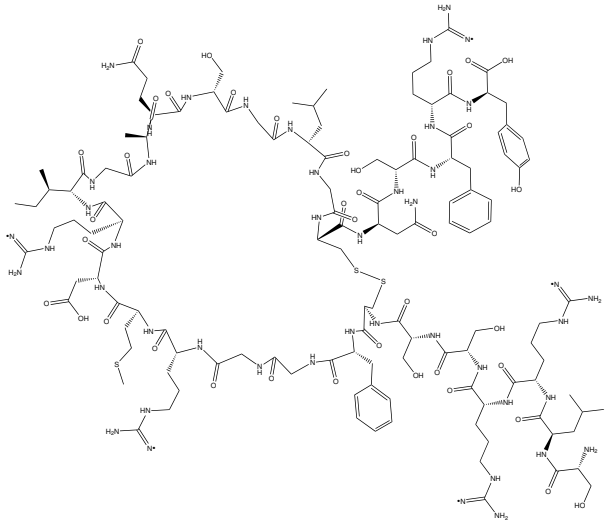


Kazius, J. et al.; *J. Med. Chem.* **2005,** *48,* 312–320

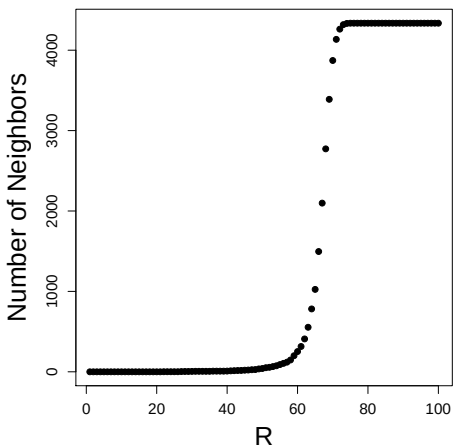
R-NN Curves and Outliers



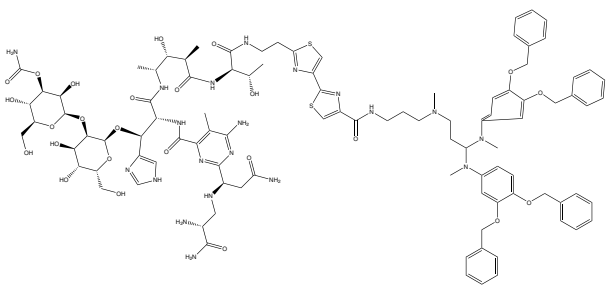
3904



3904



1861



1861

How Can We Use It For Large Datasets?

Breaking the $O(n^2)$ barrier

- Traditional NN detection has a time complexity of $O(n^2)$
- Modern NN algorithms such as kD-trees
  - have lower time complexity
  - restricted to the exact NN problem
- Solution is to use *approximate NN* algorithms such as Locality Sensitive Hashing (LSH)

Bentley, J.; *Commun. ACM* **1980,** *23,* 214–229

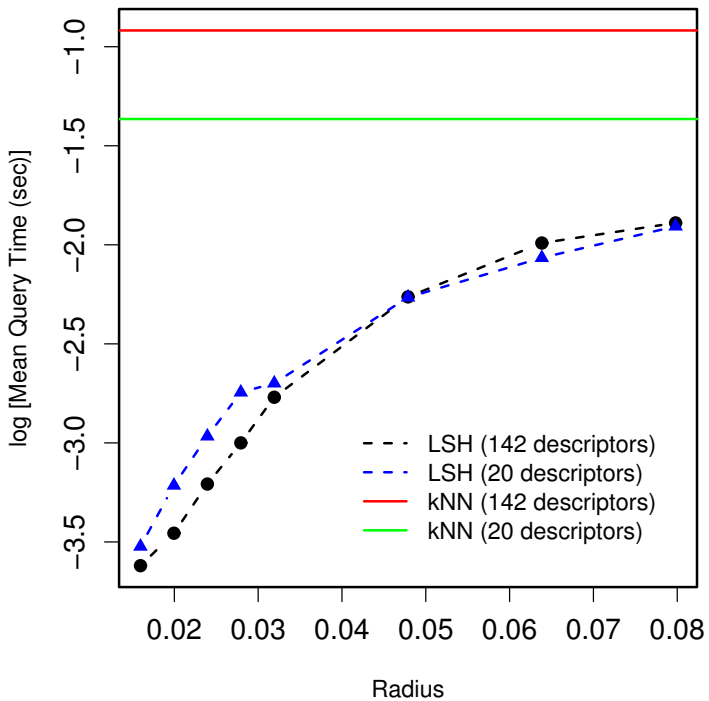
Datar, M. et al.; *SCG '04: Proc. 20th Symp. Comp. Geom.*; ACM Press, 2004

Dutta, D.; Guha, R.; Jurs, P.; Chen, T.; *J. Chem. Inf. Model.* **2006,** *46,* 321–333

How Can We Use It For Large Datasets?

Why LSH?

- Theoretically sublinear
- Shown to be 3 orders of magnitude faster than traditional *k*NN
- Very accurate ($> 94\%$)



Comparison of NN detection speed on a 42,000 compound dataset using a 200 point query set

Dutta, D.; Guha, R.; Jurs, P.; Chen, T.; *J. Chem. Inf. Model.* **2006,** *46,* 321–333

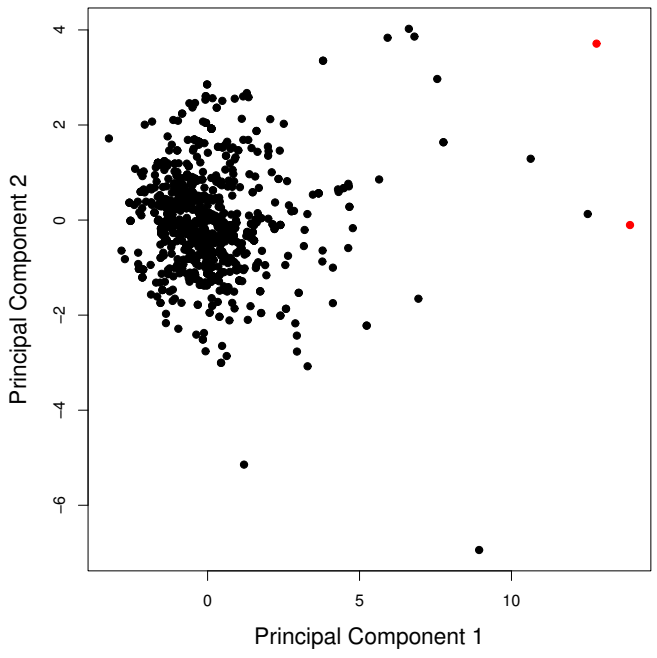
Alternatives?

Why not use PCA?

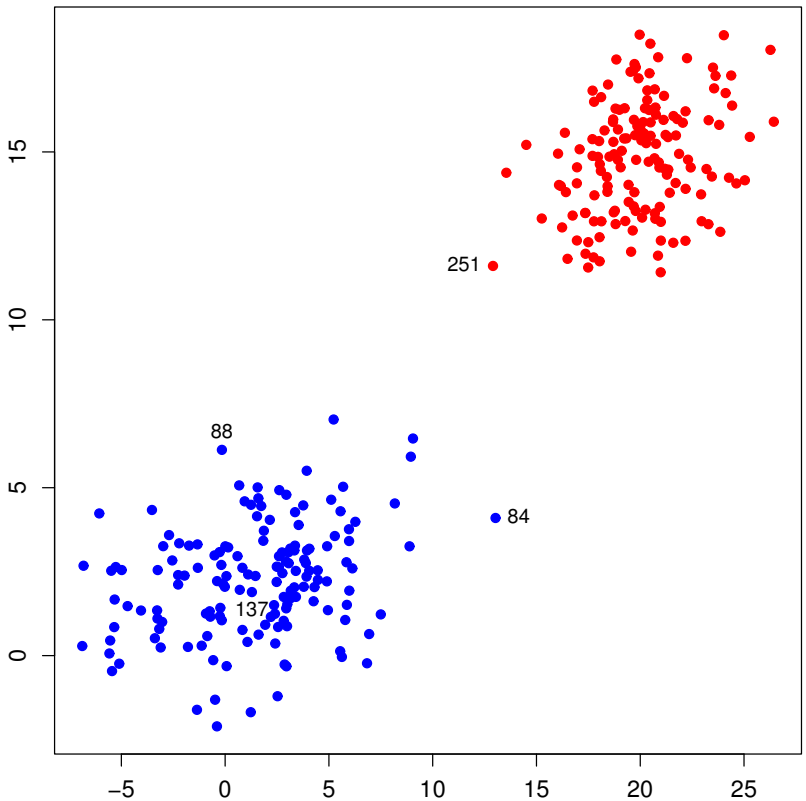
- *R*-NN curves are fundamentally a form of dimension reduction
- Principal Components Analysis is also a form of dimension reduction

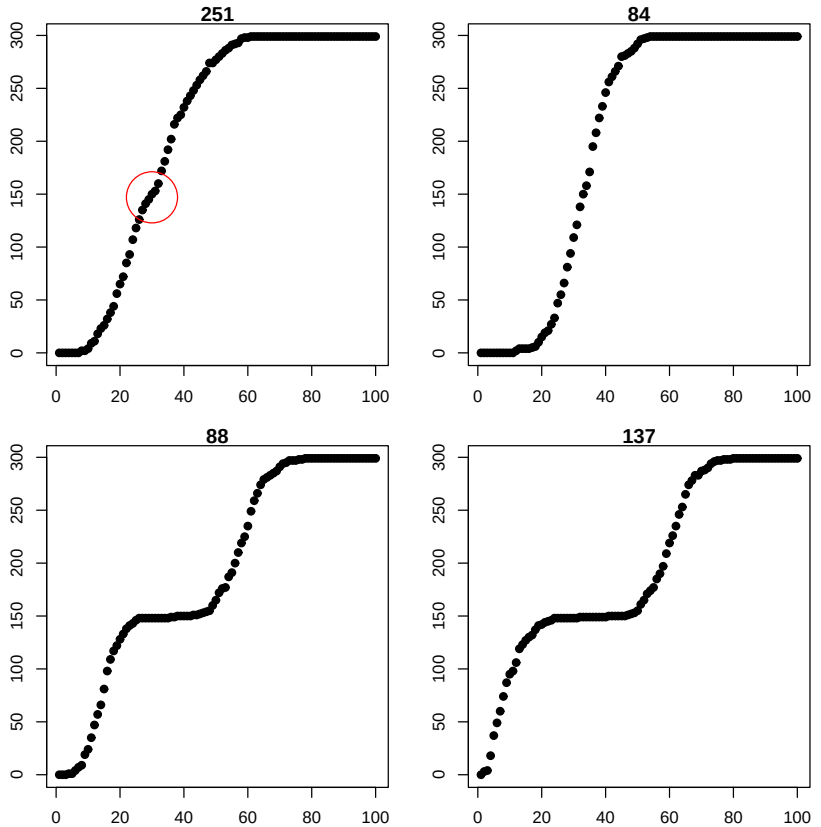
Disadvantages

- Eigendecomposition via SVD is $O(n^3)$
- Difficult to visualize more than 2 or 3 PC's at the same time
- We are no longer in the original descriptor space



R-NN Curves and Clusters





Smoothed *R*-NN Curves

R-NN curves are indicative of the number of clusters

R-NN Curves and Clusters

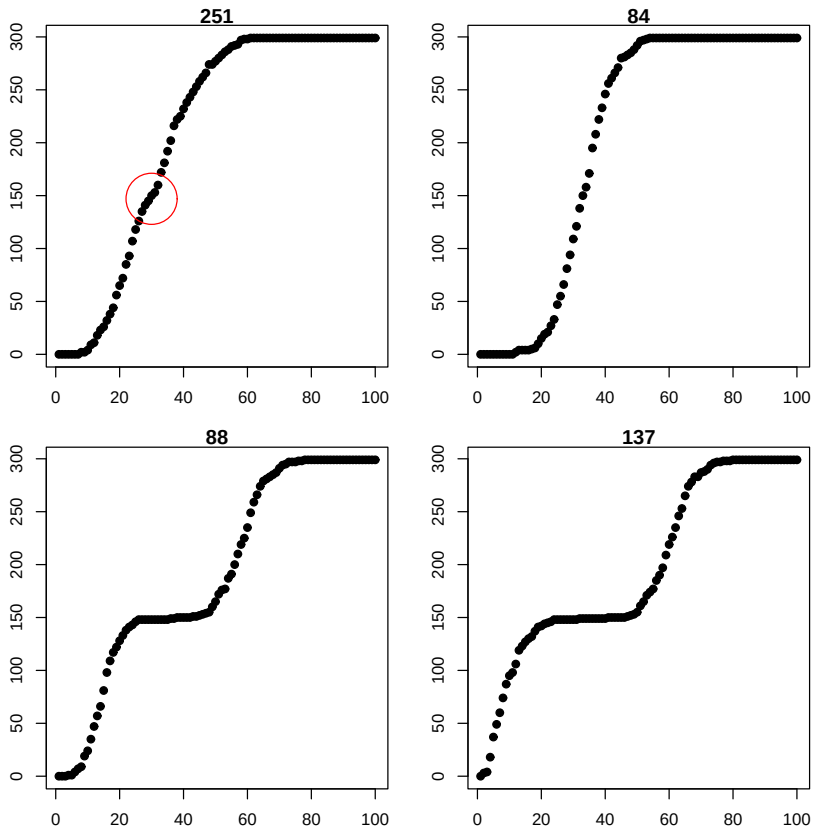
Counting the steps

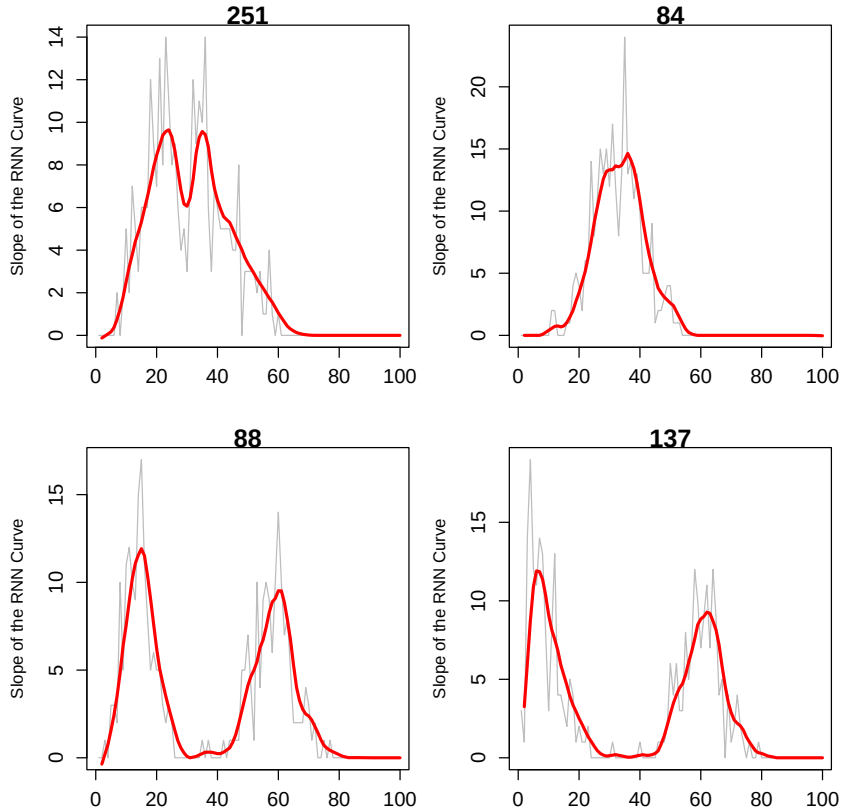
- Essentially a curve matching problem
- All points will not be indicative of the number of clusters
- Not applicable for radially distributed clusters

Approaches

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

R-NN Curves and Their Slopes

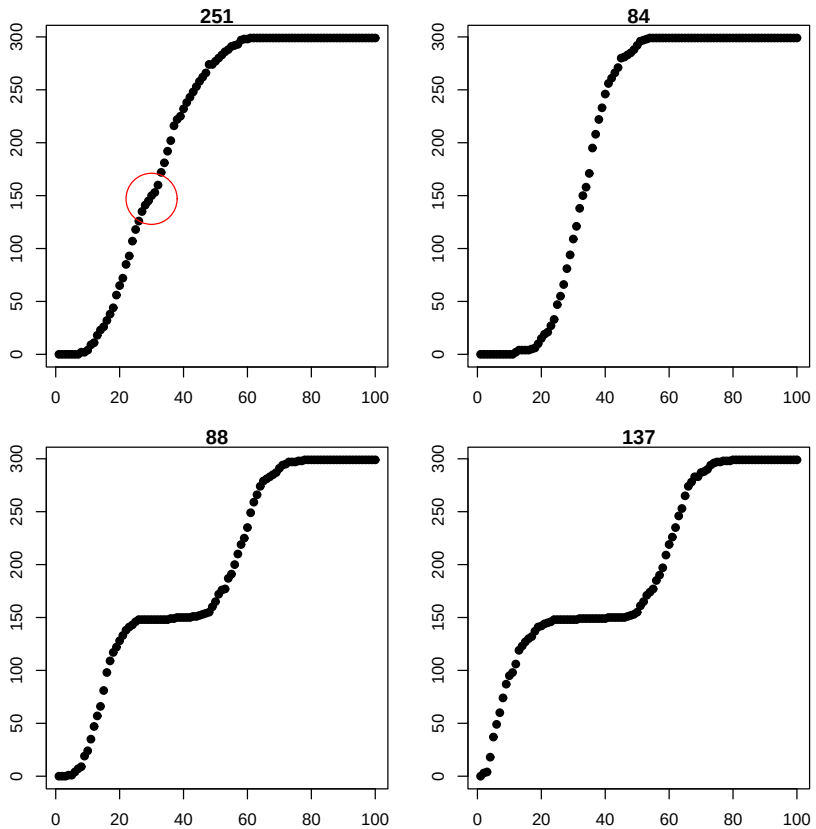


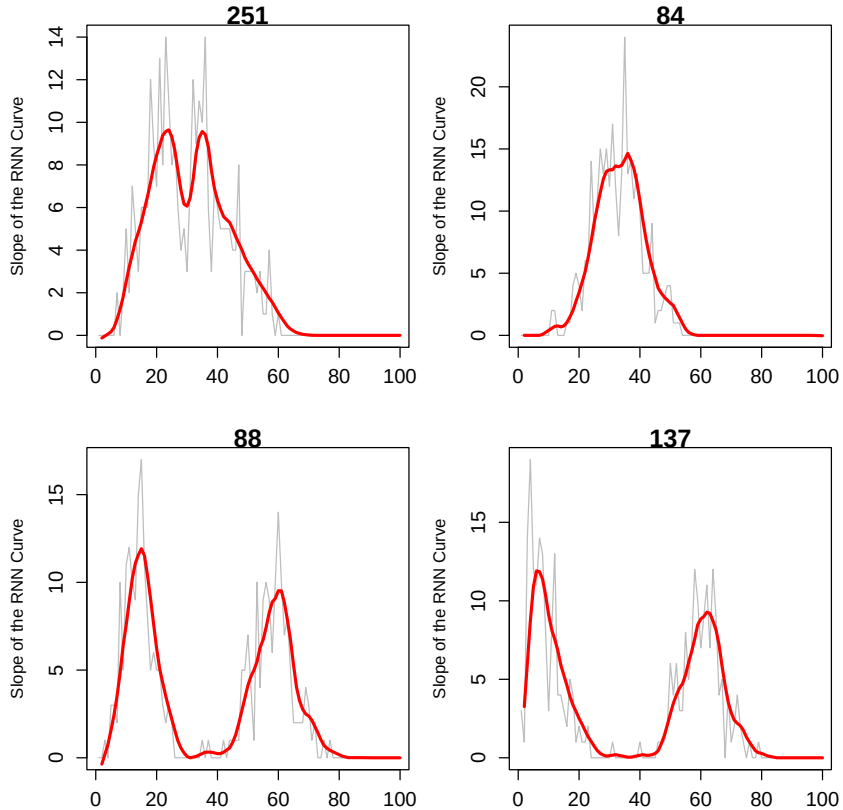


Smoothed first derivative of the *R*-NN Curves

Guha, R. *et al.*, *J. Chem. Inf. Model.*, **2007**, *47*, 1308–1318

R-NN Curves and Their Slopes





Smoothed first derivative of the *R*-NN Curves

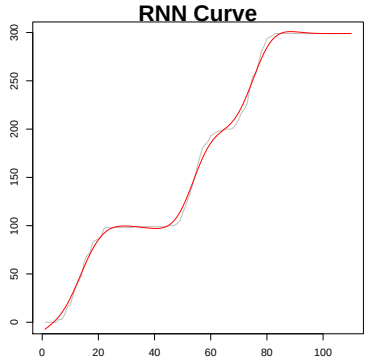
- Identifying peaks identifies the number of clusters
- Automated picking can identify spurious peaks

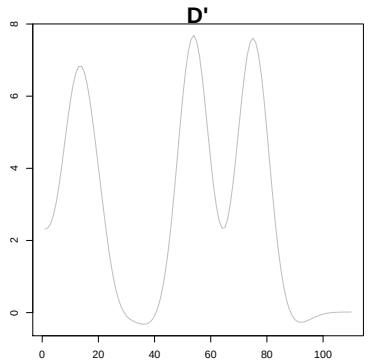
Guha, R. *et al.*, *J. Chem. Inf. Model.*, **2007**, *47*, 1308–1318

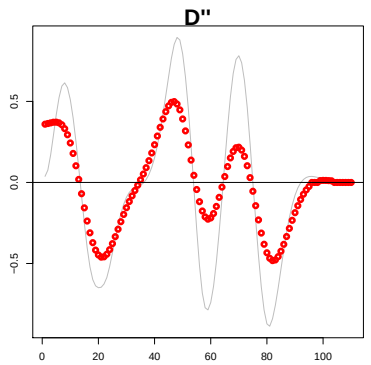
Slope Analysis of *R*-NN Curves

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
$N_{cluster} = [\max(N_{root}) + 1]/2$

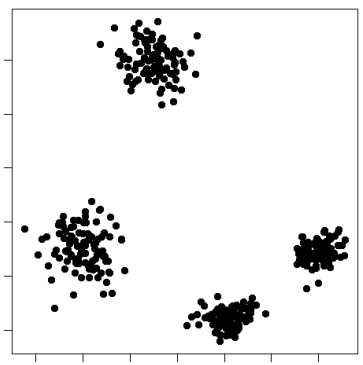




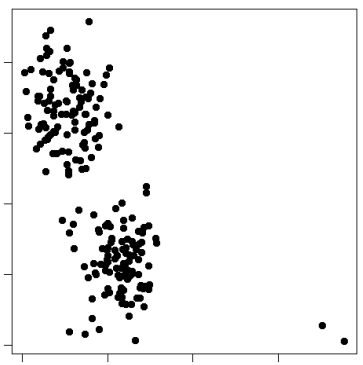


Simulated Data

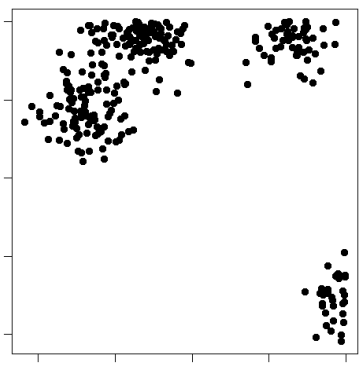
- Simulated 2D data using a Thomas point process
- Predicted k, followed by kmeans clustering using k
- Investigated similar values of k



k	ASW
4	0.61
3	0.74
5	0.70



k	ASW
3	0.44
2	0.65
4	0.47



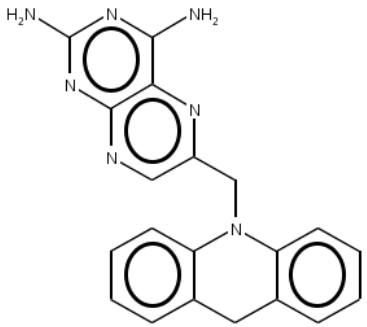
k	ASW
4	0.64
3	0.48
5	0.56

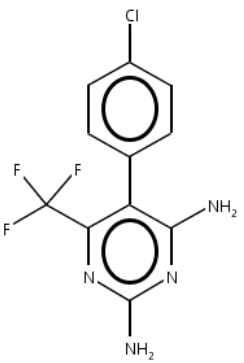
ASW - average silhouette width, higher is better; k - number of clusters

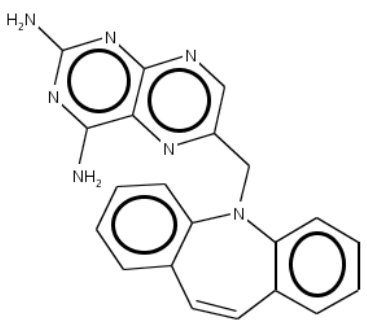
A Mixed Dataset

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





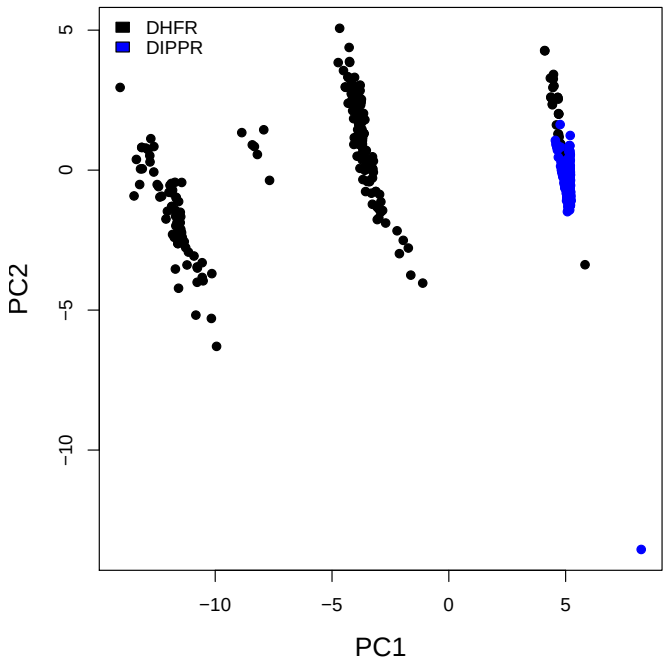


Sutherland, J. et al.; *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1906–1915

Goll, E. and Jurs, P.; *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 974–983

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

Drawbacks

- Currently works well for well separated clusters
  - But radial clusters will not be detected
- We now consider all the points in the dataset
  - Inefficient
  - Sampling experiments seem to indicate that we can make do with 45% of the points
  - Could prioritize by considering points wth low $R_{max(S)}$ values
- Small changes in local density can mislead the algorithm
  - But this is somewhat subjective

Domain Applicability for QSAR Models

What is it?

- How similar should a new structure be to the training set to get a reliable prediction

Approaches

- Descriptor based - distance from the centroid, leverage
- Structure based - fingerprint similarity
- Cluster based

Stanforth, R.W. et al., *QSAR. Comb. Sci.*, **2007**, *26*, 837–844

Netzeva, T.I. et al., *Altern. Lab. Anim.*, **2005**, *33*, 155–173

Sheridan, R.P. et al., *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1912–1928

Domain Applicability for QSAR Models

One Domain or Multiple Domains?

- It's possible to encompass the entire training set into a single domain
- This is a very broad approach
- Modeling approaches based on neighborhoods essentially consider multiple, local, domains
- Even for a global model, some observations are better predicted than others
    - Suggests that certain regions of a descriptor space are predicted better than others

Guha, R. et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1836–1847

Domain Applicability for QSAR Models

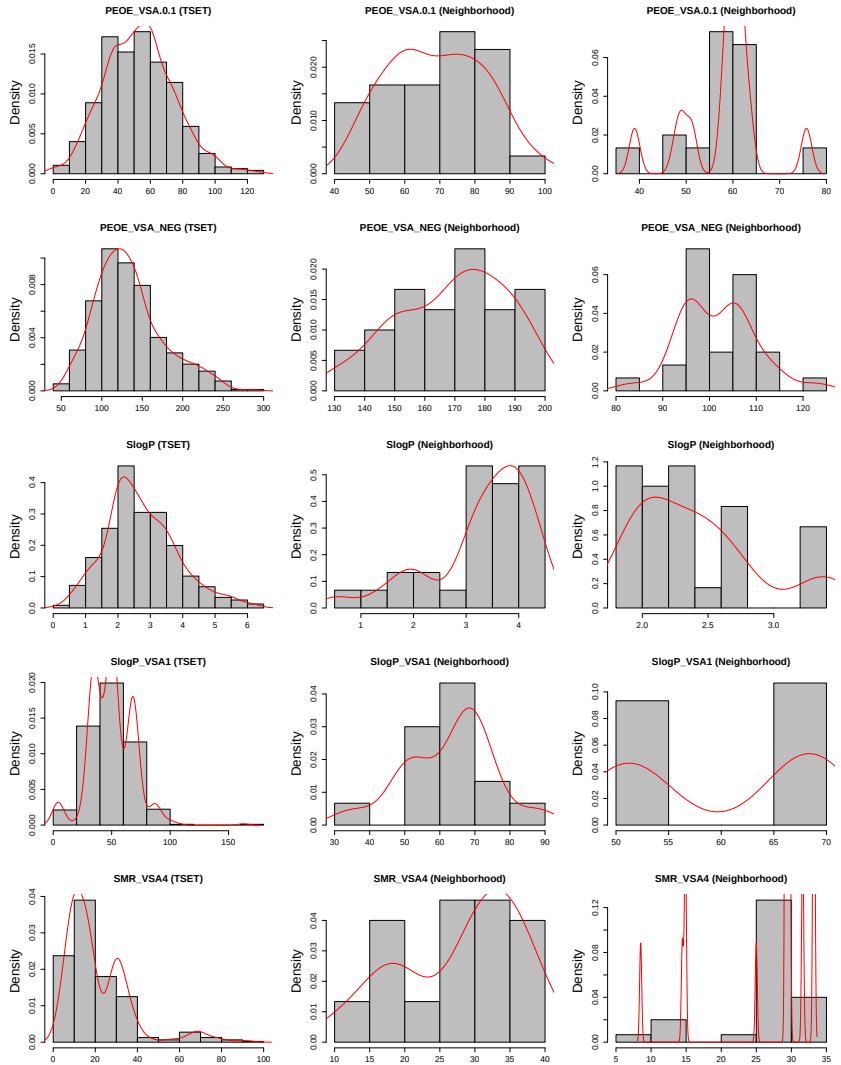
One Domain or Multiple Domains?

- It's possible to encompass the entire training set into a single domain
- This is a very broad approach
- Modeling approaches based on neighborhoods essentially consider multiple, local, domains
- Even for a global model, some observations are better predicted than others
    - Suggests that certain regions of a descriptor space are predicted better than others

Guha, R. et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1836–1847

Local Descriptor Distributions

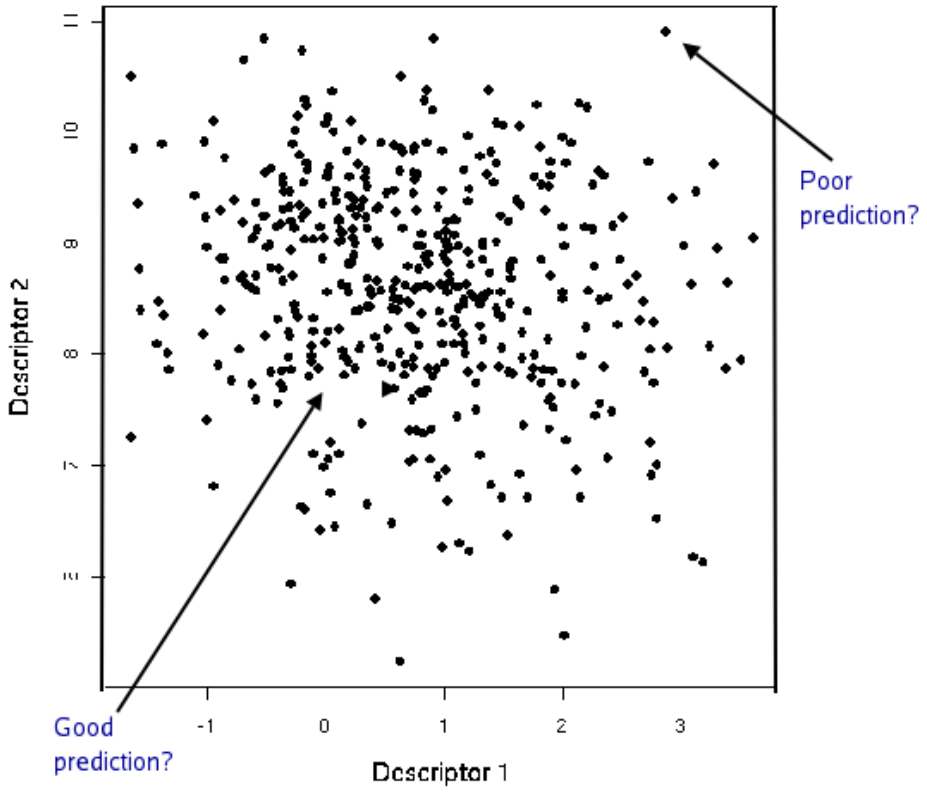
- For local regression models, the descriptor distribution in the neighborhood will affect predictive ability
- The effect is more severe when methods such as *k*NN are used which do not do any interpolation
- Nearest neighbors may not actually be nearby



Guha, R. et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1836–1847

Mapping Domain Applicability via Density

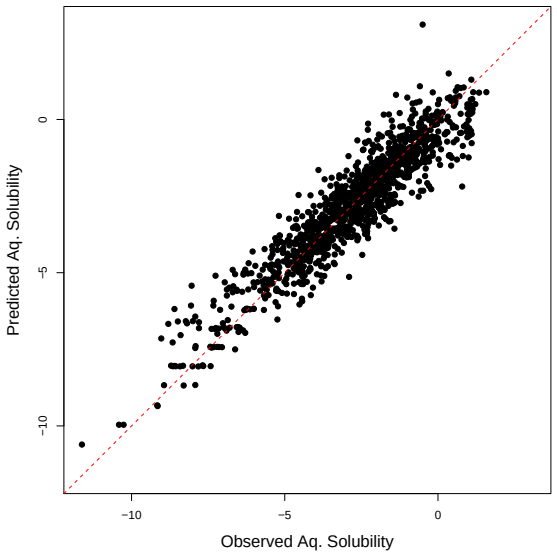
- Query molecules that occur in dense regions of of the training set descriptor space are expected to be well predicted
- Doesn't penalize for being a little far from the centroid of the space



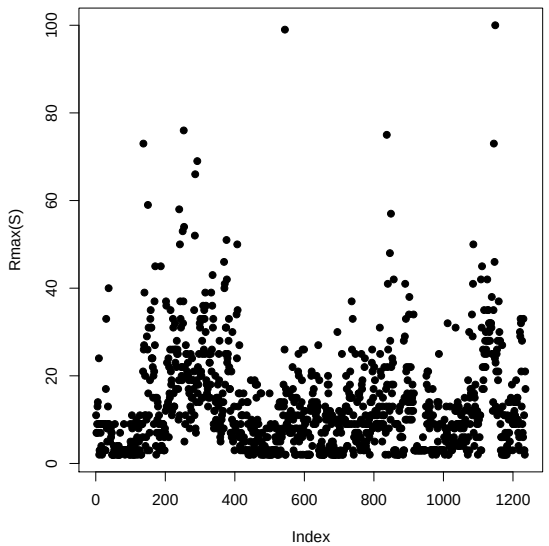
An Aqeuous Solubility Dataset

	Estimate	Std. Error	t value
(Intercept)	-2.1103	0.0909	-23.20
PEOE RPC+1	1.8754	0.2437	7.70
PEOE RPC-1	3.4954	0.1738	20.11
SlogP	-0.8199	0.0131	-62.55
CASA	0.0008	0.0001	-9.83

$RMSE = 0.7542$ $F(4, 1231) = 1993$

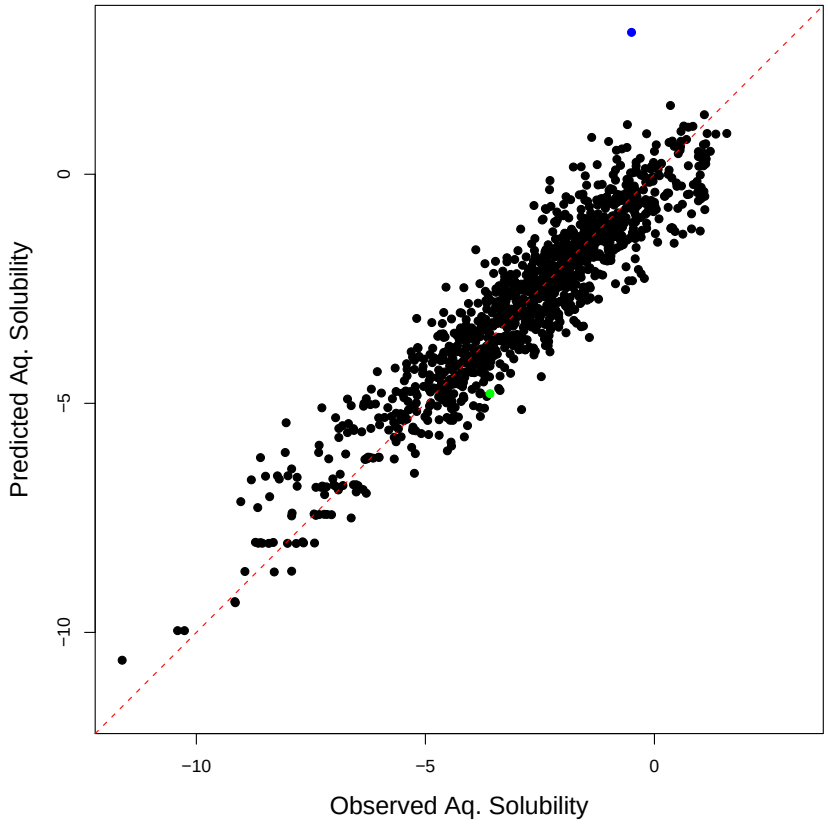


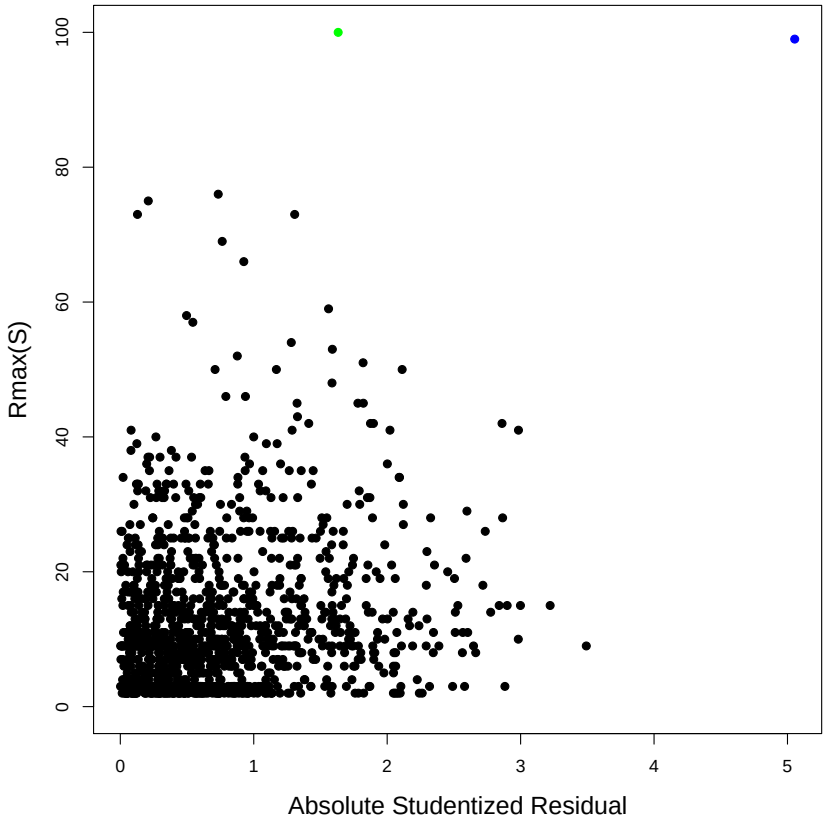
- 1236 compounds
- Evaluated 147 descriptors, reduced to 61
- Searched for a good subset using a GA



Huuskonen, J., *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 773–777

An Aqueous Solubility Dataset





- Isolated compounds can be predicted well
- The $R_{max(S)}$ values could be calculated more rigorously from a smoothed curve

Mapping Domain Applicability via Density

Caveats

- The approach assumes that similar compounds will have similar activities
- This is not always true (activity cliffs)
- Correlation between density as measured by $R_{max(S)}$ and prediction accuracy needs more investigation
  - Take into account descriptor importance via weighted Euclidean distances

Maggiora, G.M., *J. Chem. Inf. Model.*, **2006**, *46*, 1535–1535

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