Matching QSAR Sets

Statistics & Atom Pairs

Rajarshi Guha

Penn State University

Normalizing Distances & Angles

- Normalization does’nt seem to affect the histograms or the Smirnov statistic
- Autoscaling does’nt make any significant changes

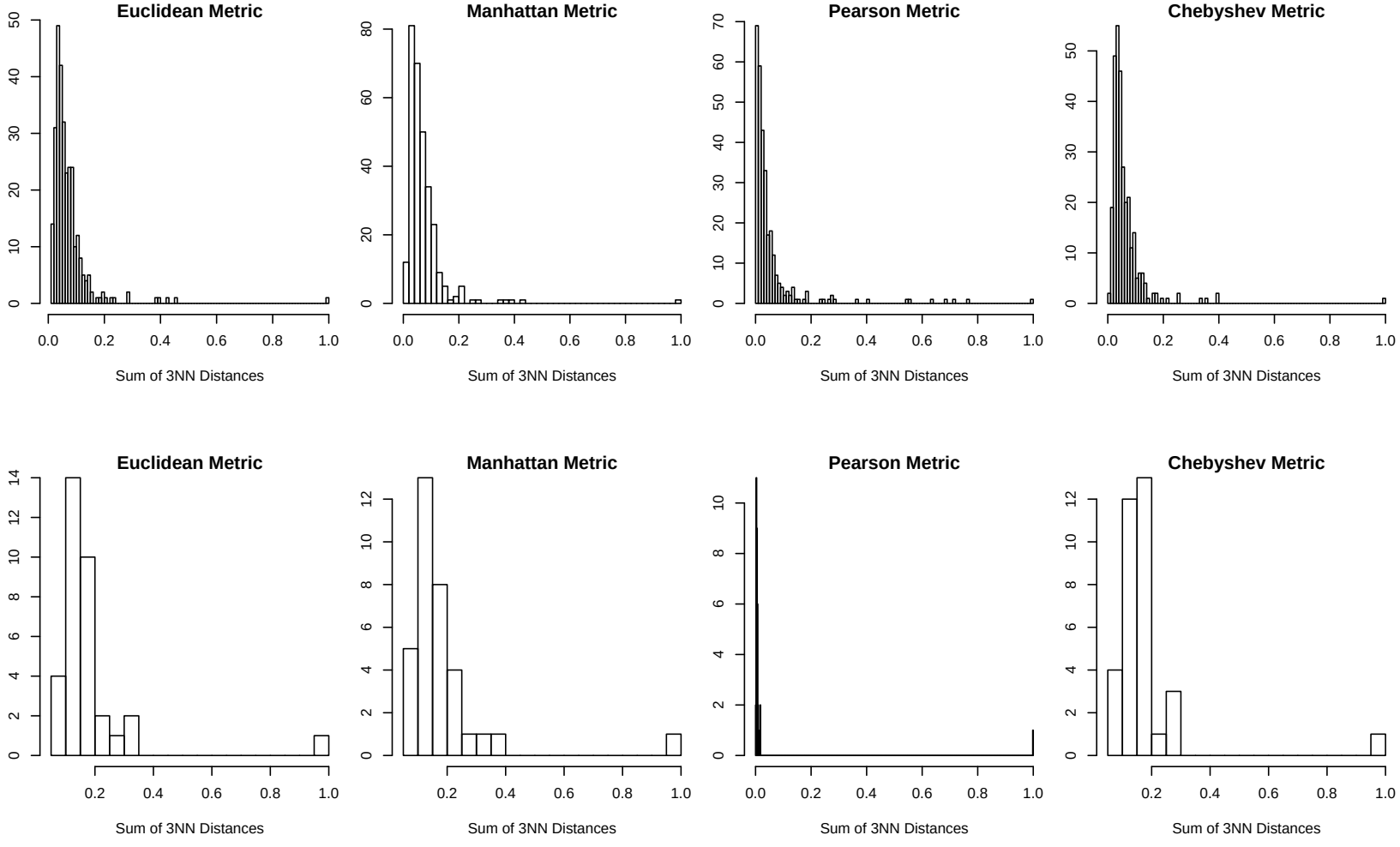
Smirnov Statistics For Normalized Sets

	Angles		
Metric	D(299,34)	Q(0.95)	H_0
Euclidean	0.2015	0.2461	Accept
Manhattan	0.1433	0.2461	Accept
Pearson	0.1513	0.2461	Accept
Chebyshev	0.2124	0.2461	Accept
	Distances		
Euclidean	0.7553	0.2461	Reject
Manhattan	0.7024	0.2461	Reject
Pearson	0.6449	0.2461	Reject
Chebyshev	0.7820	0.2461	Reject

Normalizing Distances & Angles

Sum of 3NN Distances (per molecule) for Varying Distance Metrics

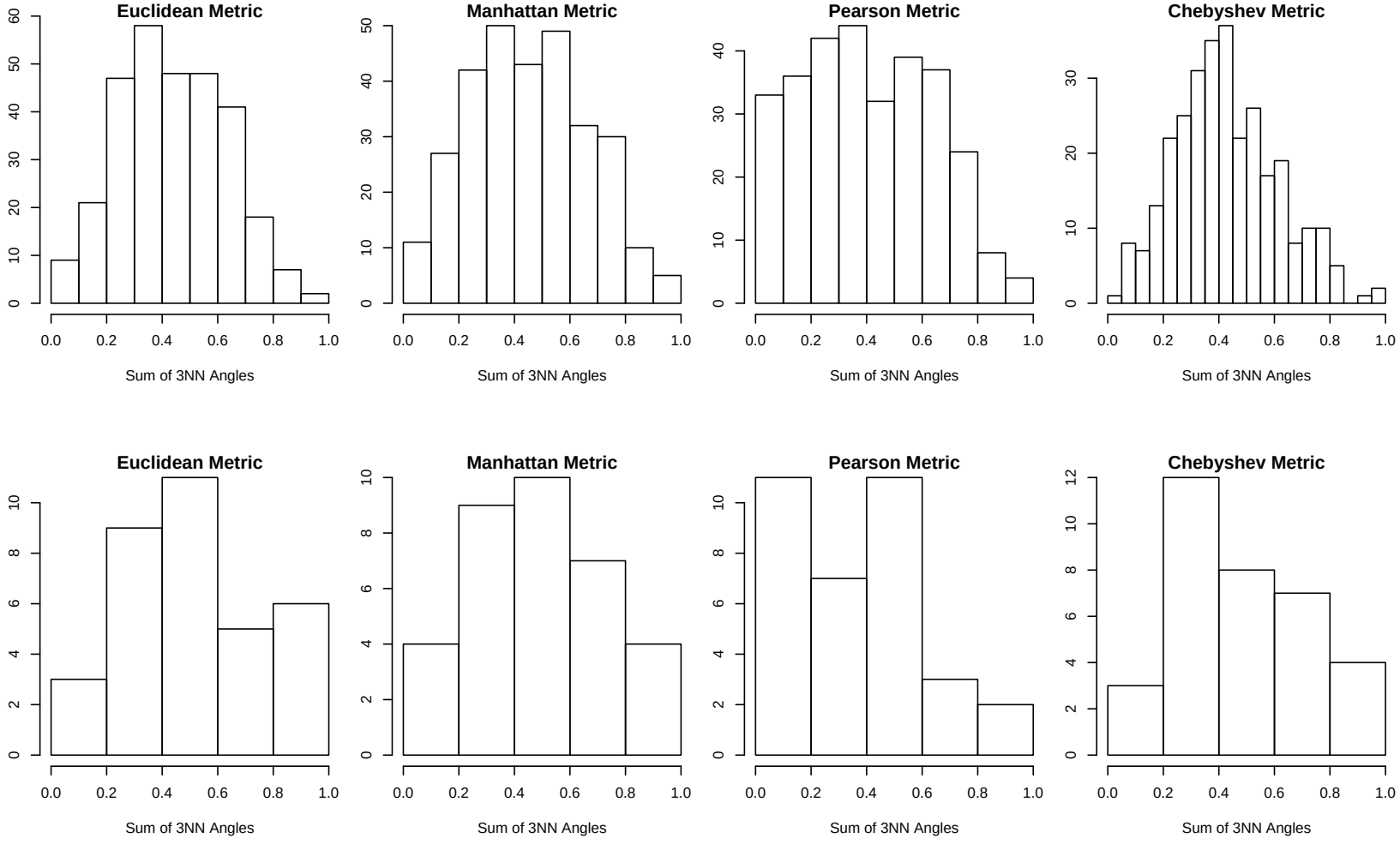
(Upper row is TSET and bottom row is PSET)



Normalizing Distances & Angles

Sum of 3NN Angles (per molecule) for Varying Distance Metrics

(Upper row is TSET and bottom row is PSET)



Confirming Dissimilarities

- To see whether the Smirnov test is actually working the PSET from the tutorial dataset was compared to TSET's from the DHFR data.
- For both angle and distance data the statistic indicates that the two sets are not similar (with p-value $\approx 10^{-16}$!) in some cases.
- However depending on the descriptor used, the test also declares the two sets to be similar

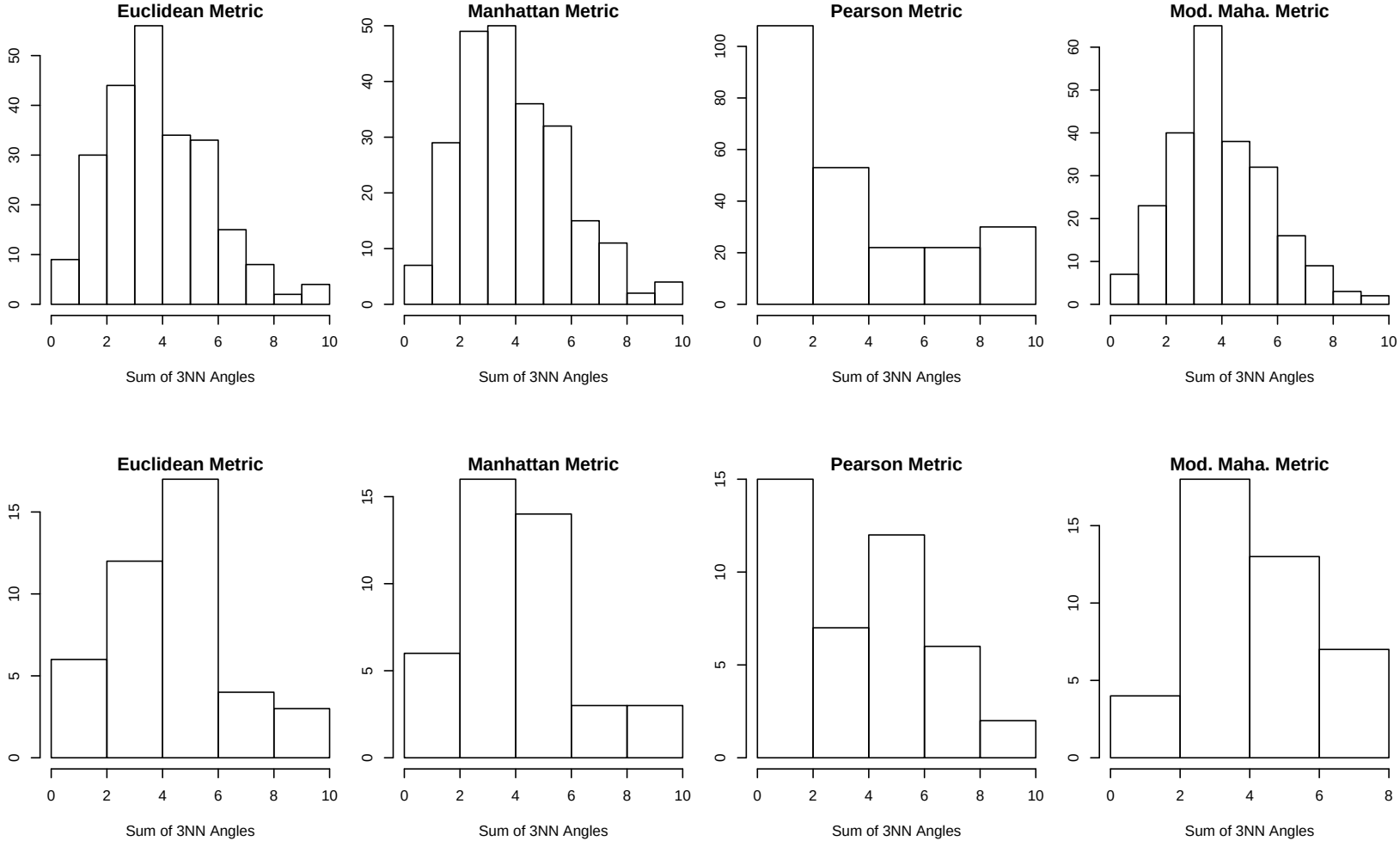
Confirming Dissimilarities

TSET	Angles		Distances	
	D	Q(0.95)	D	Q(0.95)
DHFR - BCUT & Auto	0.9617	0.2495	1.0000	0.2495
DHFR - MoRSE & Auto	0.2846	02.495	0.1554	0.2495
DHFR - Galvez	0.1249	0.2495	0.1554	0.2495
DHFR - Getaway	0.1623	0.2495	0.1810	0.2495

Examining the Mahalanobis Metric

Sum of 3NN Angles (per molecule) for Varying Distance Metrics

(Upper row is TSET and bottom row is PSET)



Atom Pairs

How Can We Use Atompairs?

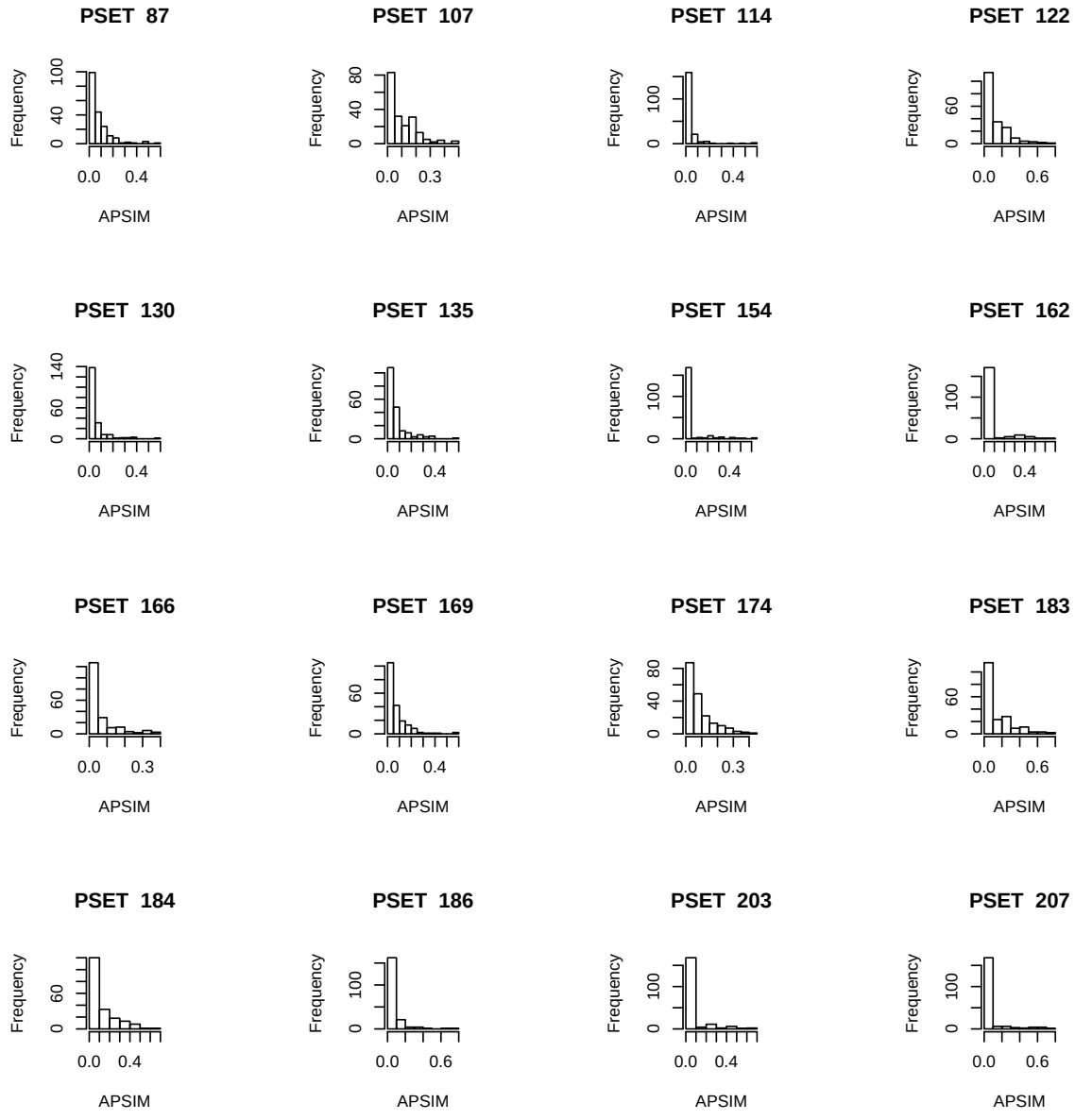
- Ideally we would like to use them as a similarity measure and correlate the similarity with the model performance
- The strategy used was to calculate similarity values (SV) for each PSET molecule with all the molecules in the TSET
- With the matrix of SV's we can investigate
  - distribution of SV's
  - relations between MLR residuals and SV's
  - partitioning the SV's

Distributions of SV's

- For the atom ID calculations, Carharts method and an atomic weight method were used
- A large number of SV's ended up being 0 in both cases

Distribution of SV's

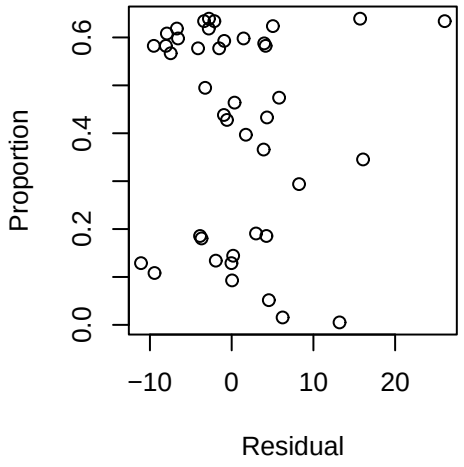
Distribution of Atom Pair Similarity Values for Each PSET Molecule with the the Whole TSET



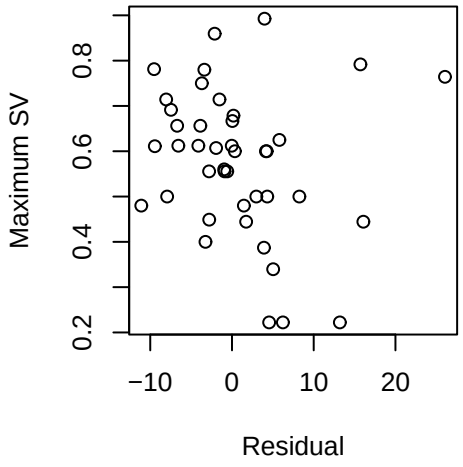
MLR Residuals & SV's

Residuals Obtained From a 4 Descriptor MLR Model

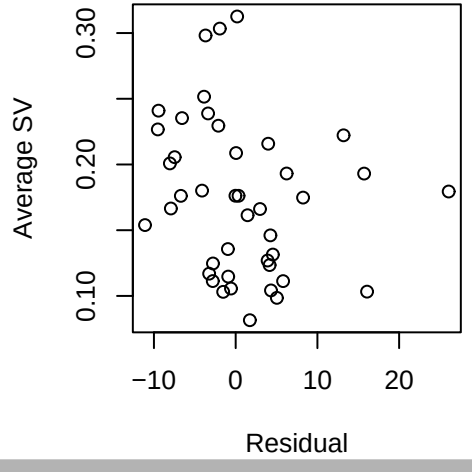
Residuals vs Proportion of Non Zero SV



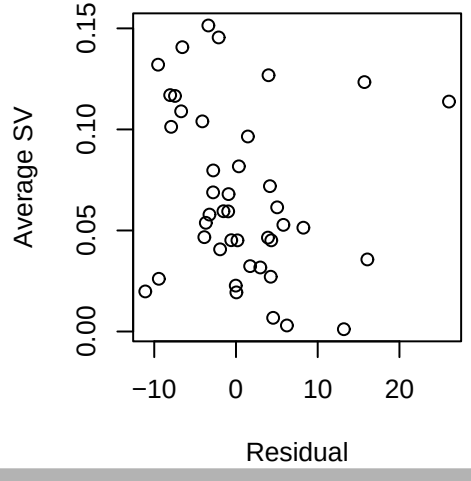
Residuals vs Maximum SV's



Residuals vs Average Of Non Zero SV's



Residuals vs Average Of All SV's



Fuzzy Analysis Clustering

- In fuzzy clustering each observation is spread out over multiple clusters
- We denote $u(i, v)$ be the membership of observation i to cluster v
- The Fanny algorithm minimizes

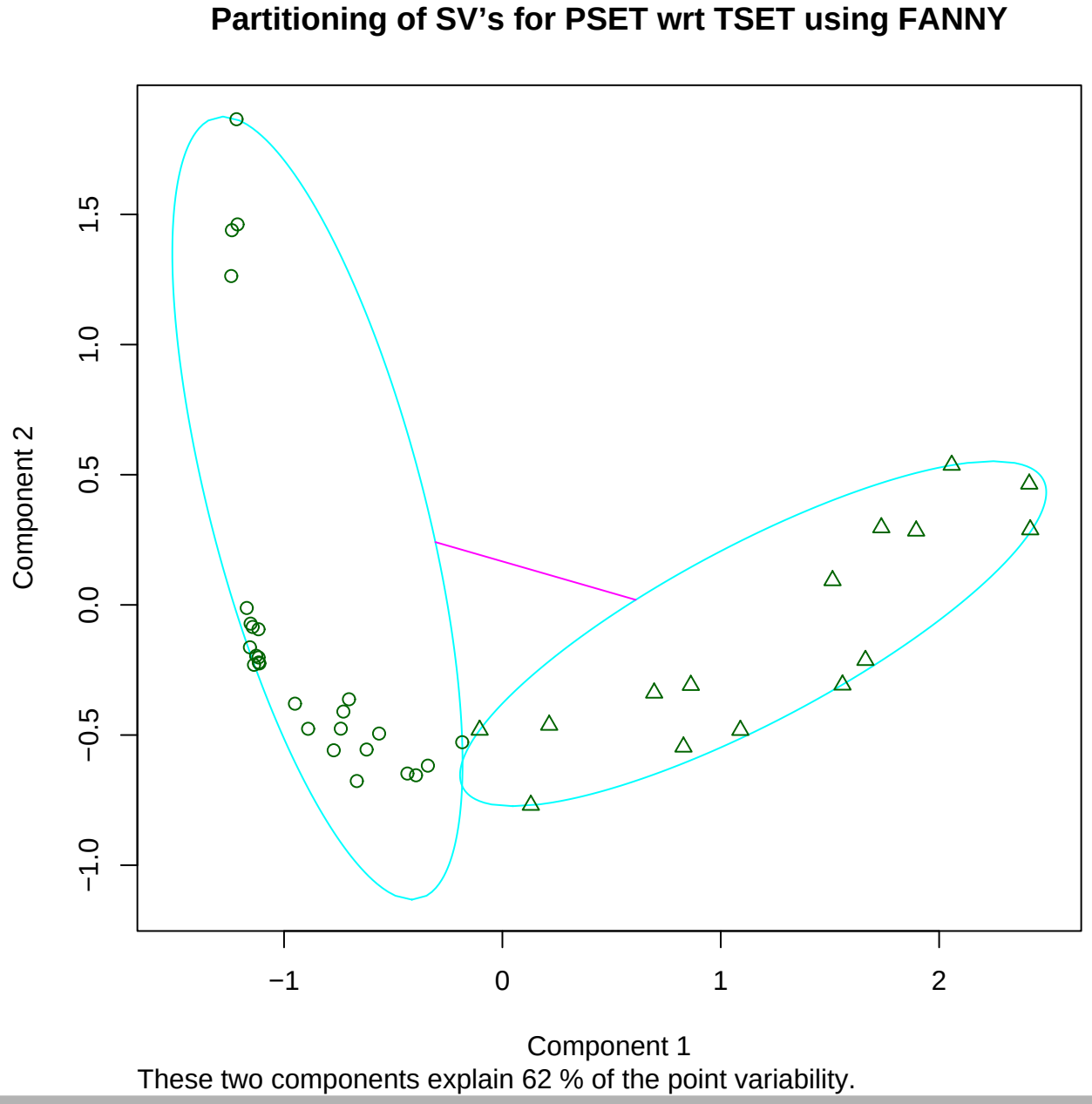
$$\sum_{v=1}^{k} \left(\sum_{i,j} u(i,v)^2 u(j,v)^2 d(i,j) / \, 2 \sum_{j} u(j,v)^2 \right)$$

where n is the number of observations, k is the number of clusters and $d(i, j)$ is the dissimilarity between observations i and j

Fuzzy Analysis Clustering

- What are we partitioning?
  - We consider the individual SV's for each PSET molecule as a variable
  - The Fanny algorithm uses these to calculate dissimilarity values
- Why partition?
  - It is possible that a group of PSET molecules appears to match the TSET in some way. Hopefully the Fanny algorithm will be able to use the similarity values to detect this
  - Investigate any correlations between cluster members and their MLR residuals

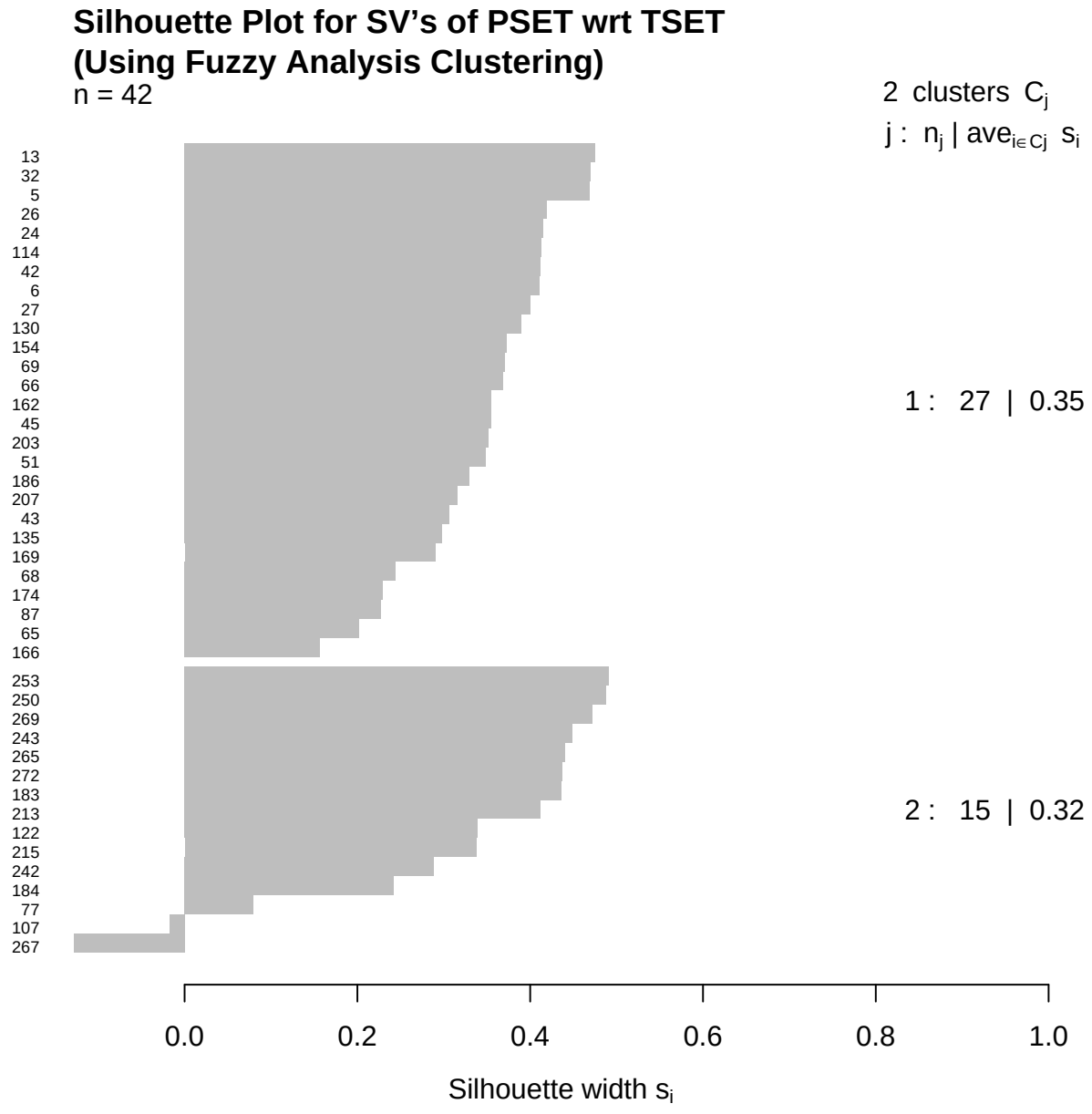
Fuzzy Analysis Clustering



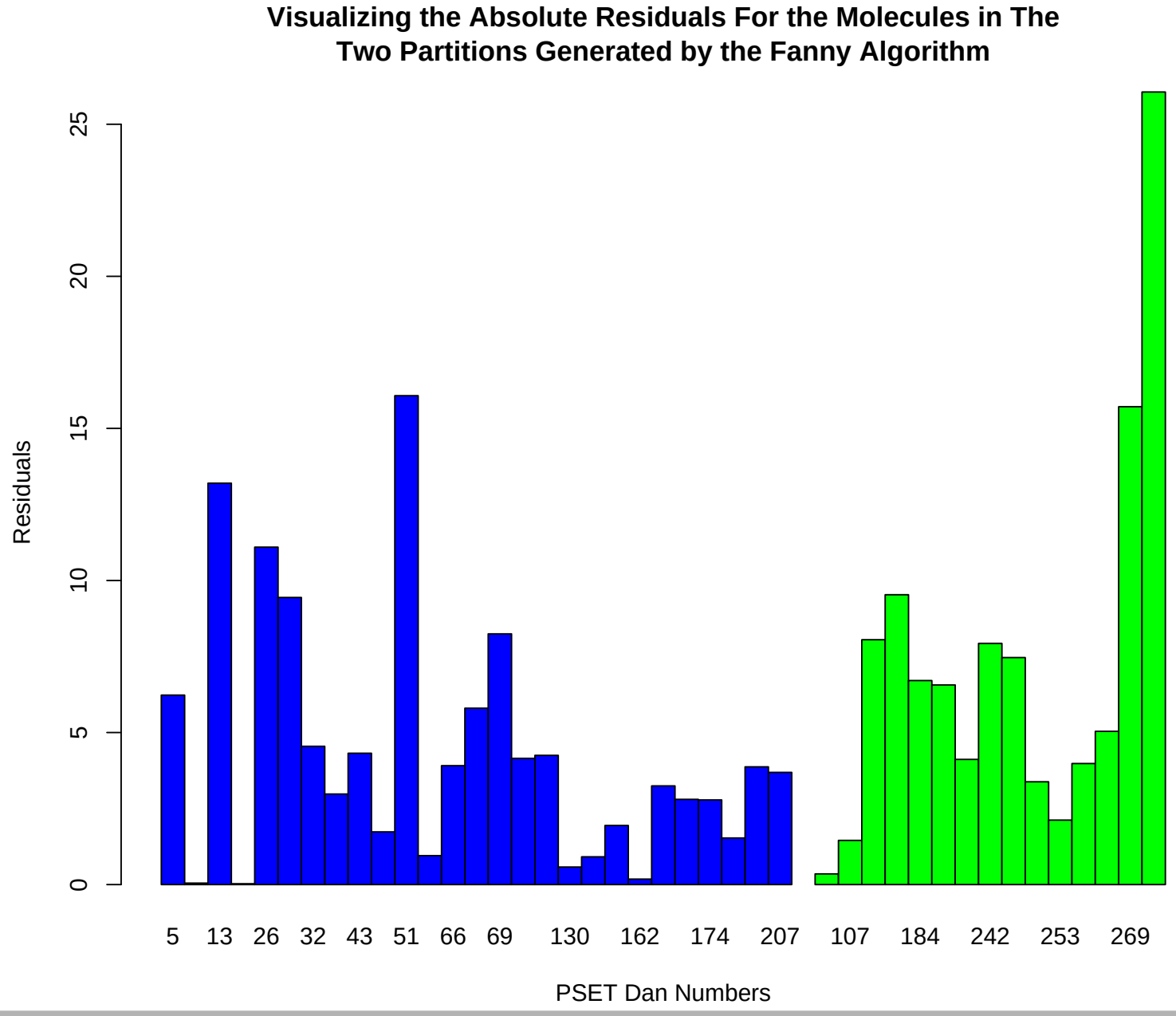
Fuzzy Analysis Clustering

- The algorithm was instructed to generate two clusters
- Increasing k to 3 or 4 did not lead to more clusters. However $k = 6$ did generate 3 clusters
- Dunn's partition coefficient was 0.5 (higher indicates crisper clustering)
- The silhouette coefficient is 0.34 indicating a weak, possibly artifical cluster structure
- There does not appear to be any distinct correlation between cluster membership and MLR residuals

Silhouette Plot of the Fanny Clusters



Cluster Memberships and MLR Residuals



Other Clustering/Partitioning Algorithms

- Methods investigated include
  - Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- All methods divided the SV's into more or less the same clusters

A Sphere Algorithm

- The general idea was to draw a TSET and a PSET sphere and investigate whether any properties of this approach would show any correlation with an MLR model

Summary of the Algorithm

- Using the molecules in the TSET
  - Find the centroid of the TSET
  - Find the TSET member that is furthest from the TSET (using Euclidean distance)
  - This distance represents the radius of the sphere enclosing the TSET (denoted as R_{tset})
- Repeat for the PSET to obtain R_{pset}

Using the Spheres

- A method similar to the *molecular diversity index* was evaluated
- For the PSET
  - Find all PSET members whose distance from the TSET centroid is less than R_{tset}
  - Evaluate the ratio of number of PSET members satisfying the above condition to the total number of PSET members
  - Denote the ratio as I_t
- Repeat the above for the TSET to obtain the ratio I_p

Results of the Sphere Approach

- In general one of I_t or I_p is always 1
- There does'nt seem to be any obvious correlation:

		R^2		I	
Data	Num. Desc.	TSET	PSET	T	P
Artemisinin	4	0.68	0.77	1.00	0.93
glass - BCUT	10	0.86	0.67	0.96	1.00
glass - BCUT	4	0.72	0.59	1.00	0.98
dhfr - BCUT	3	0.36	0.34	1.00	0.96
random	4	0.31	0.04	1.00	0.95

Silhouettes

Viewing the Quality of a Clustering

- Silhouettes are graphical means of viewing a clustering
- Each cluster is represented by a silhouette
- The silhouette shows which objects lie within the cluster and which ones are intermediate
- This method is useful when dissimilarities are on a ratio scale

Viewing the Quality of a Clustering

- For each object, i, we define a term $s(i)$
- First, let A be the cluster to which i is assigned
- The let $a(i)$ be the average dissimilarity of i to all other objects in A
- Then consider a cluster C ($C \neq A$)
- Let $d(i, C)$ be the average dissimilarity of i to all objects in C, and compute this for all $C \neq A$
- Let $b(i) = \min d(i, C)$

Viewing the Quality of a Clustering

- Finally

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

- Thus

$$-1 \leq s(i) \leq 1$$

Viewing the Quality of a Clustering

- Some properties of $s(i)$ include
  - An $s(i)$ close to 1 indicates that i was well classified
  - An $s(i)$ around 0 indicates that i's membership is not certain and a value close to -1 indicates a misclassification
  - In a silhoulette plot, wide silhouettes indicate good clustering