Navigating Molecular Haystacks: Tools & Applications

Finding Interesting Molecules & Doing it Fast

Rajarshi Guha

Department of Chemistry
Pennsylvania State University

20st April, 2006

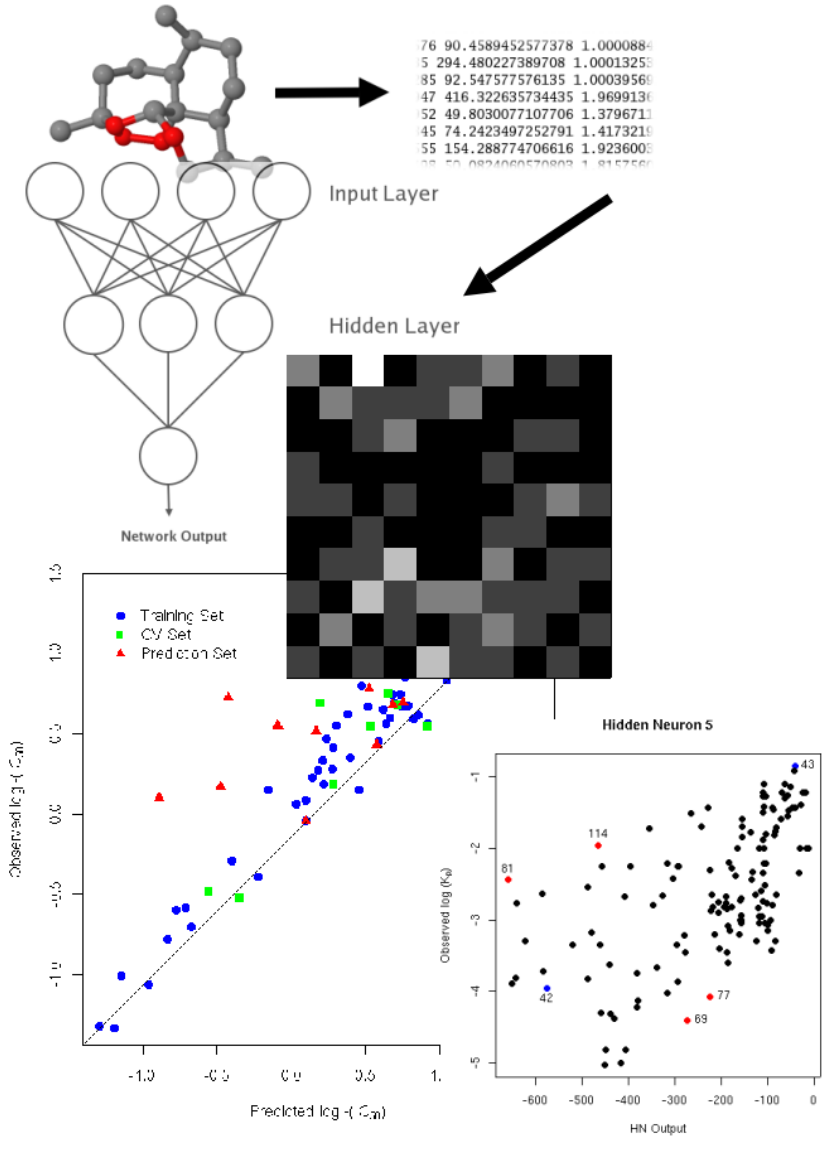
Past Work

QSAR Applications

- Artemisinin analogs
- PDGFR inhibitors
- Bleaching agents
- Linear & non-linear methods

QSAR Methods

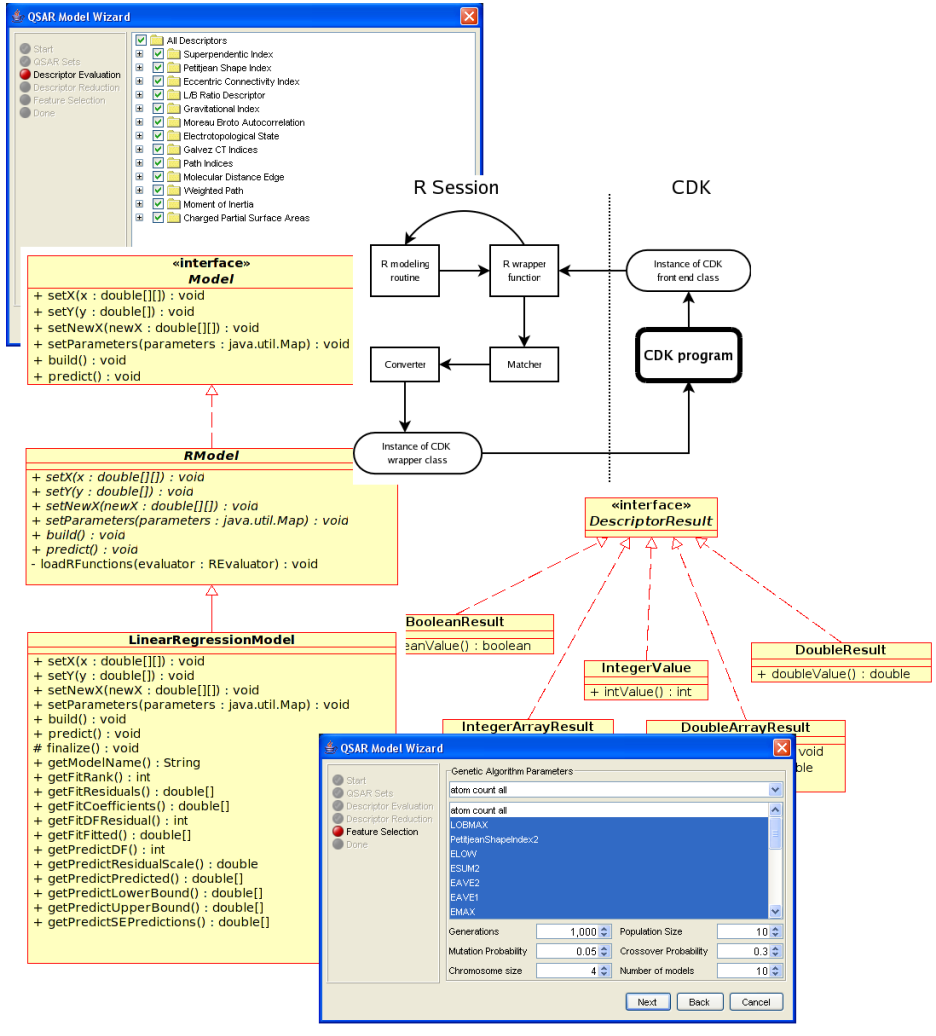
- Representative QSAR sets
- Model interpretation
- Model applicability



Past Work

Cheminformatics

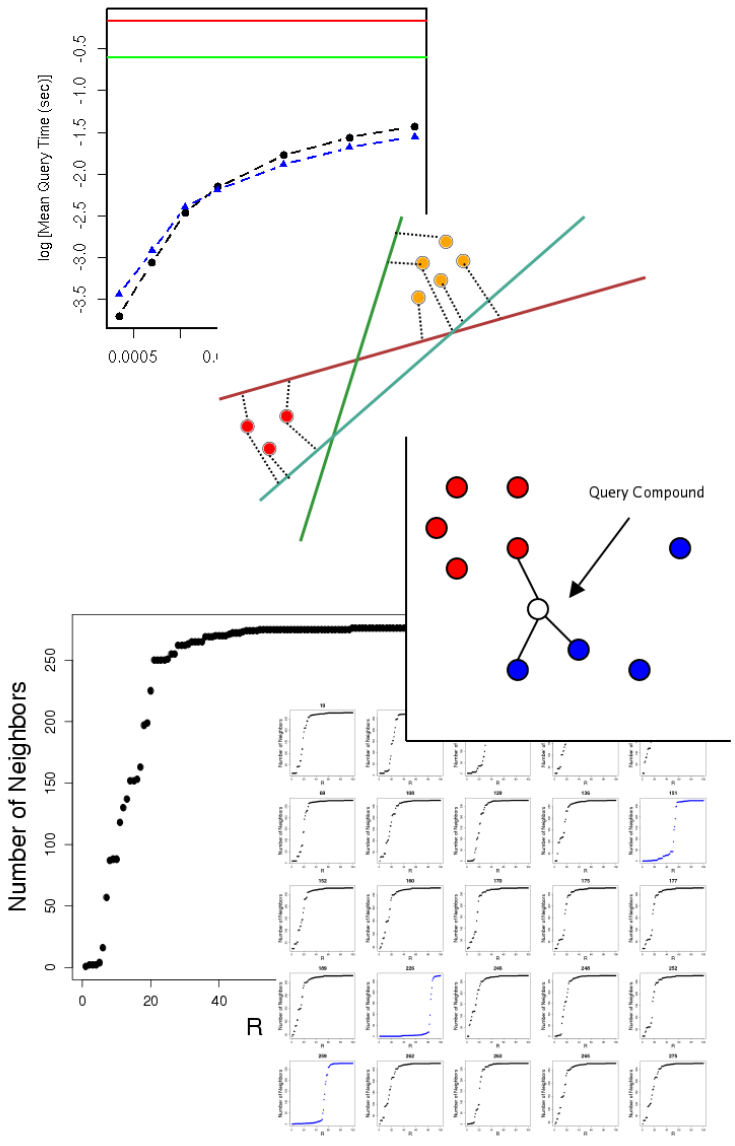
- Automated QSAR pipeline
- Contributions to the CDK
- Cheminformatics webservices
- R packages and snippets



Past Work

Chemical Data Mining

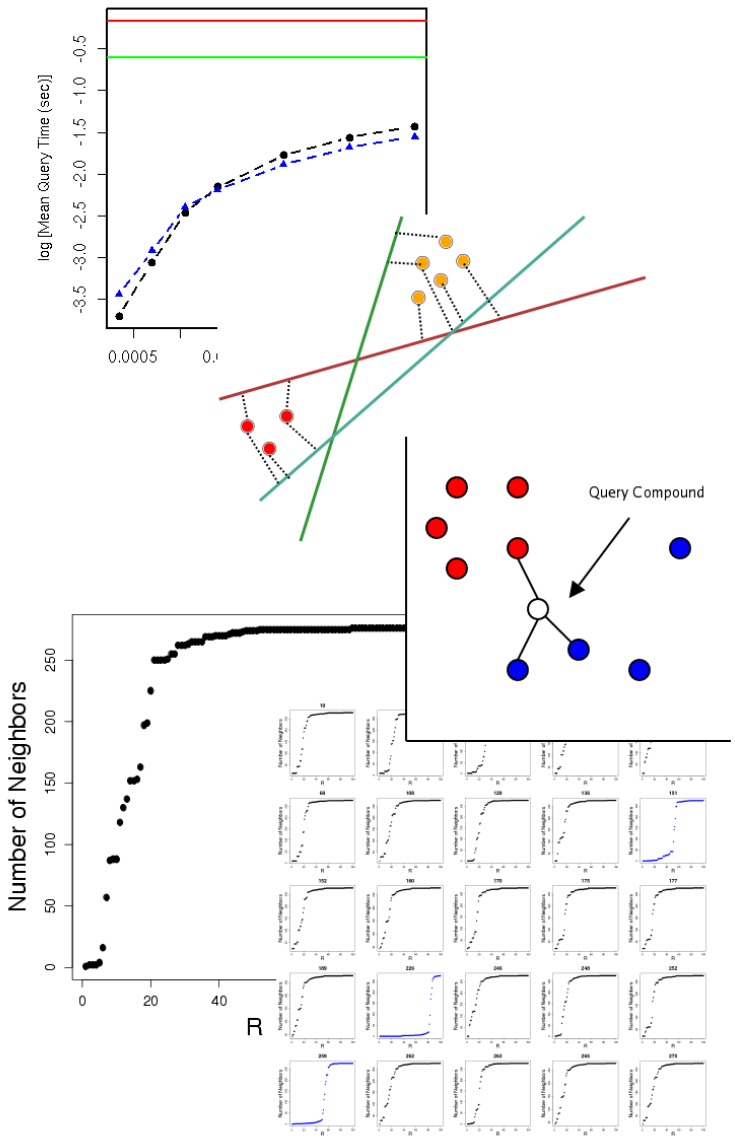
- Approximate k-NN
- Local regression
- Outlier detection
- VS protocol



Past Work

Chemical Data Mining

- Approximate k-NN
- Local regression
- Outlier detection
- VS protocol



Outline

Outline

Diversity Analysis

Why is it Important?

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

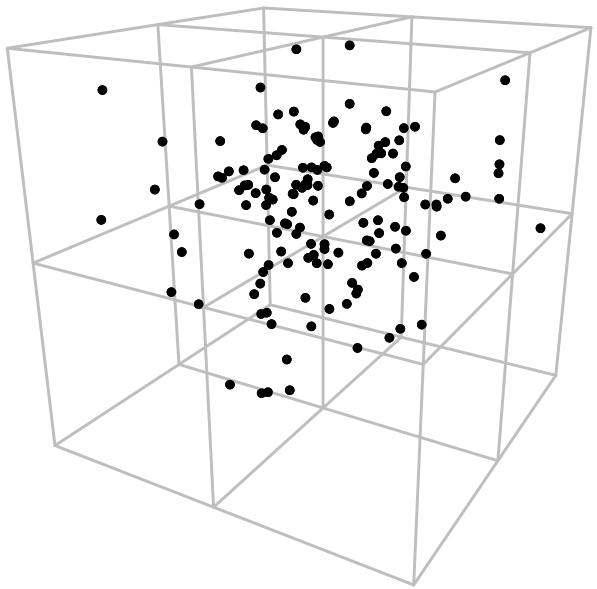
Approaches to Diversity Analysis

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



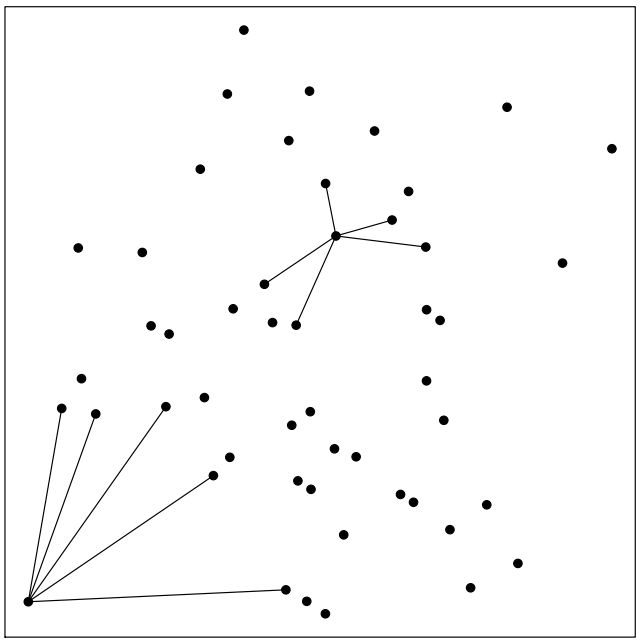
Schnur, D.; *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 36–45

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.; 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

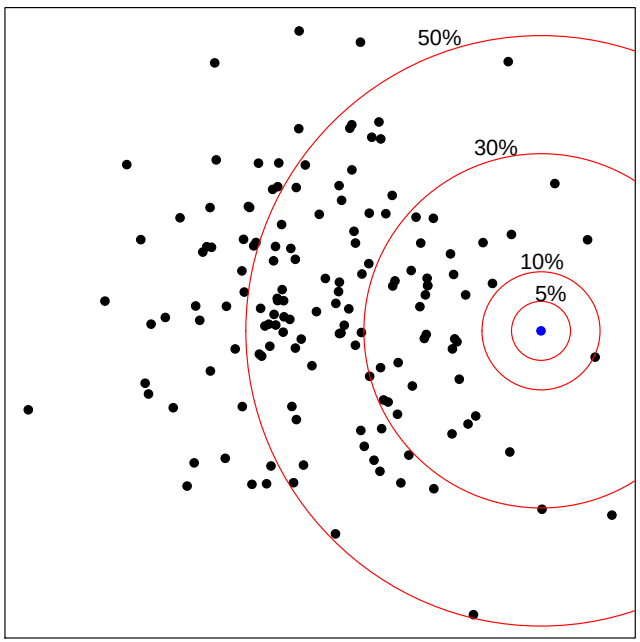
The question is ...

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

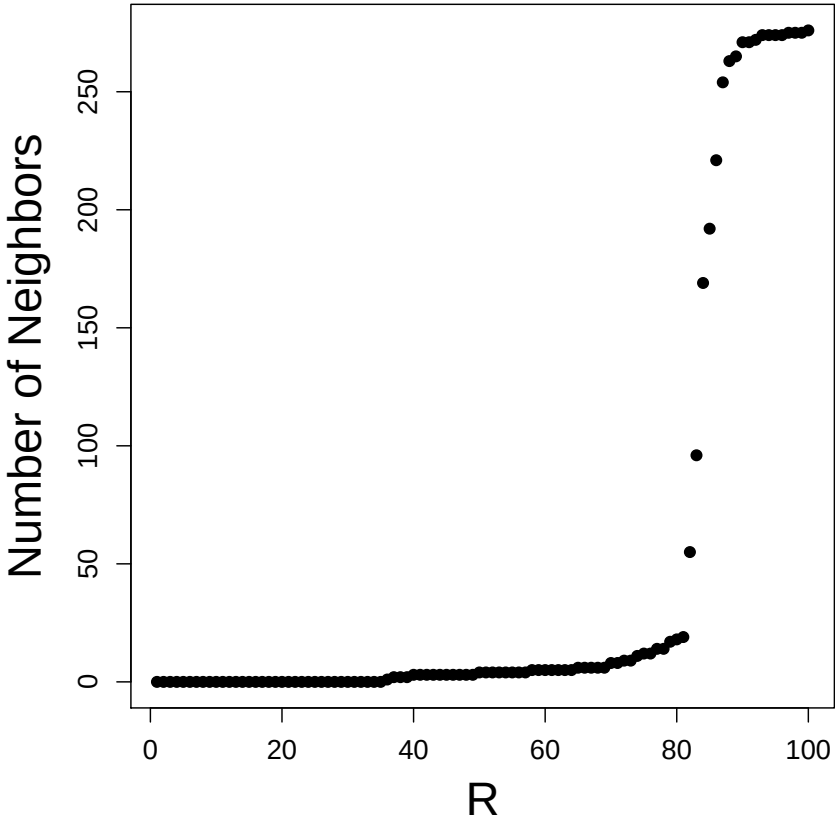
Algorithm

$D_{max} \leftarrow$ 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

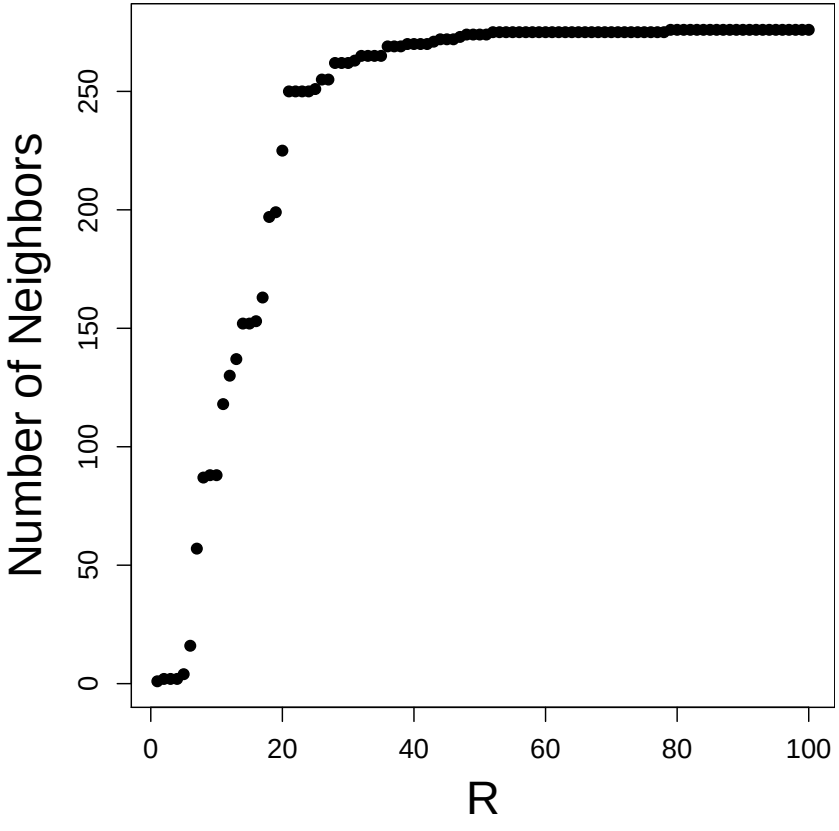


Guha, R.; Dutta, D.; Jurs, P.C.; Chen, T.; *J. Chem. Inf. Model.*, submitted

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

$$N_N = a \cdot \frac{1 + m\ e^{-R/\tau}}{1 + n\ e^{-R/\tau}}$$

- m, n should characterize the curve
- Problems
  - Two parameters
  - Non-linear fitting is dependent on the starting point
  - For some starting points, the fit 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

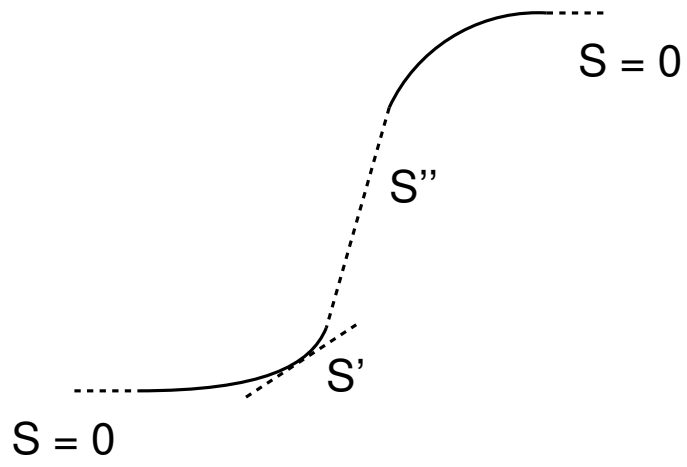
Characterizing an *R*-NN Curve

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



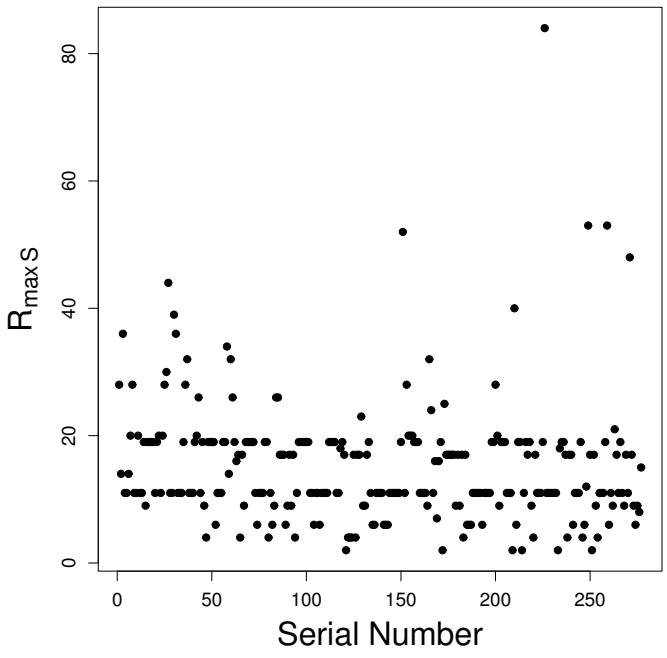
Characterizing Multiple *R*-NN Curves

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



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

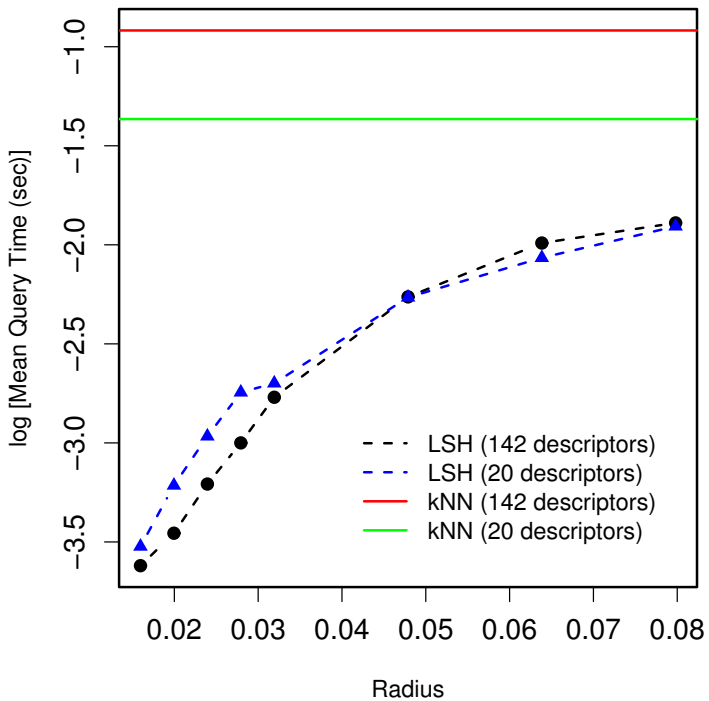
Datar, M. et al.; *SCG '04: Proc. 20th Symp. Comp. Geom.*; ACM Press, 2004

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

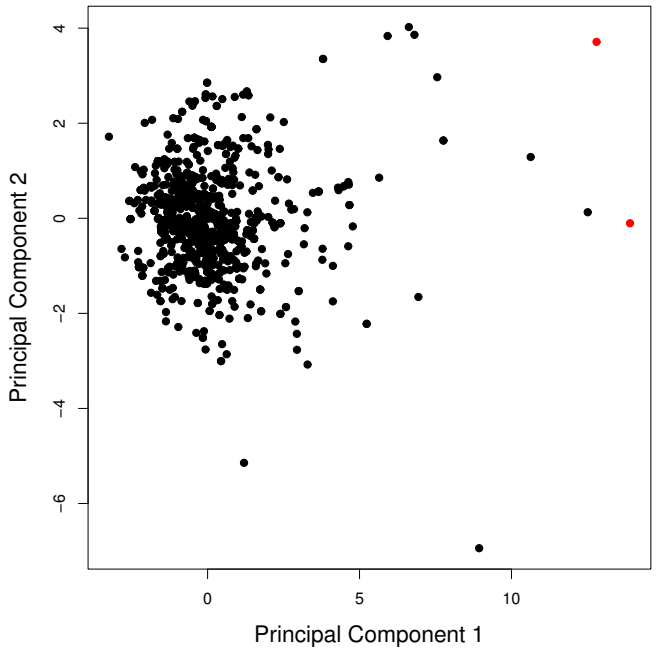
Alternatives?

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



Datasets

Boiling point dataset

- 277 molecules
- Average: $MW = 115$, $S_{tanimoto} = 0.20$
- Calculated 214 descriptors, reduced to 64

Kazius-AMES dataset

- 4337 molecules
- Average $MW = 240$, $S_{tanimoto} = 0.21$
- Known to have a number of significant outliers
- Calculated 142 MOE descriptors, reduced to 45

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

Kazius, J.; McGuire, R.; Bursi, R.; *J. Med. Chem.* **2005,** *48,* 312–320

Choosing a Descriptor Space

Boiling point dataset

- We had previously generated linear regression models using a GA to search for subsets
- Best model had 4 descriptors

Kazius-AMES dataset

- No models were generated for this dataset
- Considered random descriptor subsets
- Used a 5-descriptor subset to represent the results

The *R*-NN curve approach focuses on the distribution of molecules in a given descriptor space. Hence a *good* or *random* descriptor subset should be able to highlight the utility of the method

Choosing a Descriptor Space

Boiling point dataset

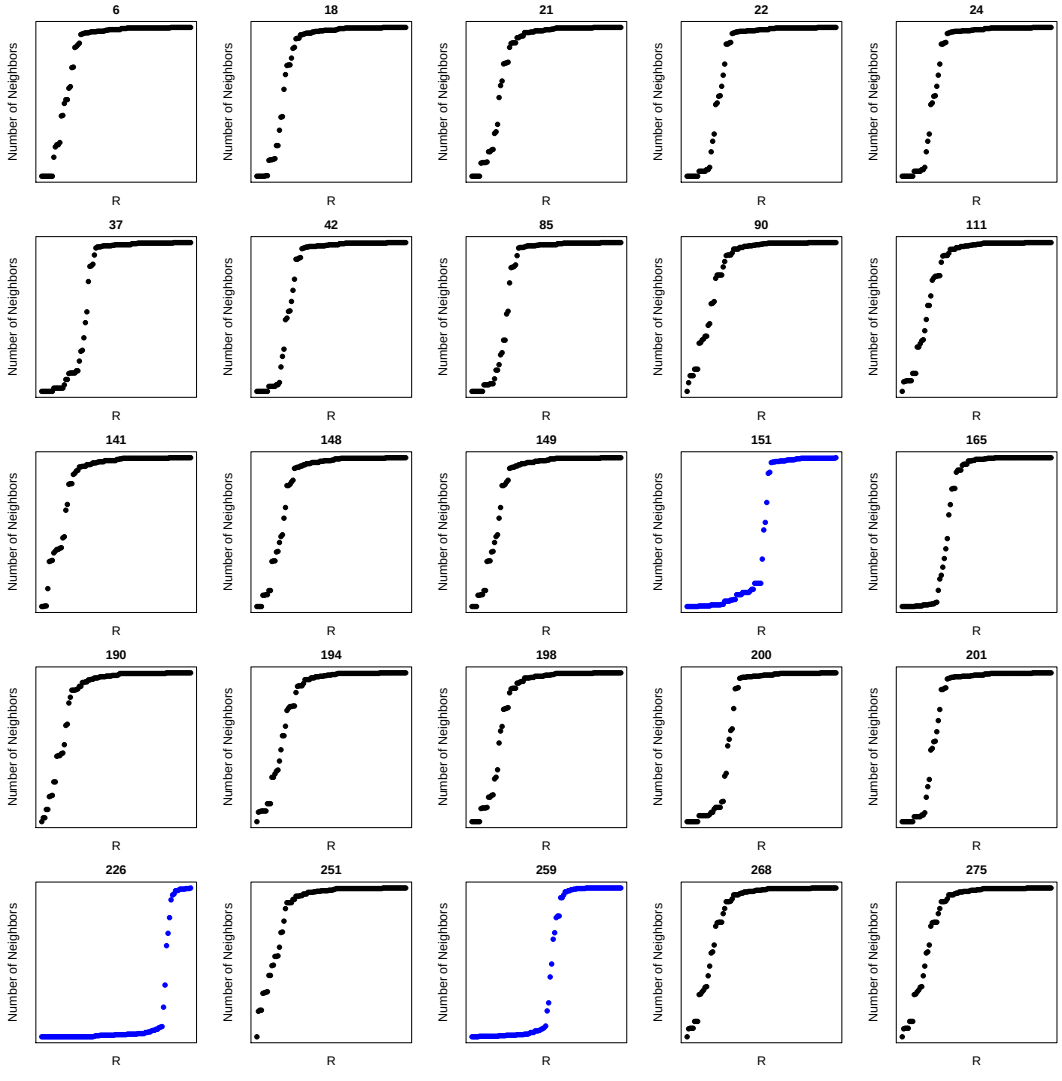
- We had previously generated linear regression models using a GA to search for subsets
- Best model had 4 descriptors

Kazius-AMES dataset

- No models were generated for this dataset
- Considered random descriptor subsets
- Used a 5-descriptor subset to represent the results

The *R*-NN curve approach focuses on the distribution of molecules in a given descriptor space. Hence a *good* or *random* descriptor subset should be able to highlight the utility of the method

R-NN Curves for the BP Dataset



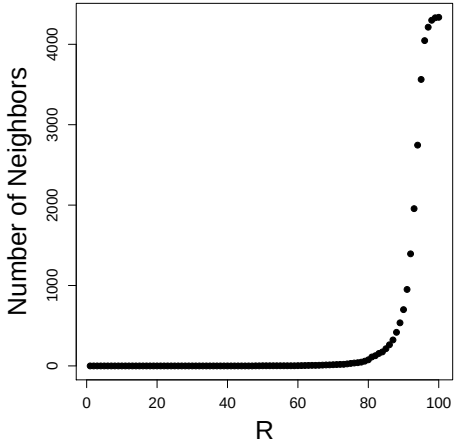
Outliers for the Kazius-AMES Dataset

Defining outliers

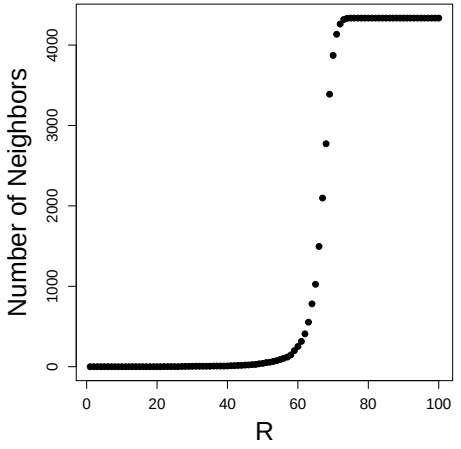
- $R = 0.5 \times D_{max}$
- NN count $\leq$ 10% of the dataset

Outliers detected

- Molecule **3904** had 2 neighbors
- Molecule **1861** had 41 neighbors
- $R = 0.4 \times D_{max} \longrightarrow$ 6 outliers
- Clearly, this process is subjective

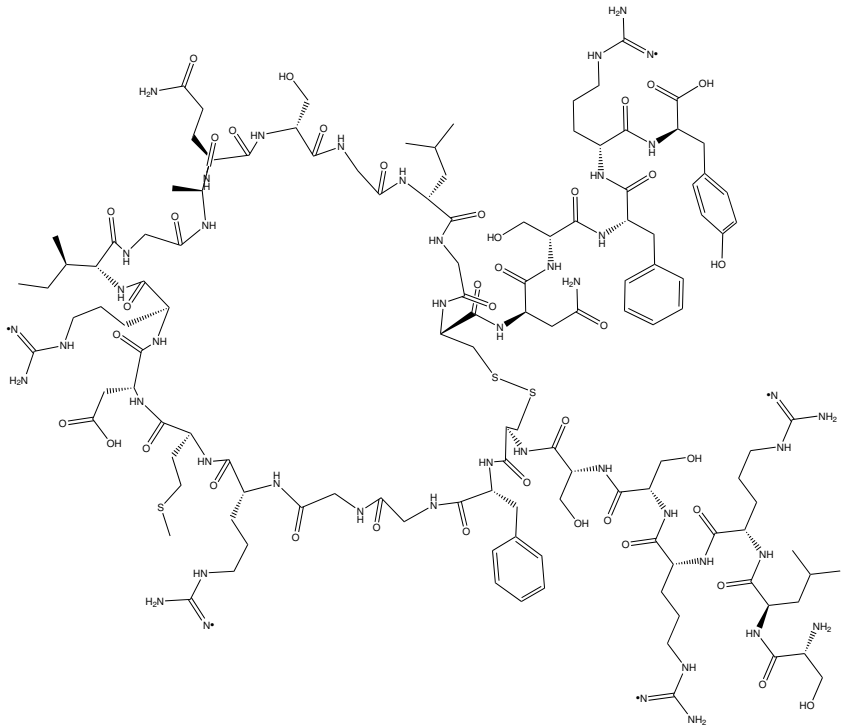


3904

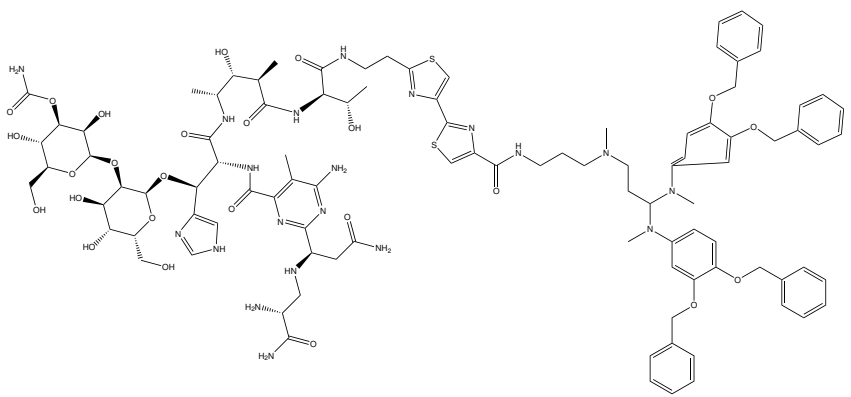


1861

Outlier Structures



3904

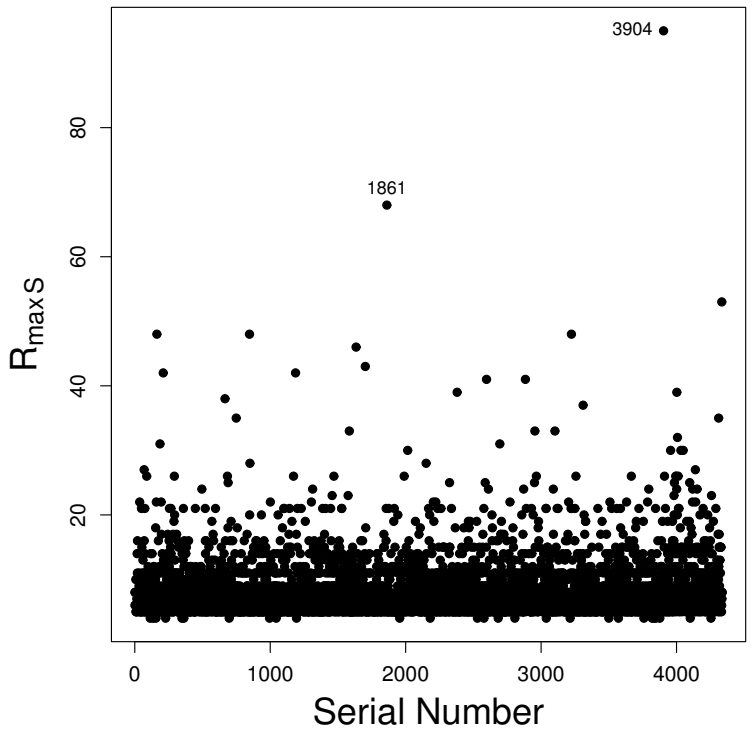


1861

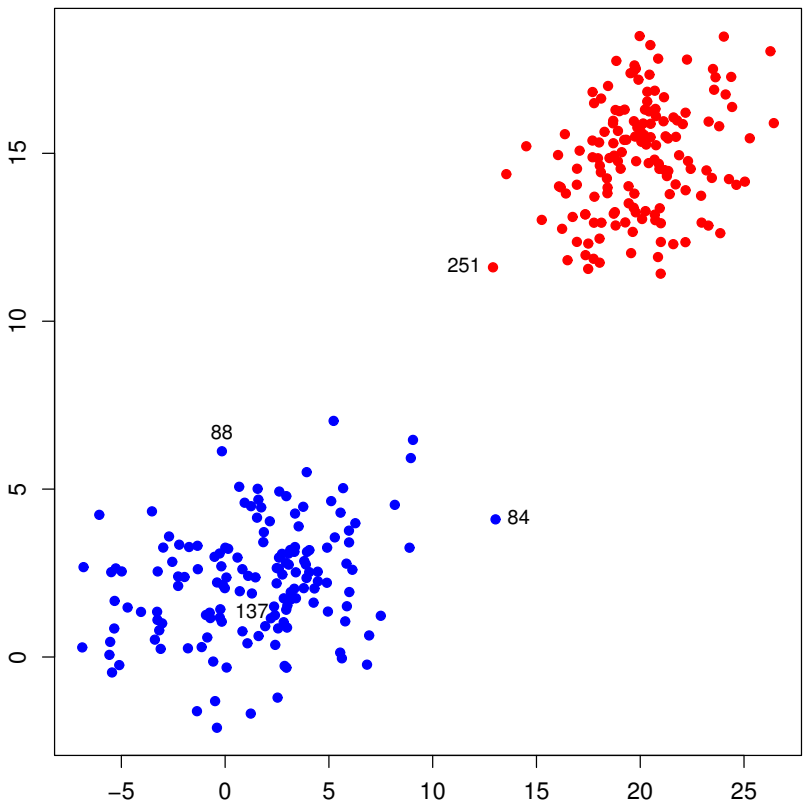
Summary of the Kazius-AMES Dataset

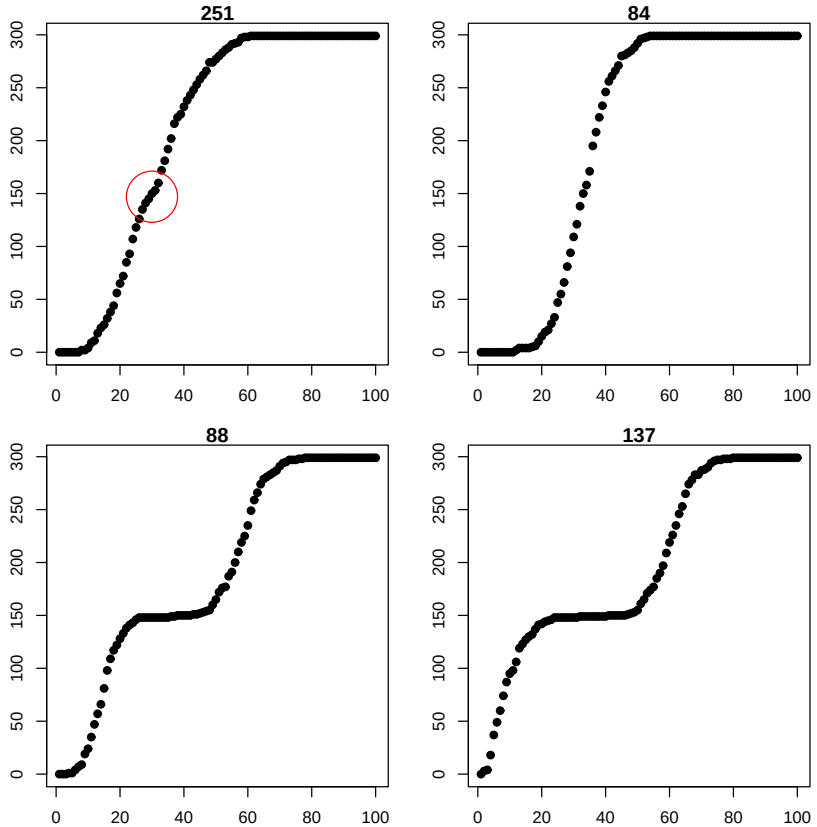
- **3904** & **1861** are immediately identifiable as outliers
- Excluding these two, there are no *significant* outliers
- The dataset appears to be relatively dense

One plot summarizes the location characteristics of *all* the molecules in the dataset

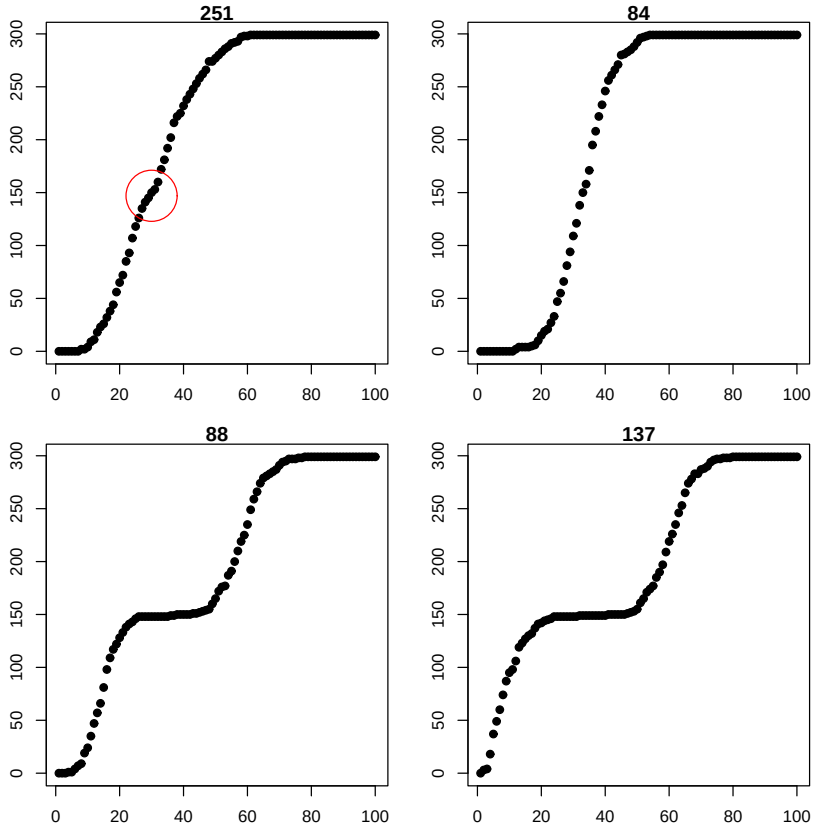


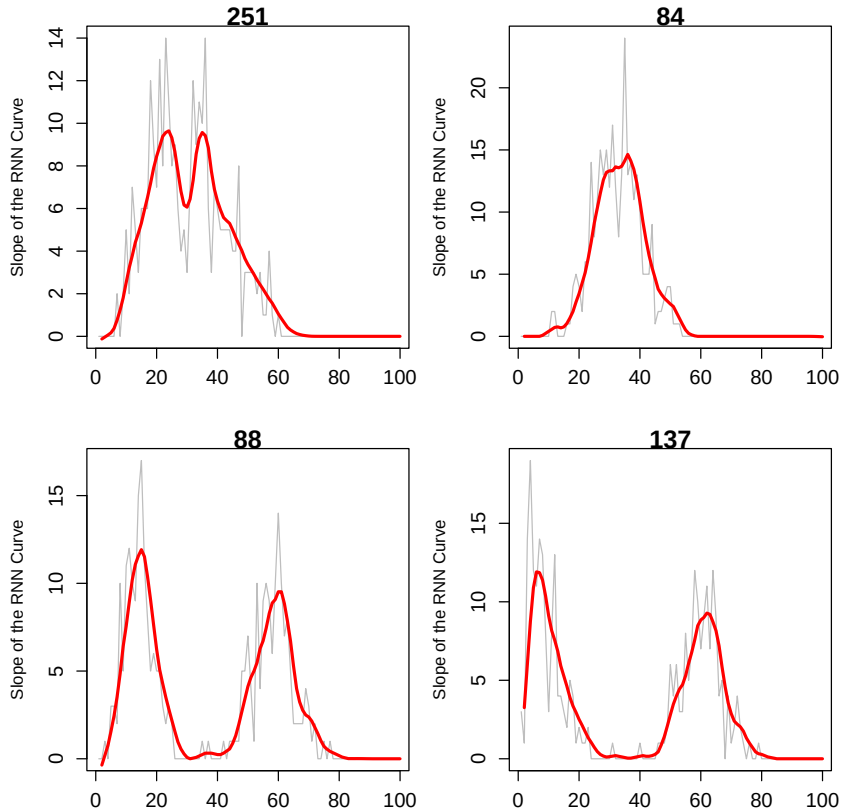
R-NN Curves & Clustering





R-NN Curves & Clustering





Smoothed with Friedmans *Super Smoother*

Friedman, J.H.; Laboratory for Computational Statistics, Stanford University, *Technical Report 5*, **1984**

Conclusions

Future Work

- Allow comparison over multiple datasets
- Automated detection of clustering

Conclusions

- Applicable to arbitrary dimensions
- Can be easily extended to large datasets
- Summarizing a dataset does not require any user defined parameters
- Simple and intuitive approach to visualizing distribution of molecules in a descriptor space

Conclusions

Future Work

- Allow comparison over multiple datasets
- Automated detection of clustering

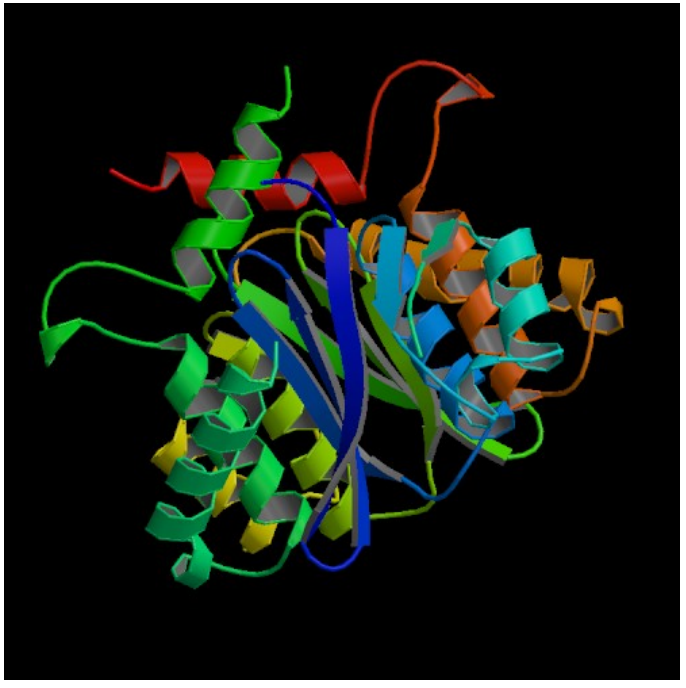
Conclusions

- Applicable to arbitrary dimensions
- Can be easily extended to large datasets
- Summarizing a dataset does not require any user defined parameters
- Simple and intuitive approach to visualizing distribution of molecules in a descriptor space

Outline

Background

- Most drugs for HIV treatment target the protease or reverse transcriptase enzymes
- Integrase is essential for the replication of HIV
- There are no FDA approved drugs that target HIV integrase



1BIS

Previous Work

Previous approaches

- Pharmacophore based models
- Docking

Disadvantages

- Requires a reliable receptor structure
- Computationally intensive
- Does not necessarily lead to a diverse hit list

Nair, V; Chi, G.; Neamati, N.; *J. Med. Chem.*, **2006**, *49*, 445–447

Dayam, R; Sanchez, T.; Neamati, N.; *J. Med. Chem.*, **2005**, *48*, 8009–8015

Deng, J. et al.; *J. Med. Chem.*, **2005**, *48*, 1496–1505

Goals

High speed

- Be able to process large libraries rapidly
- Avoid docking till required
- Try and use connectivity information only

Reliability

- Use a consensus approach for predictions
- Use molecular similarity to further prioritize predictions

Novelty

- Try to obtain hits suitable for lead hopping
- Try to obtain a diverse set of hits

Goals

High speed

- Be able to process large libraries rapidly
- Avoid docking till required
- Try and use connectivity information only

Reliability

- Use a consensus approach for predictions
- Use molecular similarity to further prioritize predictions

Novelty

- Try to obtain hits suitable for lead hopping
- Try to obtain a diverse set of hits

Goals

High speed

- Be able to process large libraries rapidly
- Avoid docking till required
- Try and use connectivity information only

Reliability

- Use a consensus approach for predictions
- Use molecular similarity to further prioritize predictions

Novelty

- Try to obtain hits suitable for lead hopping
- Try to obtain a diverse set of hits

Datasets & Tools

Training data

- Curated dataset
- 529 inactives
- 395 actives
- Binary valued activity

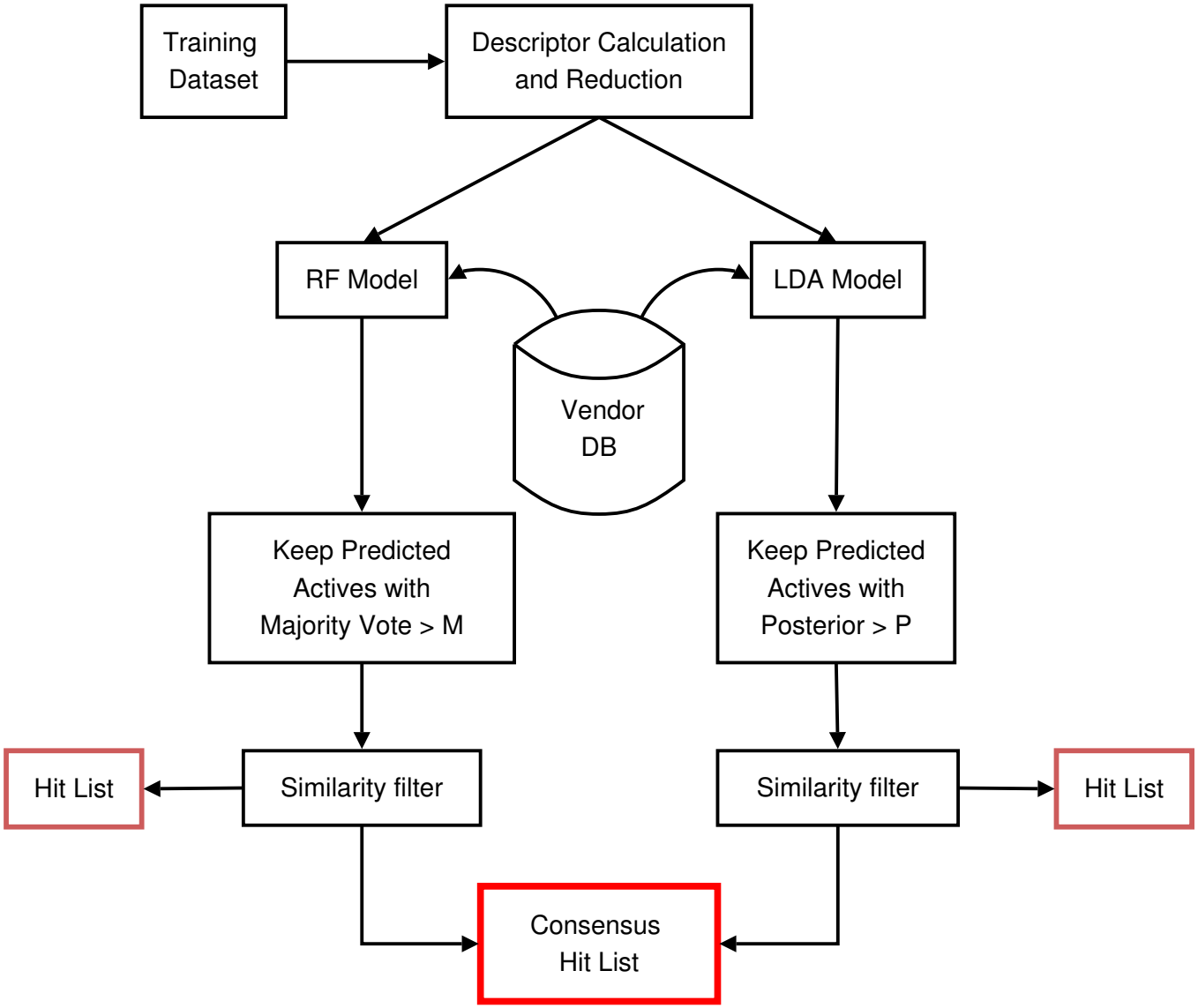
Vendor database

- 50,000 compounds
- 2D structures

Software

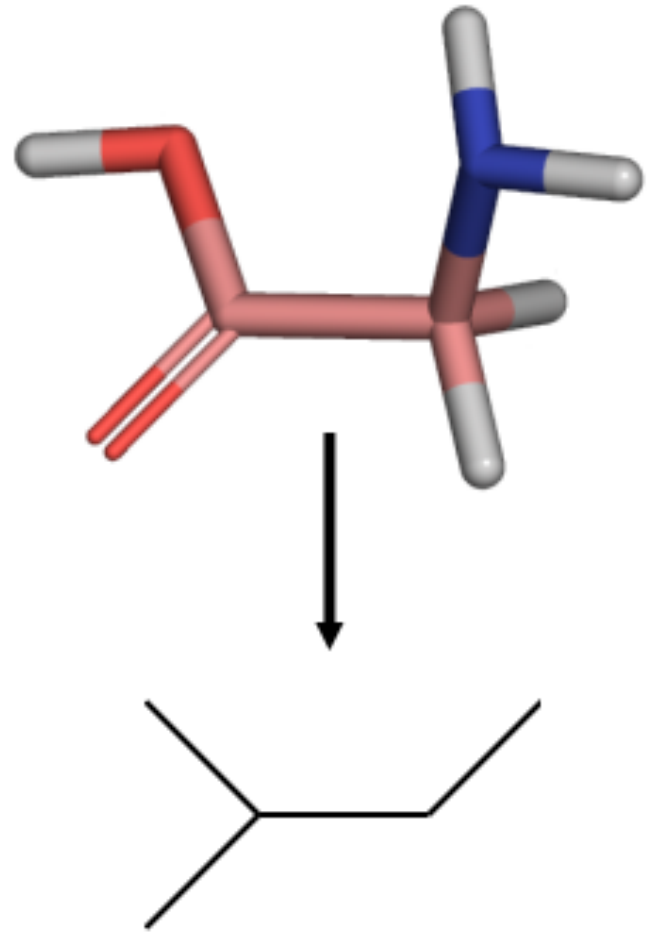
- MOE for descriptors
- R for the VS protocol
  - MASS
  - randomForest
  - fingerprint

Overview of the Screening Protocol



Descriptor Calculations

- Restricted ourselves to topological & *2.5D* descriptors
- 147 descriptors were evaluated
- Correlated and low variance descriptors were removed resulting in a reduced pool of 45 descriptors



Predictive Models - LDA

Why?

- Simple
- May be sufficient

How?

- Use a GA to search for good descriptor subsets
- Used a 6-descriptor model

Accuracy?

	Percent Correct
On whole dataset	72
On TSET/PSET	72 / 71
With leave-10%-out	71

Predictive Models - Random Forest

Why?

- No feature selection required
- Does not overfit
- Useful for non-linearities in the dataset

How?

- Number of trees = 500
- Number of features sampled = 6

Accuracy

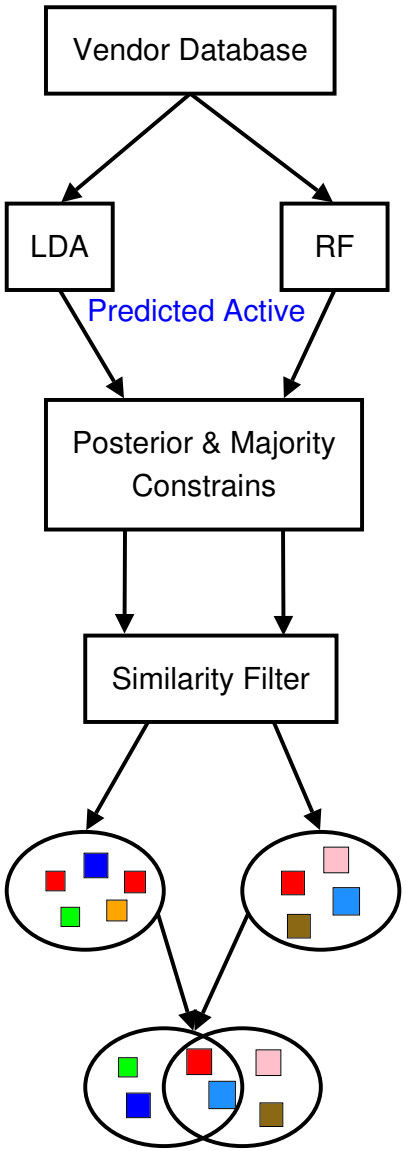
	Percent Correct
On whole dataset	72
On TSET/PSET	75 / 70

Consensus Predictions

Goal

Try and be sure that a predicted active is active

- Consider compounds predicted active by *both* models
  - LDA: consider compounds with posterior $>$ P
  - RF: consider compounds with majority vote $>$ M
- Apply similarity filter
- We now have 2 hit lists
- A final hit list is obtained from the intersection of the individual hitlists

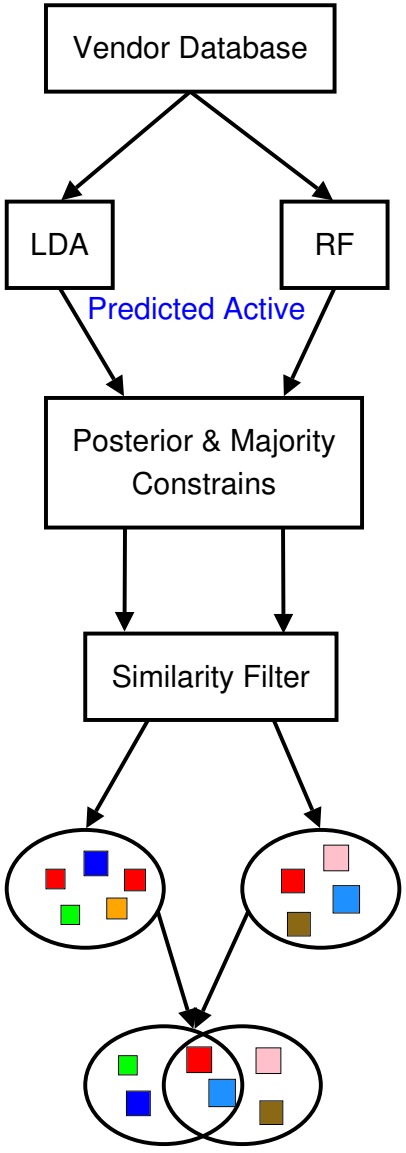


Consensus Predictions

Goal

Try and be sure that a predicted active is active

- Consider compounds predicted active by *both* models
  - LDA: consider compounds with posterior $>$ P
  - RF: consider compounds with majority vote $>$ M
- Apply similarity filter
- We now have 2 hit lists
- A final hit list is obtained from the intersection of the individual hitlists



Similarity Filter

Goal

Predict actives from vendor database that are more similar to actives than inactives

- Evaluate 166 bit MACCS fingerprints
- Evaluate average Tanimoto similarity between
  - Vendor compound and TSET actives (S_1)
  - Vendor compound and TSET inactives (S_2)
- Select compounds where

$$S_1 - S_2 \geq \epsilon$$

Similarity Filter

Goal

Predict actives from vendor database that are more similar to actives than inactives

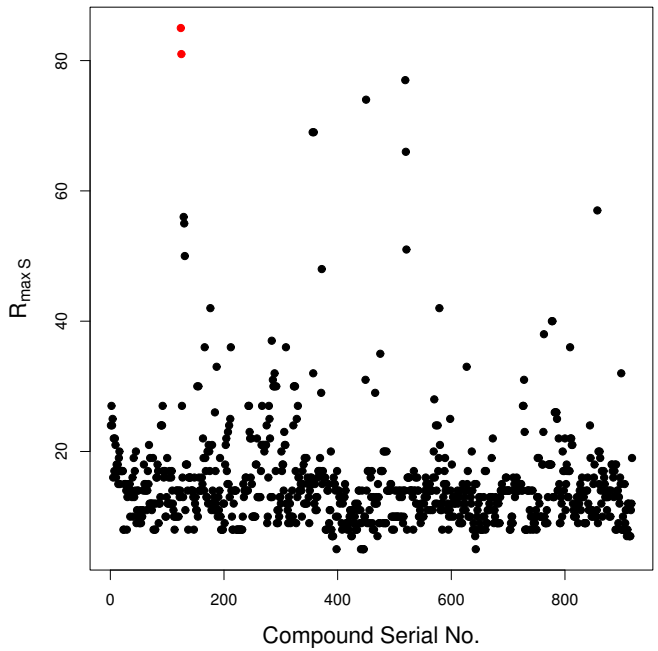
- Evaluate 166 bit MACCS fingerprints
- Evaluate average Tanimoto similarity between
  - Vendor compound and TSET actives (S_1)
  - Vendor compound and TSET inactives (S_2)
- Select compounds where

$$S_1 - S_2 \geq \epsilon$$

Suggesting Good Leads

Can we prioritize further?

- Look for TSET actives that are outliers
- These compounds may be good starting points for further modifications
- Evaluate similarity of hit list compounds to these isolated TSET compounds
- Hit list compounds most similar to the most isolated TSET active may be good leads



The points marked red are the most outlying compounds in the TSET and are also active

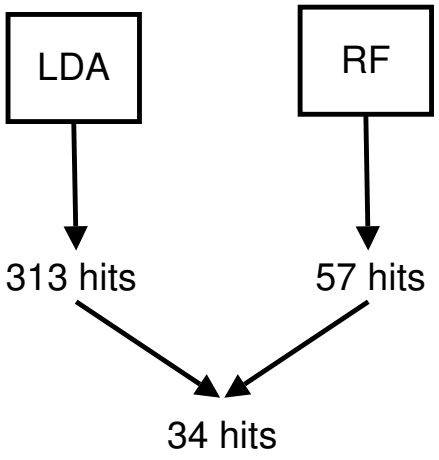
Saeh, J.C.; Lyne, P.D.; Takasaki, B.K.; Cosgrove, D.A.; *J. Chem. Inf. Comput. Sci.*, **2005**, *45*, 1122-1133

Guha, R.; Dutta, D.;Jurs, P.C.; Chen, T; *J. Chem. Inf. Model.*, *submitted*

Results

Summary of the hits

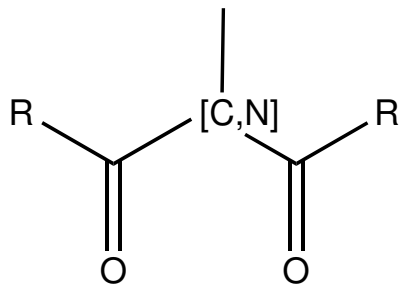
- Parameter settings
  - $P, M > 0.7$
  - $\epsilon \geq 0.01$
- Of the 34 hits, 7 had a similarity > 0.5 with the most outlying TSET active
- We obtained 66 more hits from an ensemble of LDA models



- Average similarity $= 0.64$
- Not significantly diverse
- None were in common with those predicted by the pharmacophore models

The Search for β-diketones

- A recently published inhibitor exhibited a β-diketone moiety
- None of our hits had this moiety

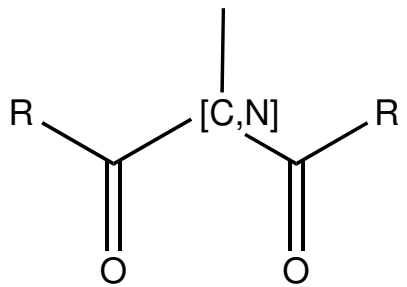


- Searched the vendor DB for structures with this feature and predicted them
  - LDA model $\rightarrow$ 244 actives (posterior > 0.8)
  - RF model $\rightarrow$ 4 actives (score > 0.85)
  - The intersection set contained 4 compounds
- We missed these hits due to our similarity constraints

Nair, V; Chi, G.; Neamati, N.; *J. Med. Chem.*, **2006**, *49*, 445–447

The Search for β-diketones

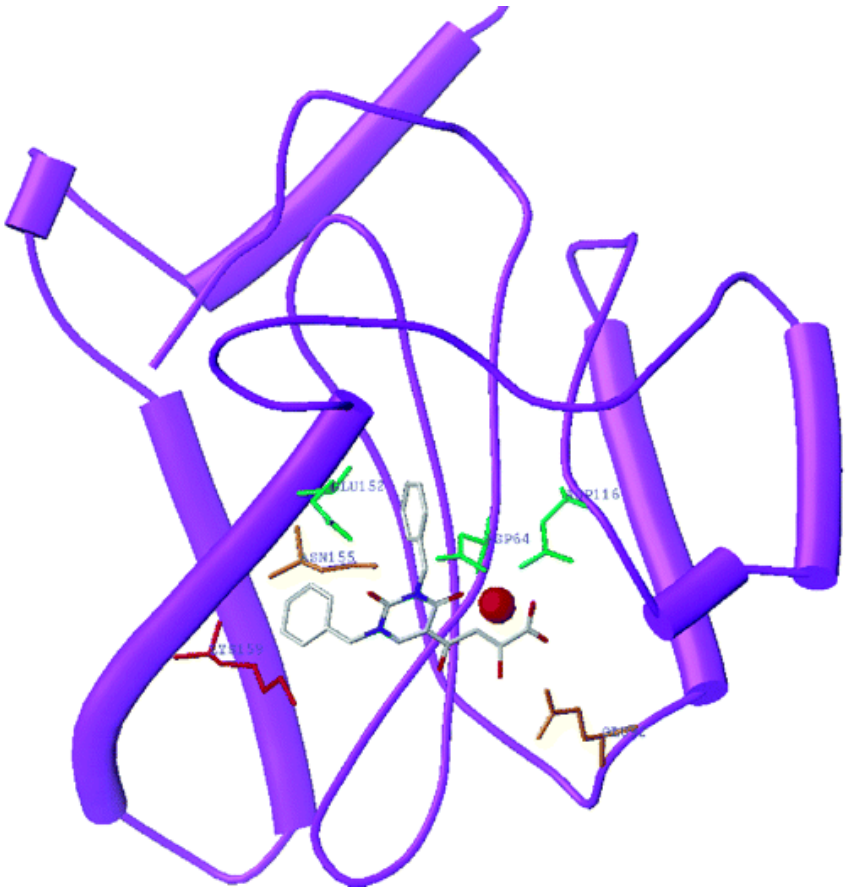
- A recently published inhibitor exhibited a β-diketone moiety
- None of our hits had this moiety

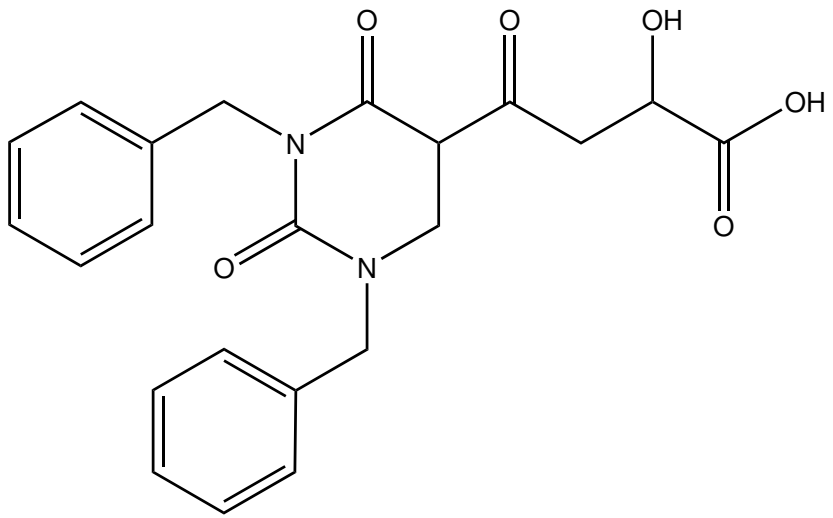


- Searched the vendor DB for structures with this feature and predicted them
  - LDA model $\rightarrow$ 244 actives (posterior > 0.8)
  - RF model $\rightarrow$ 4 actives (score > 0.85)
  - The intersection set contained 4 compounds
- We missed these hits due to our similarity constraints

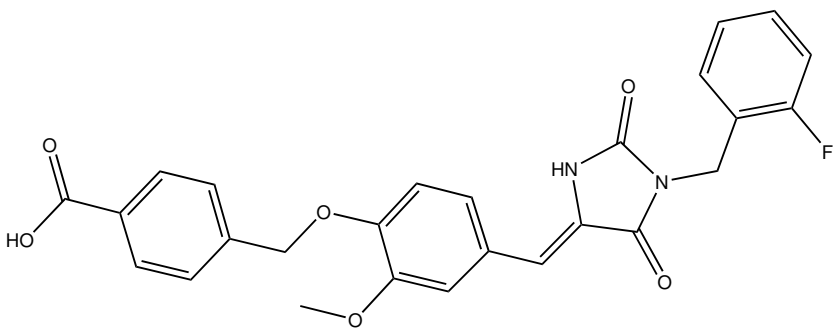
Nair, V; Chi, G.; Neamati, N.; *J. Med. Chem.*, **2006**, *49*, 445–447

Similarity to the Published Compound





Published



Predicted

Nair, V; Chi, G.; Neamati, N.; *J. Med. Chem.*, **2006**, *49*, 445–447

Future Work

Reliability

- Consider similarity to known actives in terms of pharmacophores
- Dock our best hits (may not be conclusive)
- Improve performance with *local* methods

Diversity

- Investigate distribution of vendor compounds in the descriptor space
- Cluster the vendor DB and predict representative members of the clusters

Assay!

Future Work

Reliability

- Consider similarity to known actives in terms of pharmacophores
- Dock our best hits (may not be conclusive)
- Improve performance with *local* methods

Diversity

- Investigate distribution of vendor compounds in the descriptor space
- Cluster the vendor DB and predict representative members of the clusters

Assay!

Future Work

Reliability

- Consider similarity to known actives in terms of pharmacophores
- Dock our best hits (may not be conclusive)
- Improve performance with *local* methods

Diversity

- Investigate distribution of vendor compounds in the descriptor space
- Cluster the vendor DB and predict representative members of the clusters

Assay!

Outline

Summary

- The outlier detection scheme provides a way to analyze spatial distributions of molecules in arbitrary descriptor spaces
- The method can be extended to large datasets using approximate NN algorithms
- The technique results in a single intuitive plot that can be used to easily identify outliers

- A consensus approach to predictive models allows us to increase confidence in activity predictions
- Prioritization based on similarity is intuitive, but may not work well with homogenous datasets
- The protocol appears to have identified possible leads which await confirmation

Summary

- The outlier detection scheme provides a way to analyze spatial distributions of molecules in arbitrary descriptor spaces
- The method can be extended to large datasets using approximate NN algorithms
- The technique results in a single intuitive plot that can be used to easily identify outliers

- A consensus approach to predictive models allows us to increase confidence in activity predictions
- Prioritization based on similarity is intuitive, but may not work well with homogenous datasets
- The protocol appears to have identified possible leads which await confirmation

Acknowledgements

- Prof. P. C. Jurs
- Dr. D. Dutta
- Prof. N. Neamati
- Dr. R. Dayam

Why are R-NN Curves Sigmoidal?

- Let the neighbor density be ρ_0
- At some critical radius, r_0, the neighbor density changes to ρ_1, $\rho_1 >> \rho_0$
- After a certain radius, R, the number of NN's becomes constant, N
- For a radius $r < r_0$, consider a small change dr
- The number of NN's on the strip is $2\pi\rho_0 dr$ which on integration gives $2\pi\rho_0 r_0^2$

Why are R-NN Curves Sigmoidal?

- Let the neighbor density be ρ_0
- At some critical radius, r_0, the neighbor density changes to ρ_1, $\rho_1 >> \rho_0$
- After a certain radius, R, the number of NN's becomes constant, N
- For a radius $r < r_0$, consider a small change dr
- The number of NN's on the strip is $2\pi\rho_0 dr$ which on integration gives $2\pi\rho_0 r_0^2$

Why are R-NN Curves Sigmoidal?

- For a radius $r_0 < r < R$, the number of NN's is given by

$$2\pi\rho_0 r_0^2 + \int_{r_0}^{r} 2\pi r \rho_1 dr$$

- The number of nearest neighbors as a function of r is

$$NN(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ 2\pi\rho_0 r^2 & \text{if } 0 < r \leq r_0 \\ 2\pi\rho_0 r_0^2 + 2\pi\rho_1(r^2 - r_0^2) & \text{if } r_0 < r < R \\ N & \text{if } r \geq R \end{cases}$$

- From Zhang et al., the above function is an approximation of the logistic function

Zhang, M.; Vassiliadis, S.; Delgado-Frias, J.; *IEEE Trans. Comput.*, **1996**, *45*, 1045–1049

Why are R-NN Curves Sigmoidal?

- For a radius $r_0 < r < R$, the number of NN's is given by

$$2\pi\rho_0 r_0^2 + \int_{r_0}^{r} 2\pi r \rho_1 dr$$

- The number of nearest neighbors as a function of r is

$$NN(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ 2\pi\rho_0 r^2 & \text{if } 0 < r \leq r_0 \\ 2\pi\rho_0 r_0^2 + 2\pi\rho_1(r^2 - r_0^2) & \text{if } r_0 < r < R \\ N & \text{if } r \geq R \end{cases}$$

- From Zhang et al., the above function is an approximation of the logistic function

Zhang, M.; Vassiliadis, S.; Delgado-Frias, J.; *IEEE Trans. Comput.*, **1996**, *45*, 1045–1049

Why are R-NN Curves Sigmoidal?

- For a radius $r_0 < r < R$, the number of NN's is given by

$$2\pi\rho_0 r_0^2 + \int_{r_0}^{r} 2\pi r \rho_1 dr$$

- The number of nearest neighbors as a function of r is

$$NN(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ 2\pi\rho_0 r^2 & \text{if } 0 < r \leq r_0 \\ 2\pi\rho_0 r_0^2 + 2\pi\rho_1(r^2 - r_0^2) & \text{if } r_0 < r < R \\ N & \text{if } r \geq R \end{cases}$$

- From Zhang et al., the above function is an approximation of the logistic function

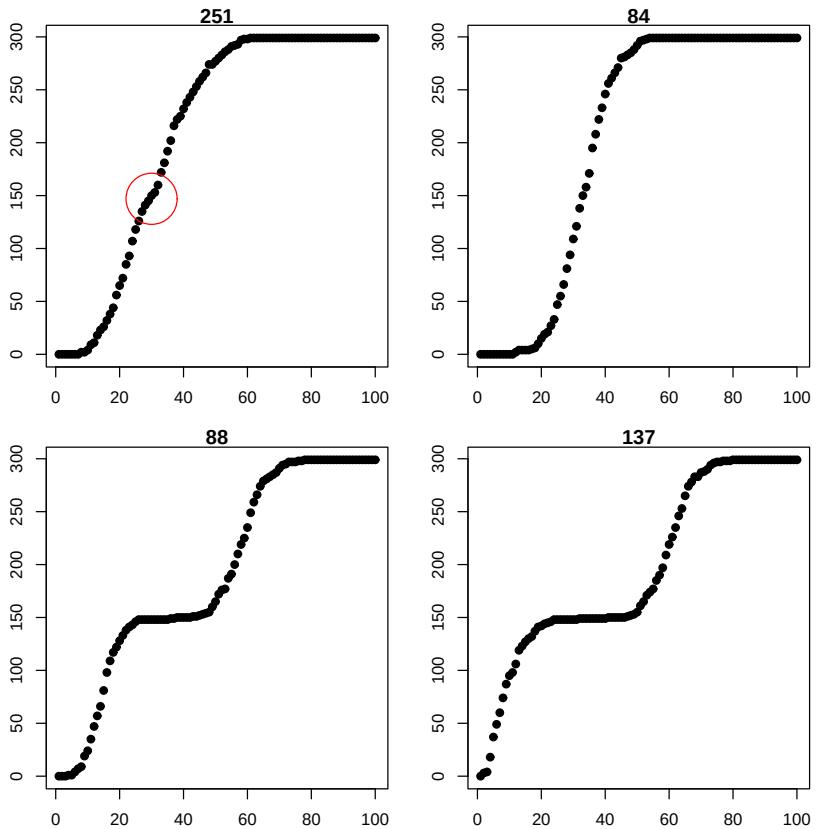
Zhang, M.; Vassiliadis, S.; Delgado-Frias, J.; *IEEE Trans. Comput.*, **1996**, *45*, 1045–1049

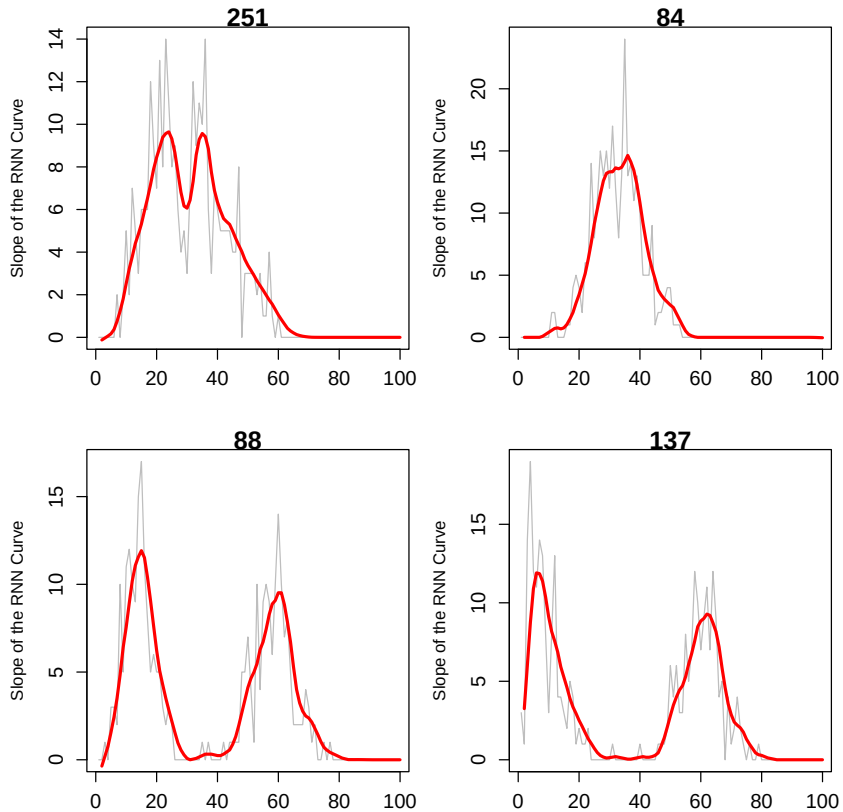
R-NN Curves & Clustering

How Do We Characterize Curves?

- ✗ RMSE
- ✗ Generating a mean curve
- ✓ Examine slopes
- ✓ Piecewise linear approximation
- ✓ Compare distance matrices
- ✓ Shape matching
  - Hausdorff distance
  - Fréchet distance

R-NN Curves & Clustering





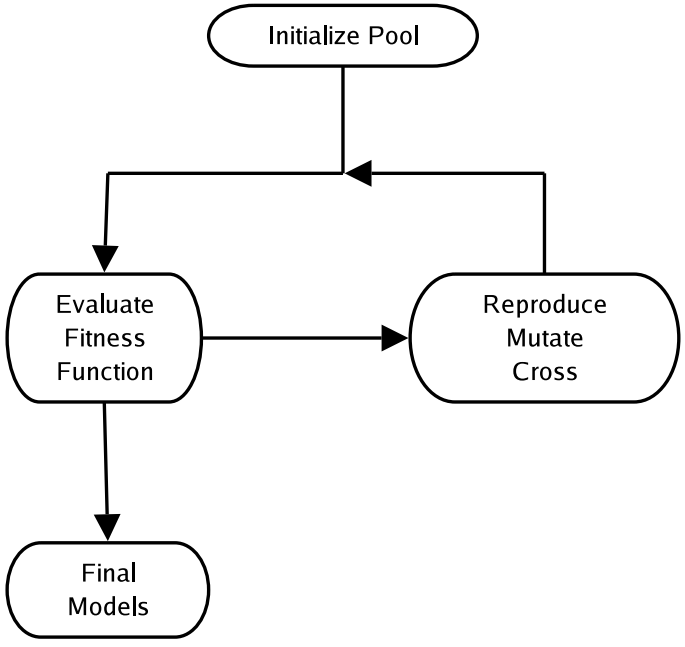
Smoothed with Friedmans *Super Smoother*

Friedman, J.H.; Laboratory for Computational Statistics, Stanford University, *Technical Report 5*, **1984**

Genetic Algorithms

Features

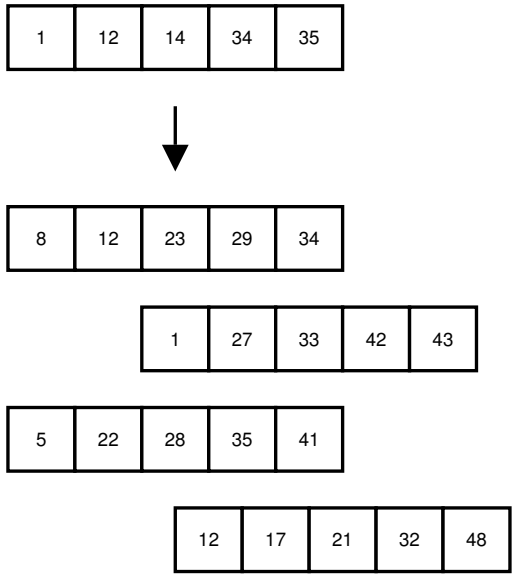
- Stochastic optimization procedure
- Based on evolutionary principles
- Applicable to large search spaces
- Not always guaranteed to find the optimal solution
- Fitness is evaluated by the *objective function*



Genetic Algorithms

Initialization

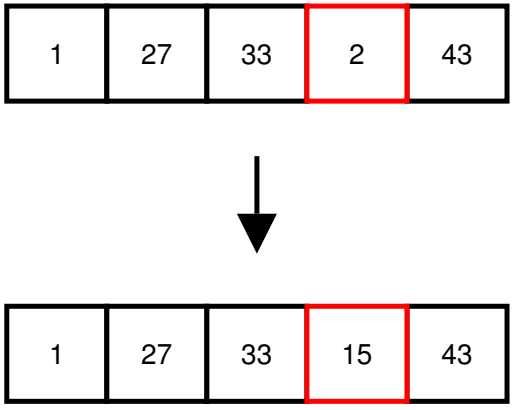
- Initially random chromosomes
- Chromosome length is the size of the descriptor subset
- Genes are the descriptors
- Larger pool sizes allow for a wider search



Genetic Algorithms

Mutation

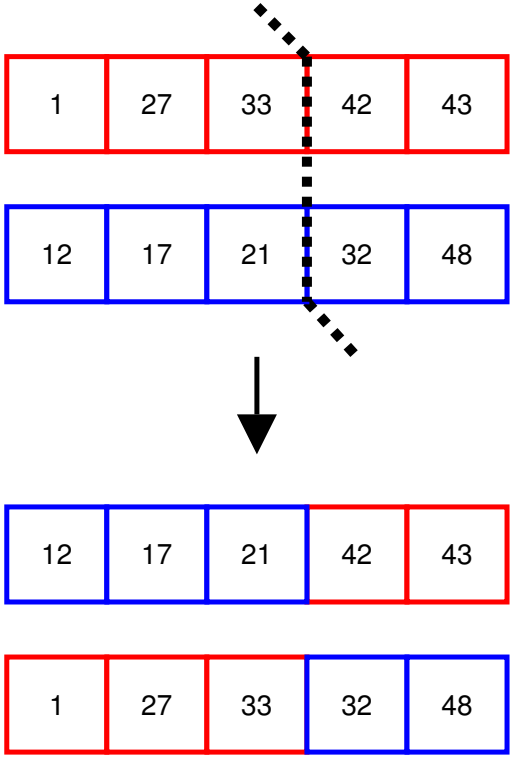
- Mutation frequency is relatively low
- Mutations randomly change a single descriptor
- Helps to get out of local minima



Genetic Algorithms

Crossover

- Crossover results in swapping of genes
- Crossover between fit individuals should lead to children with good aspects of both parents
- In single point crossover
  - Choose crosspoint
  - Swap corresponding sections
  - Results in two new children



Random Forest

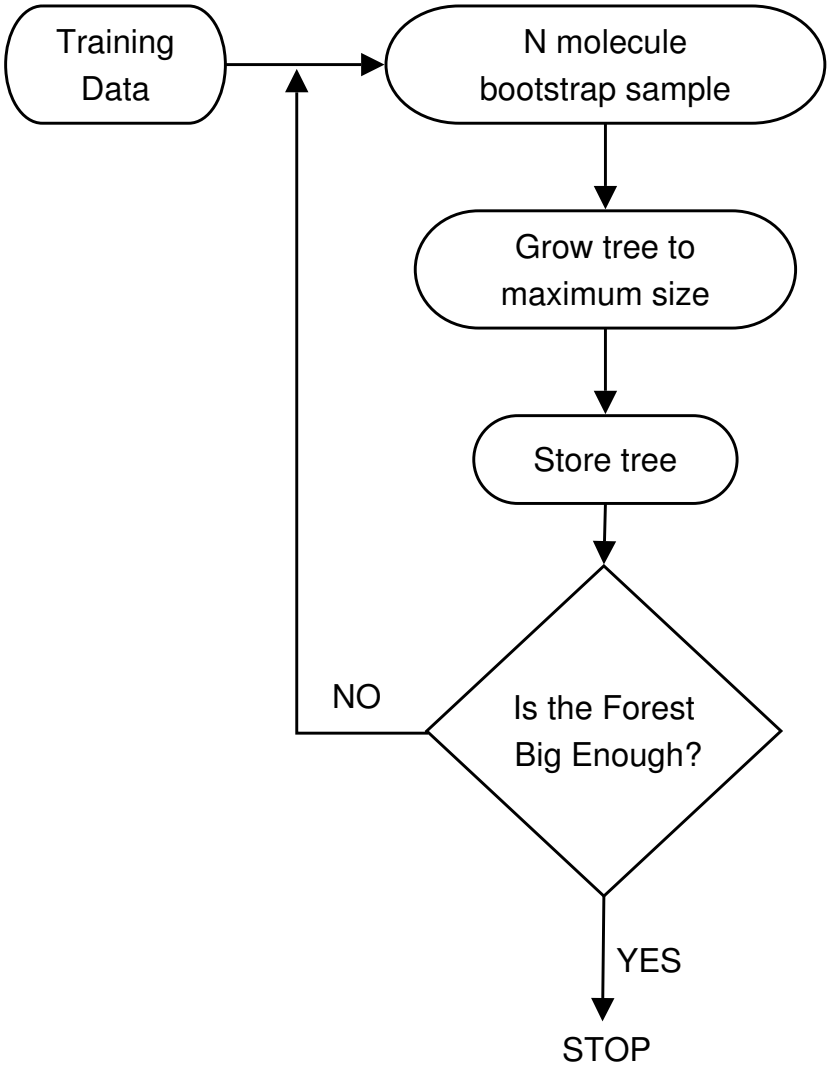
Features

- Built in estimation of prediction accuracy
- Measure of descriptor importance
- Measure of similarity

Random Forest vs. Decision Tree

- RF is faster because it searches fewer descriptors at each split
- A decision tree must be pruned via cross validation. RF trees are grown to full depth
- RF is much more accurate than a decision tree

Random Forest



OOB Samples

What is OOB?

- Due to bootstrap sampling, each tree uses $\approx 2/3$ of the TSET
- The remaining 1/3 is the Out Of Bag sample
- The OOB sample can be used to estimate performance during training or measure descriptor importance

Descriptor Importance

- Decision trees create explicit relationships between descriptors and predictions
- These relationships are hidden in a RF model
- A randomization procedure on OOB samples can be used to rank descriptors in importance

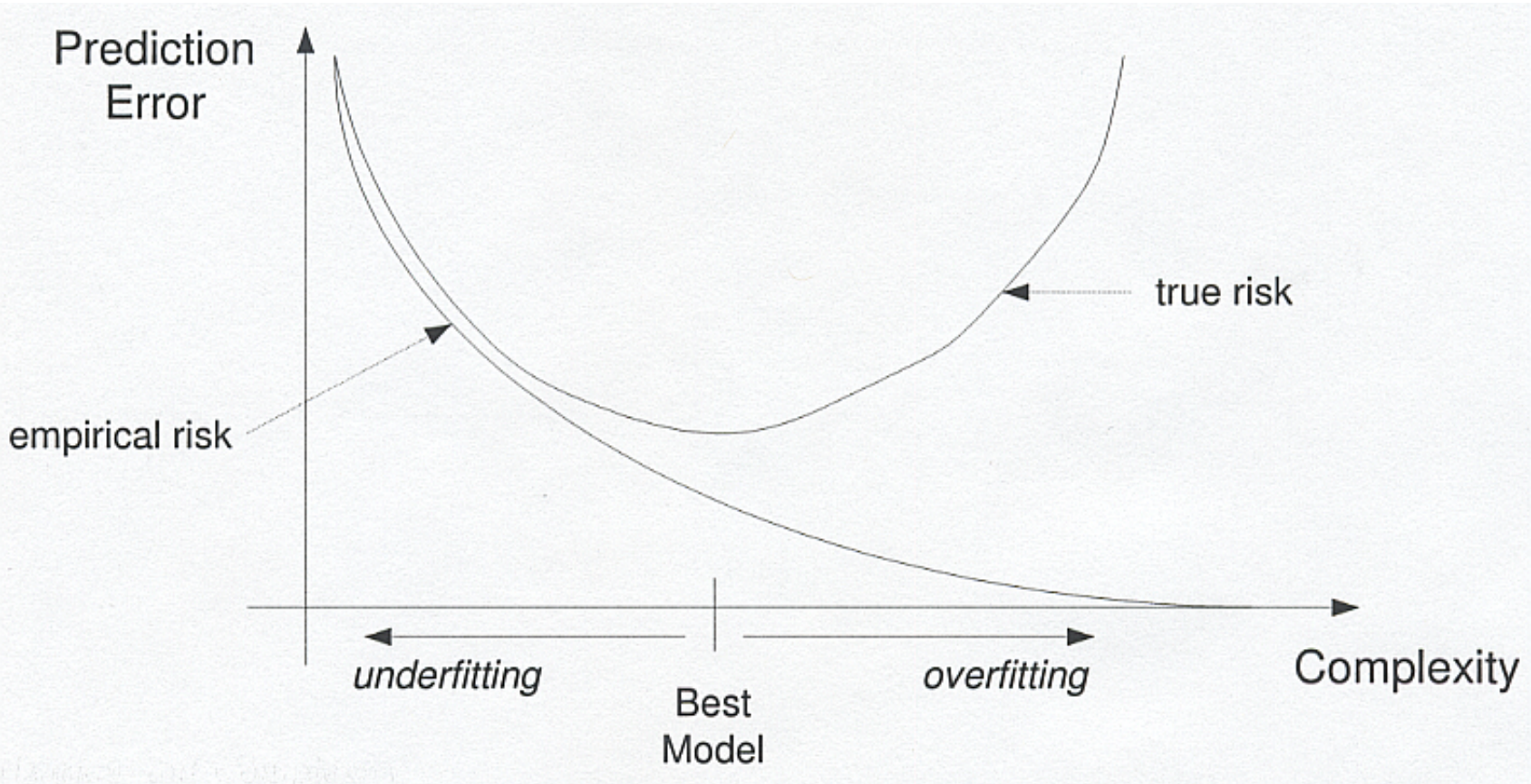
Model Complexity

- Very simple models or very complex models may perform poorly
- The optimal performance and complexity are obtained from a trade off between *bias* and *variance*

$$\textit{Mean Squared Error} = \textit{Var}(\hat{y}) + \textit{Bias}^2(\hat{y}), \quad \text{where}$$

$$\textit{Var} \propto \frac{1}{\textit{complexity}} \quad \text{and} \quad \textit{Bias} \propto \textit{complexity}$$

Model Complexity



http://www.ece.wisc.edu/~nowak/ece901/lecture3.pdf