Matching QSAR Sets

Sphere Algorithms & Atom Pairs

Rajarshi Guha

Penn State University

Sphere Algorithm

- Based on a modification of the diversity index
- In the original algorithm spheres are drawn around the TSET points
- Its aim was to pick PSET points near TSET points

Sphere Algorithm - Modification

- In this version spheres are drawn around PSET points
- It assumes that these spheres will occupy *most* of the volume encompassing the TSET points
- For a given sphere look at the TSET points lying inside it
- Can we get any useful information from these included points related to some feature of the PSET point in question?

Sphere Algorithm - Terminology

- Centreset: Set of points in ND space at which spheres are drawn
- Checkset: Set of points which are considered to lie within the spheres centered at centerset points
- APS: Atom Pair Similarity
- SEoP: Standard Error of Prediction

Sphere Algorithm - Method

- Evaluate the volume enclosed by the TSET points

$$V = \prod_{j=1}^{k} \left(X_{max,j} - X_{min,j} \right)$$

- Find the volume for one of the points
- Assuming this is an N dimensional sphere caclulate its radius. Optionally scale it (our friend c)
- Once we have the radius we can use the points in the centerset and draw spheres around them
- Find out how many checkset points are present in each sphere and evaluate a density for each sphere

Sphere Algorithm - Method

- The problem with this is that it is possible that a PSET point lies outside of the volume encompassing the TSET points
- So we calculate V using the whole dataset
- A further improvement is to use an occupied volume rather than raw volume
- Occupied volume is calculated using a Monte Carlo approach

The Sphere Algorithm With the Artemisinin Dataset

Sphere Algorithm - Artemisinin Dataset

- The dataset consisted of 179 molecules and 65 descriptors in the reduced pool
- The TSET had 161 molecules and PSET 18 molecules

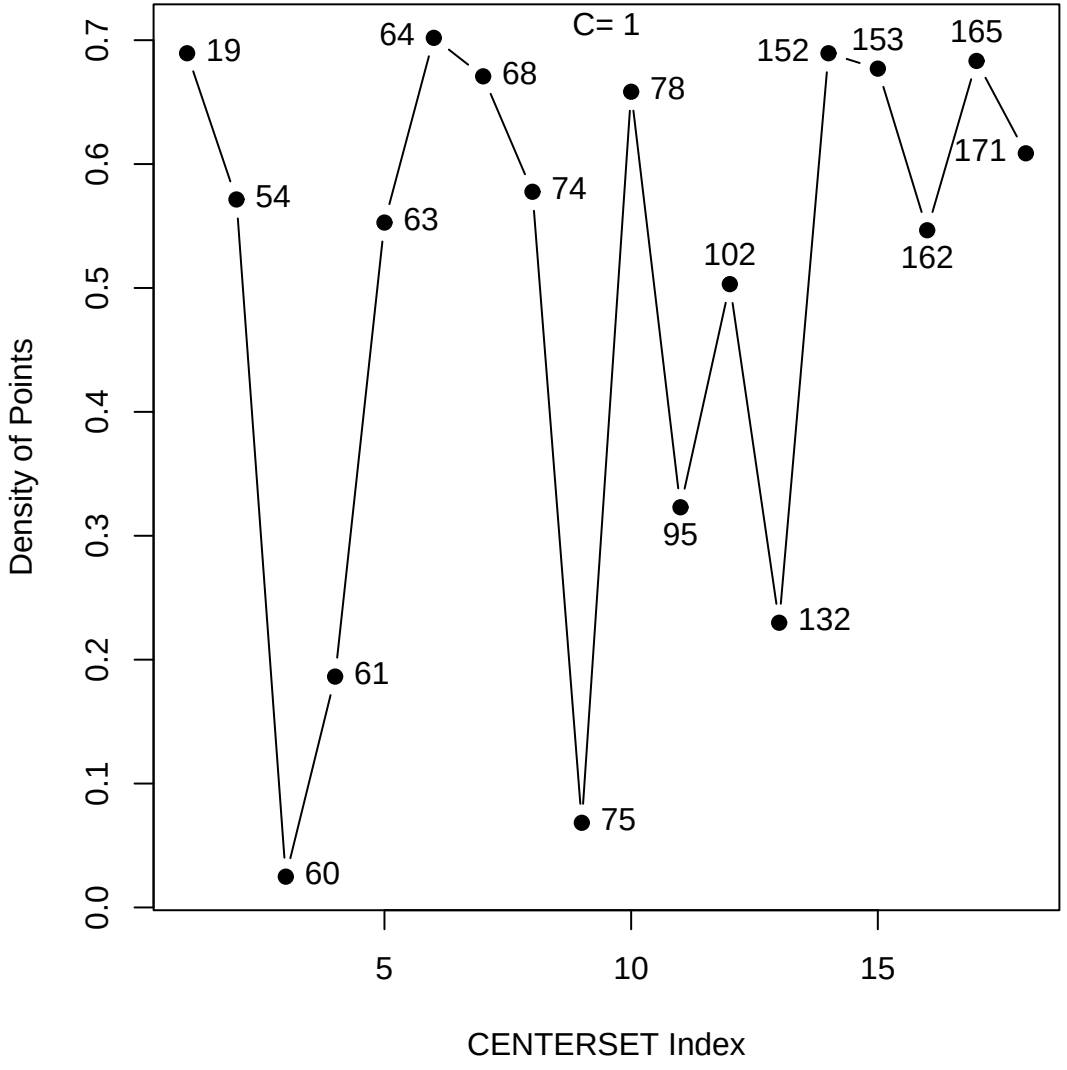
Sphere Algorithm - Artemisinin Model

- Bets model in terms of statistics. But not very impressive visually!

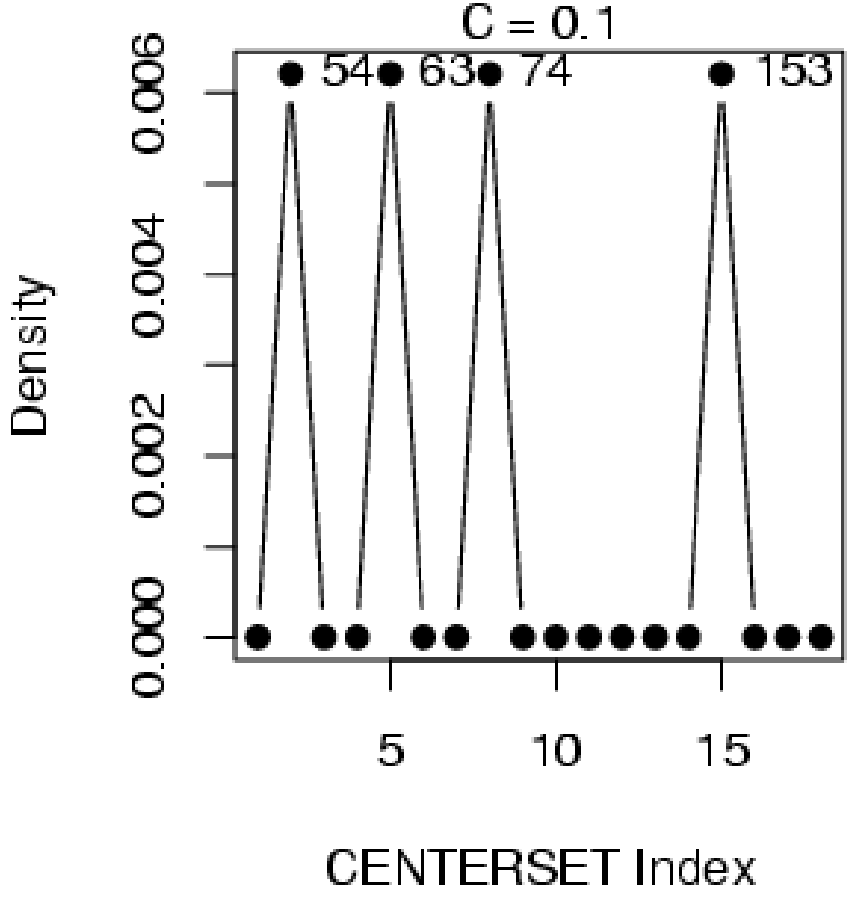
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.55550    5.29283 -11.819  < 2e-16 ***
dN7CH.20     -0.22214    0.01311 -16.949  < 2e-16 ***
dNSB.12       0.22423    0.02308   9.717  < 2e-16 ***
dWTPT.2      28.90461    2.61434  11.056  < 2e-16 ***
dMDE.14       0.13231    0.02759   4.795 3.77e-06 ***

Residual standard error: 0.887 on 156 degrees of freedom
Multiple R-Squared: 0.7096,     Adjusted R-squared: 0.7021
F-statistic: 95.28 on 4 and 156 DF,  p-value: < 2.2e-16

Variance Inflation Factors:
  dN7CH.20  dNSB.12  dWTPT.2  dMDE.14
  1.624504 1.346523 1.466014 1.549218
```

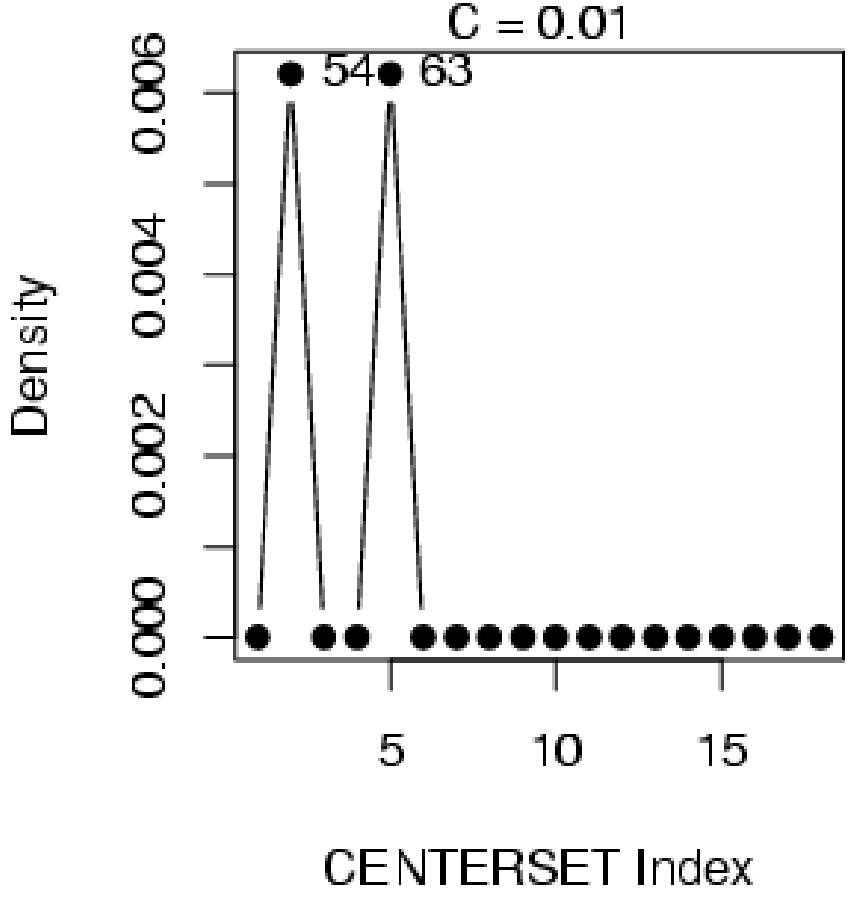
Sphere Algorithm - Densities

Plot of Sphere Density for Each PSET Point vs. PSET Index



Sphere Algorithm - Scaling the Radius



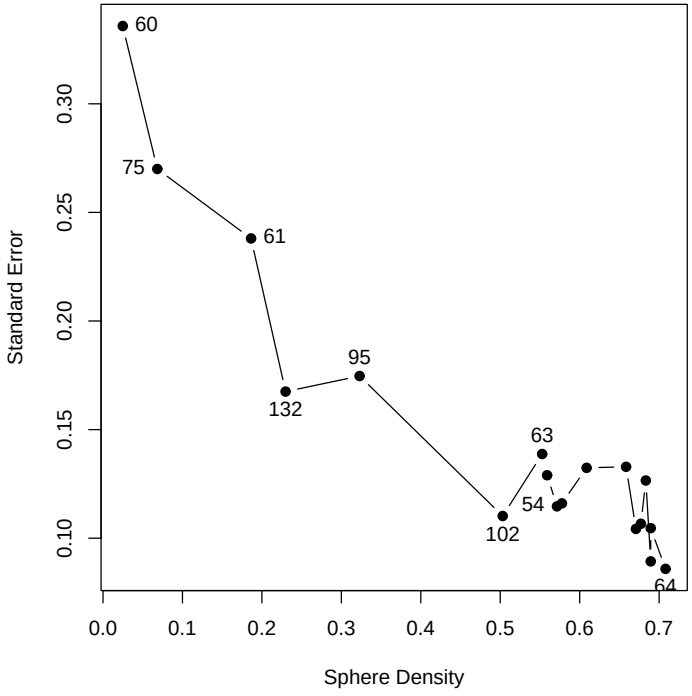


Sphere Algorithm - Scaling the Radius

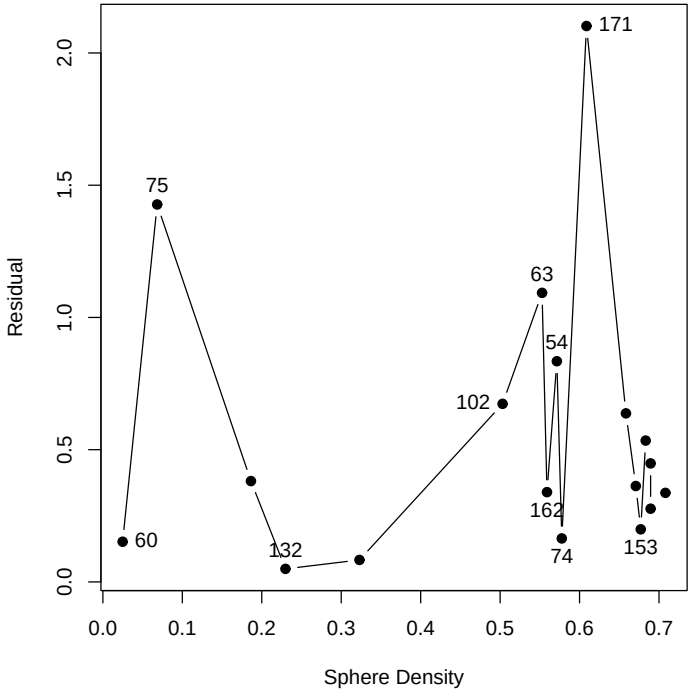
- Decreasing the radius implies that fewer TSET points will be present in the PSET spheres
- Thus as we scale the sphere radius down, only PSET molecules that are surrounded by lots of TSET molecules will have non-zero density
- Thus: as we scale the radius down, the PSET molecules that are 'closest' to the TSET should show non-zero densities
- **Will these PSET points have lower SE of prediction or residuals?**

Sphere Algorithm - Density, Residuals & SE's

Plot of Sphere Density for Each PSET Sphere vs. Standard Error of Prediction for PSET Points



Plot of Sphere Density for Each PSET Sphere vs. MLR Residual for PSET Points



- It appears that residuals don't really correlate well with sphere density
- SE of predictions seem to follow an inverse trend

Sphere Algorithm - Comment

- SE's can be considered as an indication of reliability of the prediction
- Large SE's indicate that the confidence limits of the prediction are large
- Thus: lower sphere densities might indicate that the prediction will not be reliable. Matches intuition!

The Sphere Algorithm With a Toy Dataset

Toy Dataset - Overview

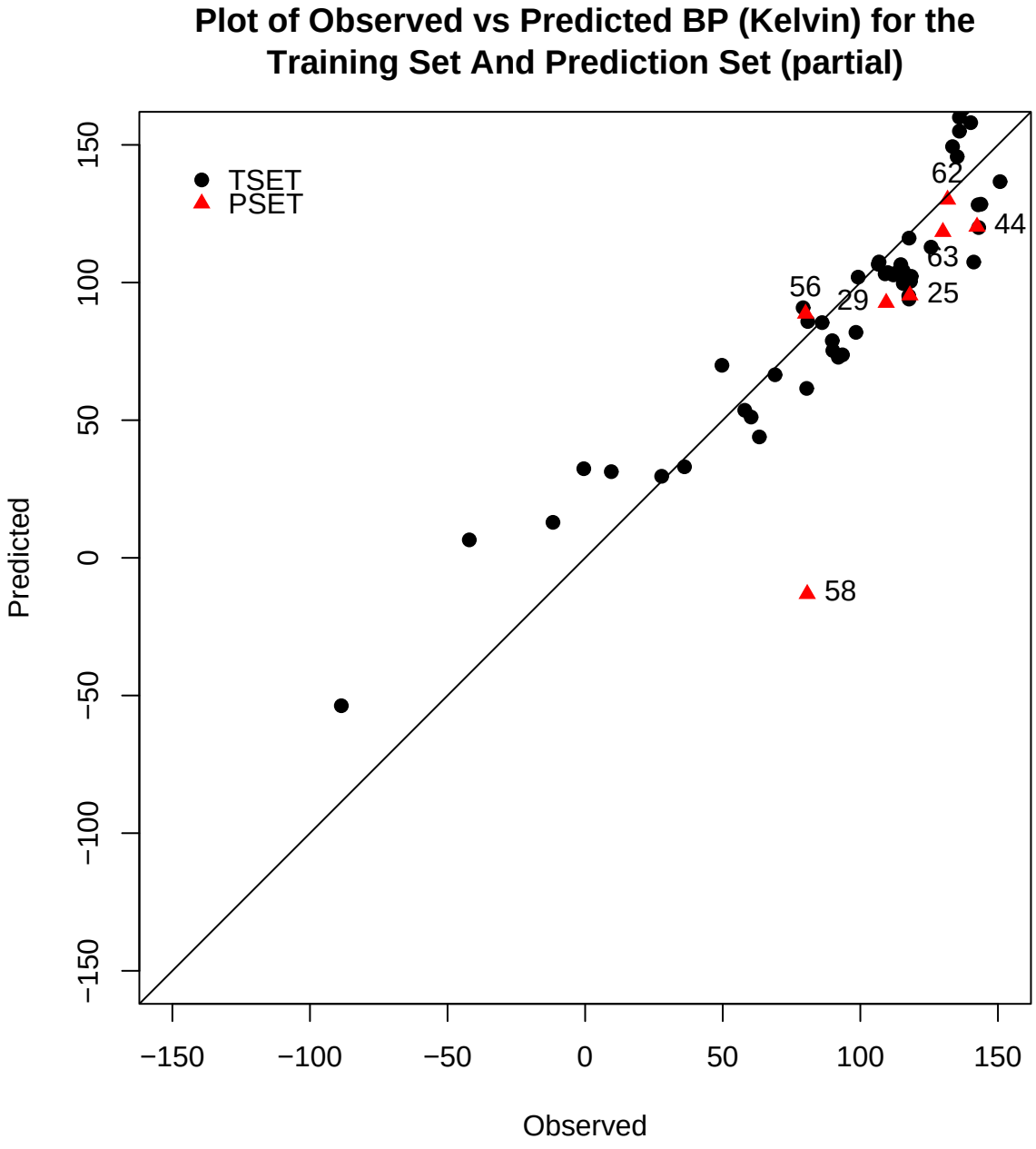
- This was designed to provide distinct outliers as well as a few well predicted points
- 65 molecules. 57 taken from JCICS, **1998**, *38*, 387-394 and were all straight chain alkanes and associated isomers
- 8 molecules were randomly selected to be as different from the 57 as possible (eg. benzene, pyrrole, anthracene)
- Out of the 57 molecules, 52 were placed in the training set and 5 were placed along with the 8 external compounds in the prediction set
- The dependent variable was boiling point.

Toy Dataset - Linear Model

- A 4 descriptor linear model was generated

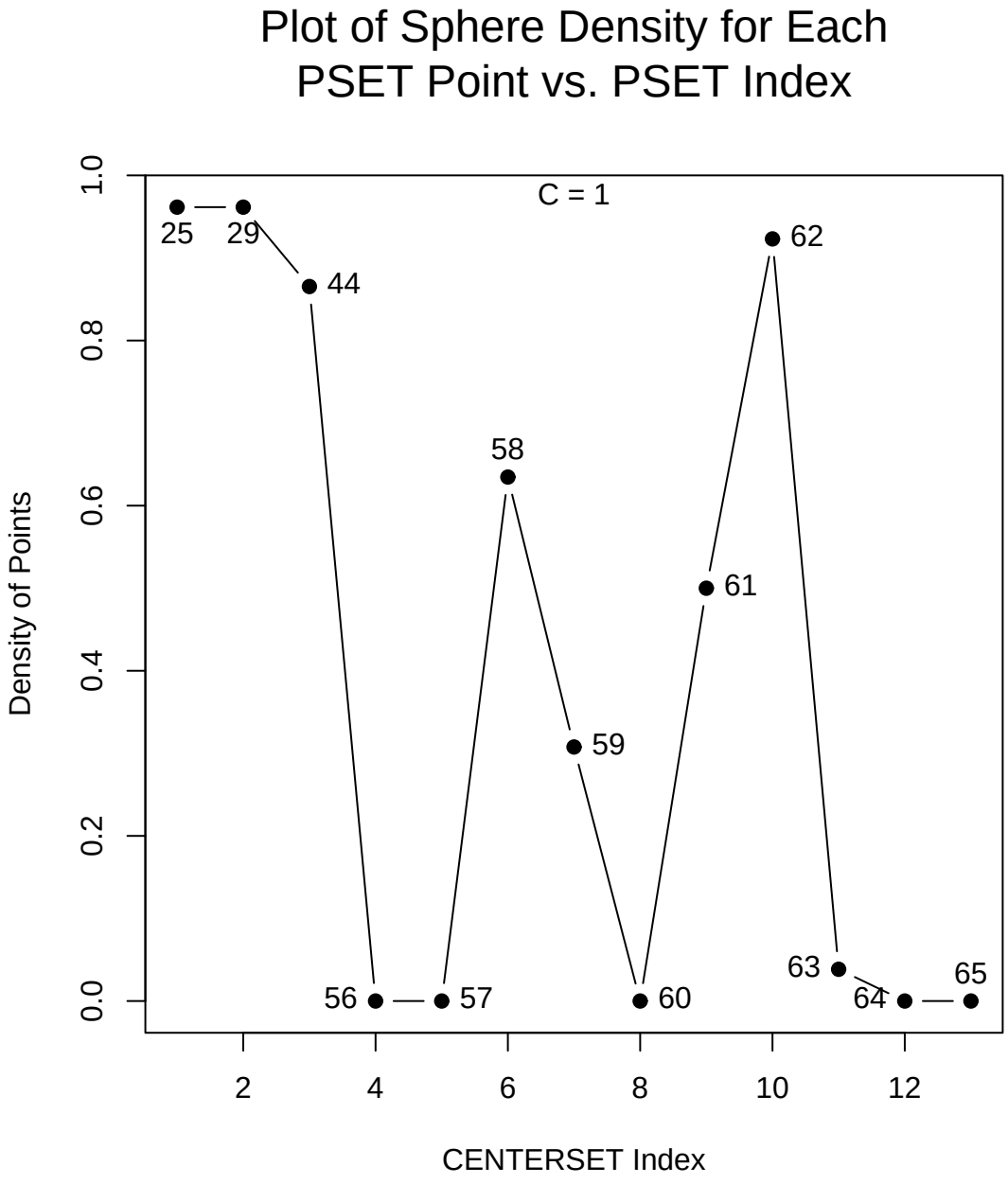
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -381.6960    60.3677  -6.323 8.72e-08 ***
dEMIN.1      -43.2189     9.1003  -4.749 1.95e-05 ***
dEMAX.1       88.8862    10.4446   8.510 4.46e-11 ***
dECCN.1        1.2717     0.1052  12.089 4.99e-16 ***
dSHDW.6      501.1936   136.7371   3.665 0.000627 ***

Residual standard error: 19.82 on 47 degrees of freedom
Multiple R-Squared: 0.905,      Adjusted R-squared: 0.8969
F-statistic: 111.9 on 4 and 47 DF,  p-value: < 2.2e-16

Variance Inflation Factors:
   dEMIN.1  dEMAX.1  dECCN.1  dSHDW.6
   1.114556 1.486670 1.205632 1.241283
```

Toy Dataset - Plot



Toy Dataset - Plot

Sphere Algorithm - Densities



Sphere Algorithm - Zero Density Compounds

- The compounds with zero densities are expected to have *very* little in common with the majority of the dataset
- The table below indicates that this seems to be true (they are all from the 8 external compounds)

Dan	Name
56	benzene
57	benzoic acid
60	bromomethane
64	anthracene
65	acetic acid

Sphere Algorithm - High Density Compounds

- The compounds with the highest densities are expected to be much more similar to the majority of the dataset
- From the plot, the highest compounds turn out to be from the original dataset

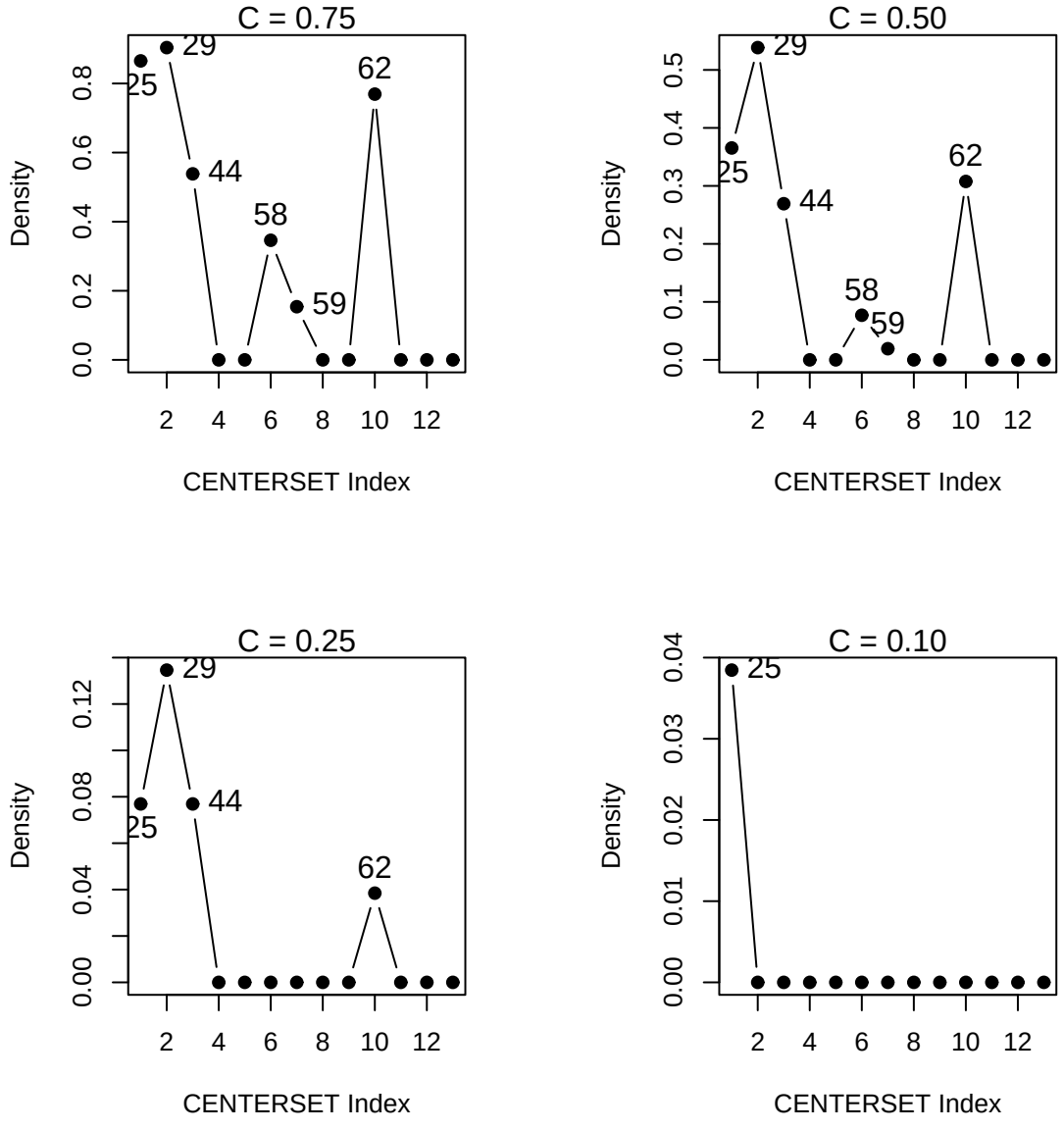
Dan	Name
25	3-methylheptane
29	2,4-dimethylhexane
44	4-methyloctane
62	2,2,3-trimethylhexane

Sphere Algorithm - High Density Compounds

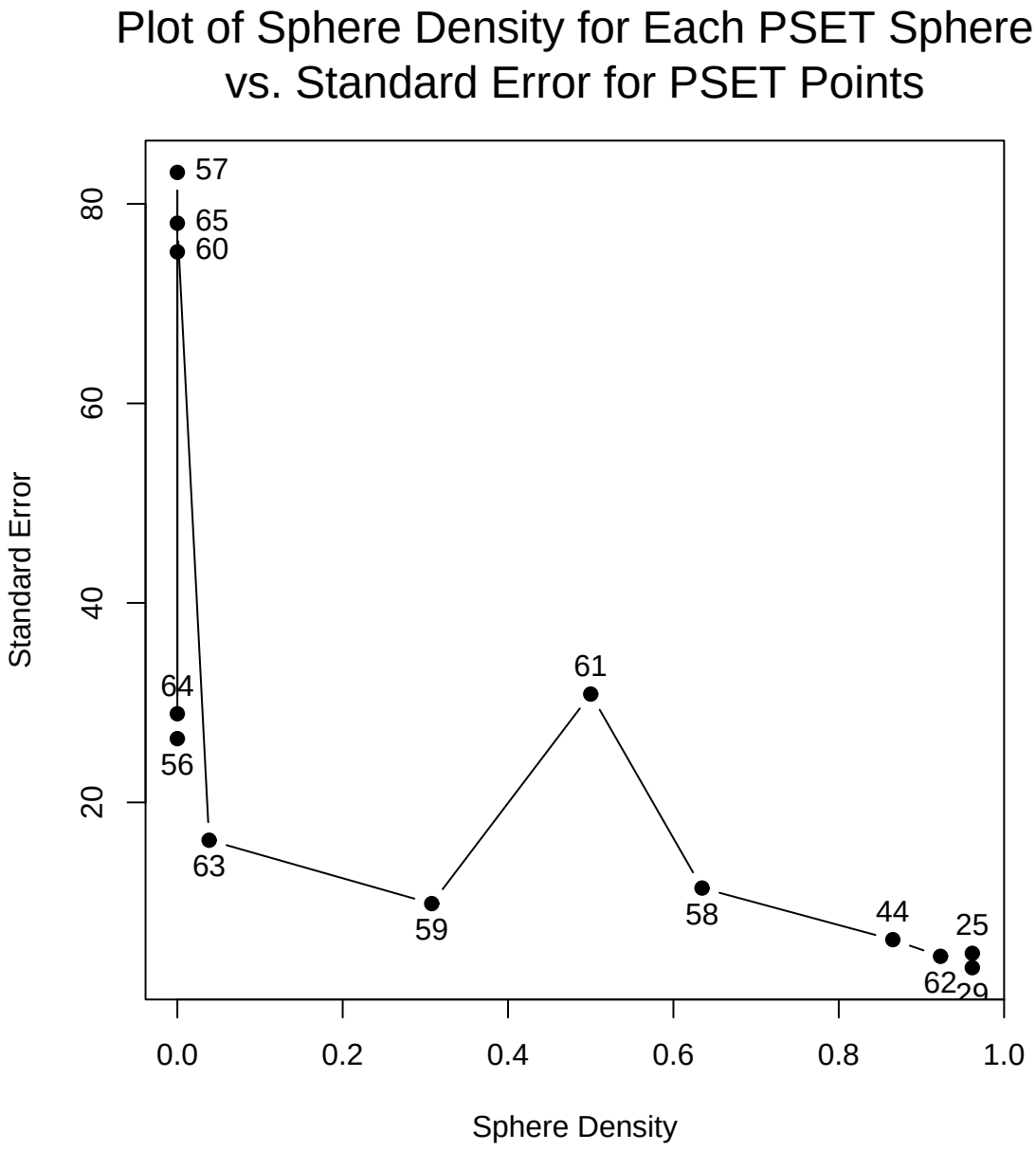
- It is interesting to see that decane (59) has an average density
- Encouraging since the training set contains upto C_9 but decane is still a straight chain alkane and should not be entirely unrecognizable
- Hence the medium density value

Sphere Algorithm - Scaling the Radius

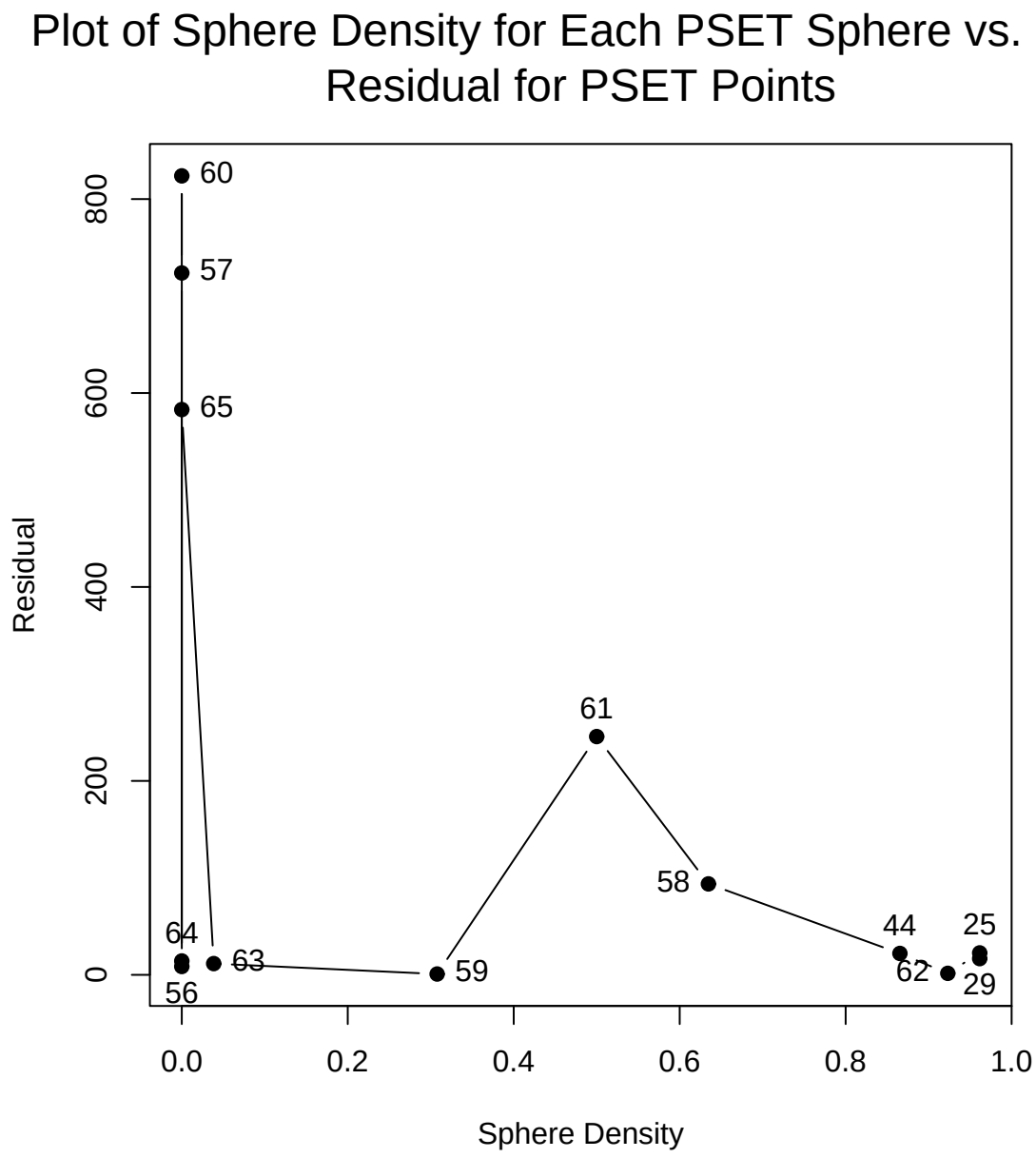
Variation of Density With Scaled Radii



Sphere Algorithm - Density & Errors



Sphere Algorithm - Density & Residual



Sphere Algorithm - Comments

- Unlike the artemisinin dataset, the behavior of sphere density with residual and SEoP for the toy dataset are quite similar.
- The general trend of lower density and higher residual/error is present
- However this trend is quite obscured in the case of the toy dataset

The Sphere Algorithm & Atom Pairs

Sphere Algorithm - Using Atom Pairs

- The sphere algorithm provides us with a set of TSET points surrounding a PSET point
- Ideally, we would like to avoid use of specific descriptors when making the spheres and analyzing their contents
- Atom pairs allow us to look at the contents of the spheres

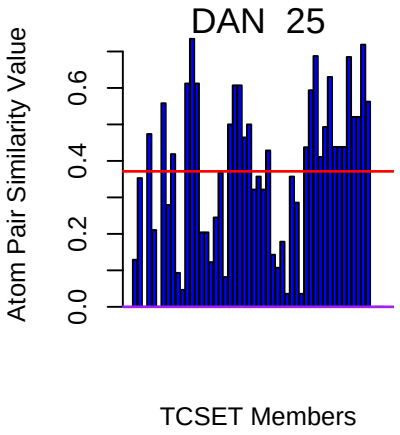
Sphere Algorithm - Atom Pair Method

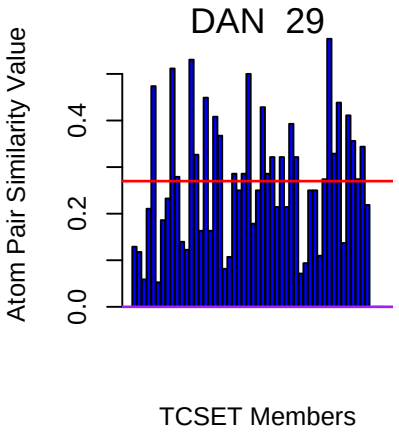
- Create spheres as before
- Calculate atom pair similarities between the PSET point and all the TSET points in the sphere
- Calculate atom pair similarities between the PSET point and all TSET points outside the sphere
- It is expected that

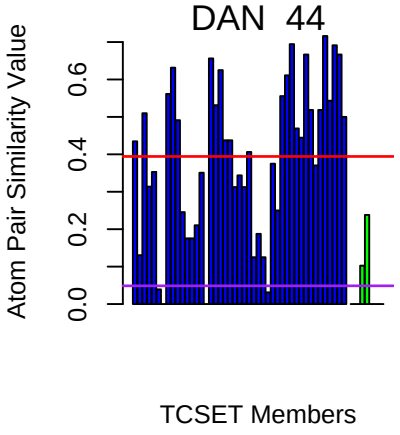
$$\overline{AP}_{inside} > \overline{AP}_{outside}$$

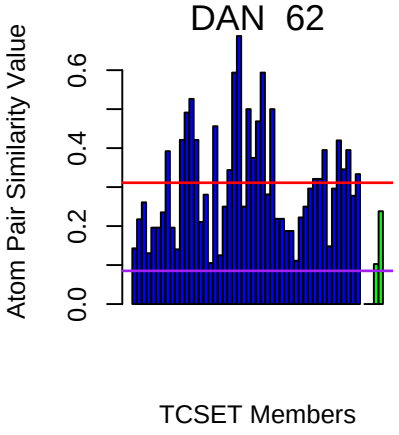
Sphere / AP - Toy Dataset

Barplots of AP Similarity Values Between Each PSET Point
And TSET Members Inside And Outside its Sphere





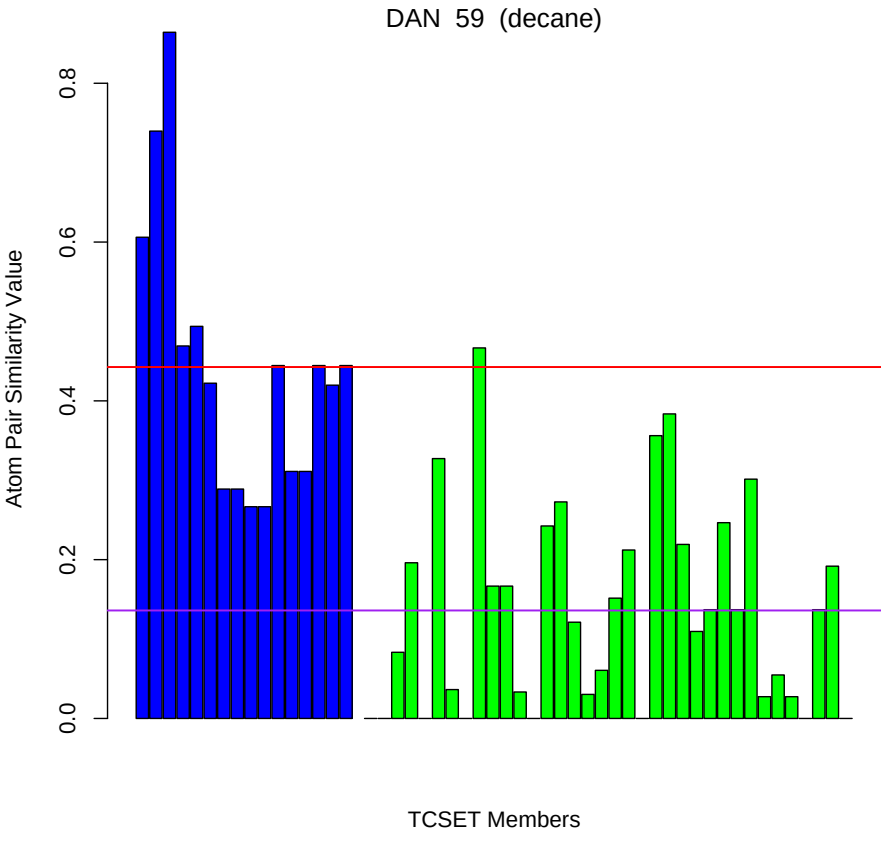




Sphere / AP - Comments on Distribution

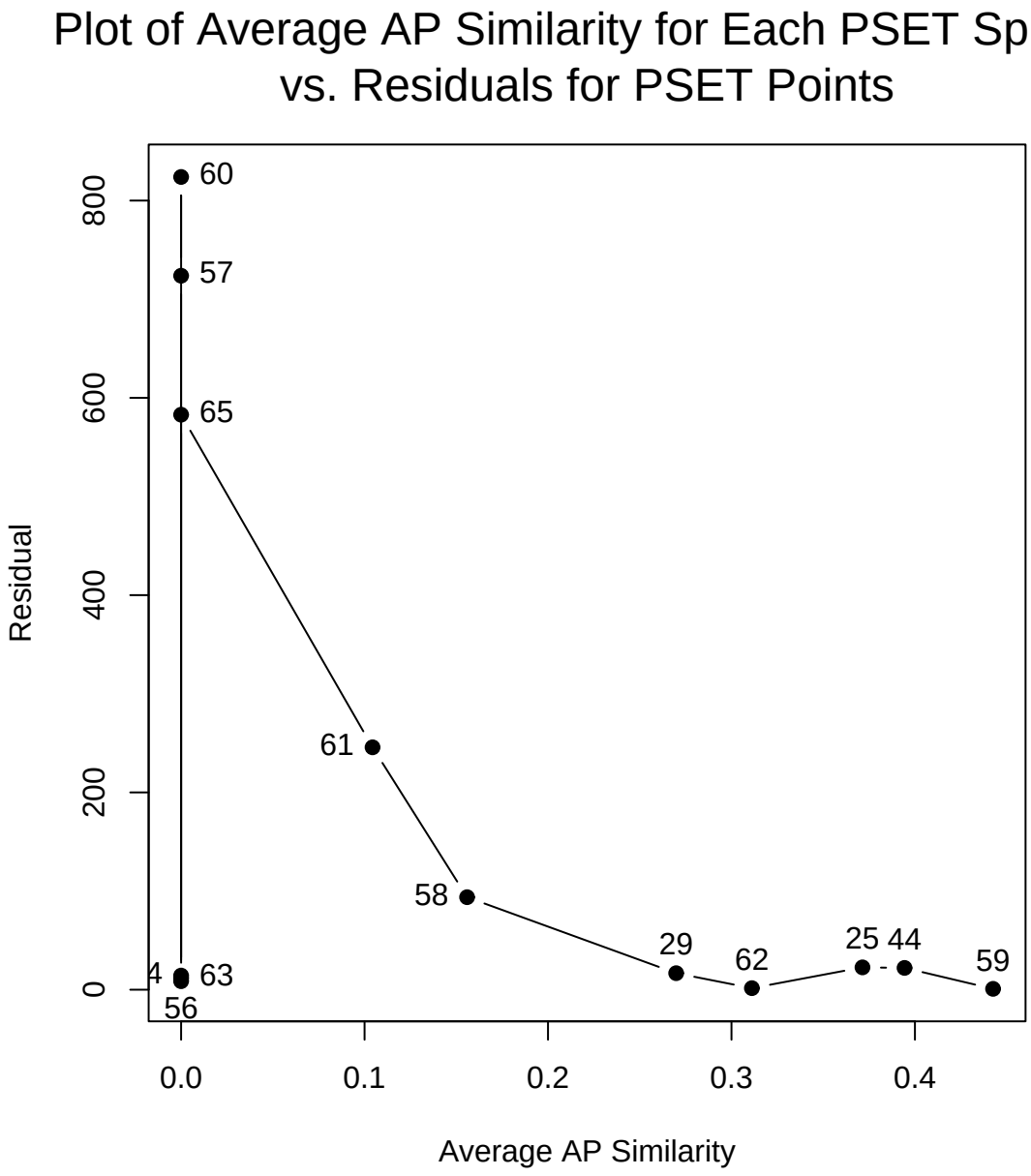
- Only the distributions for the 4 PSET members that had the most TSET members in their spheres are shown
- A number of PSET members had *no* TSET members in their spheres (cf. density plots)
- Dan 63 was anomalous since it only had 2 points within its sphere and the AP similarity values between both points was 0 (in fact all APS values are 0 for 63 - pyrrole)

Sphere / AP - Decane

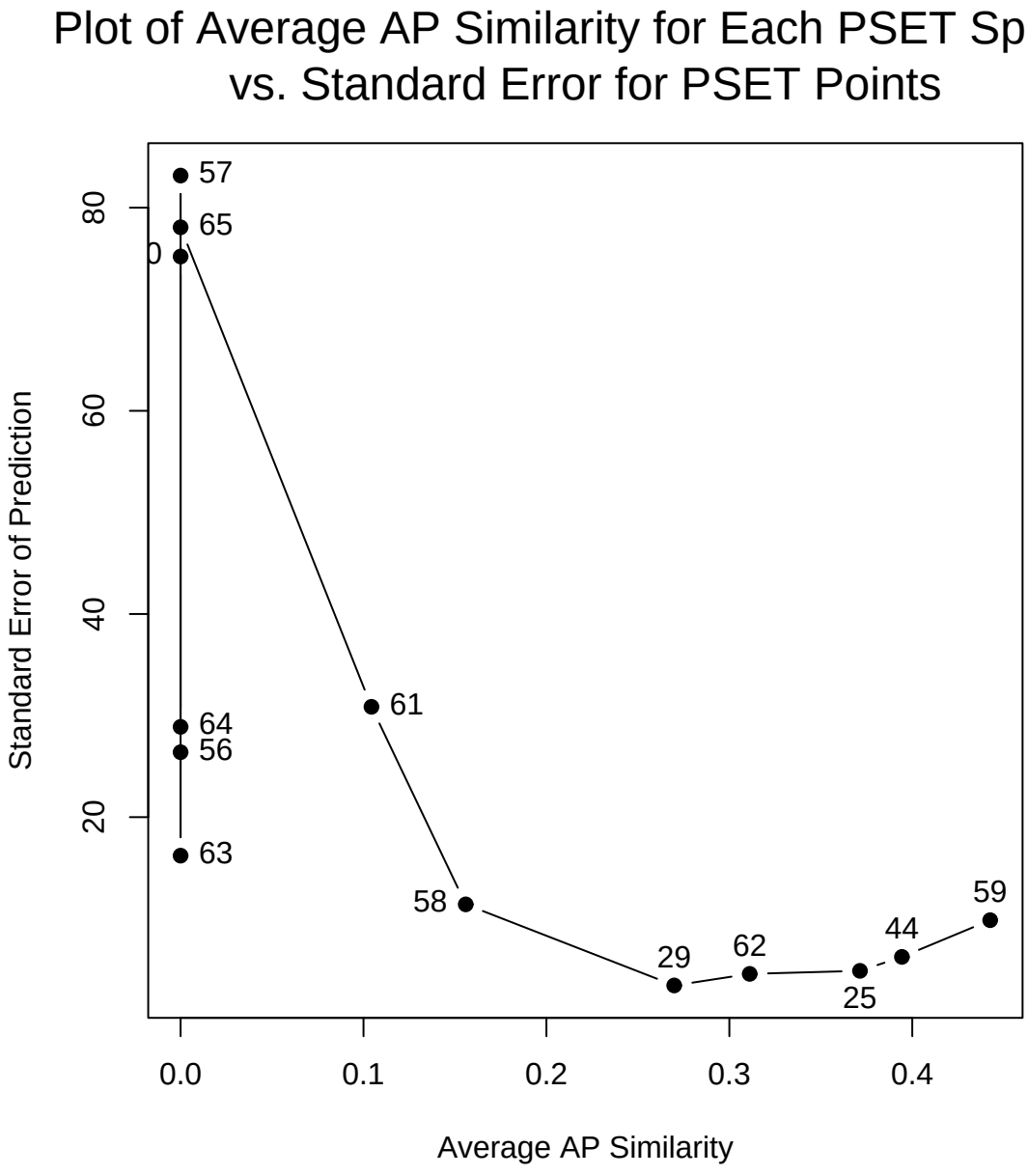


- The plot for decane shows that even though it has average TSET density the average APS value is still greater inside the sphere than outside
- The real test is whether the averaged APS values can correlate with residuals or SE's of predictions

Sphere / AP - Correlation to Residuals



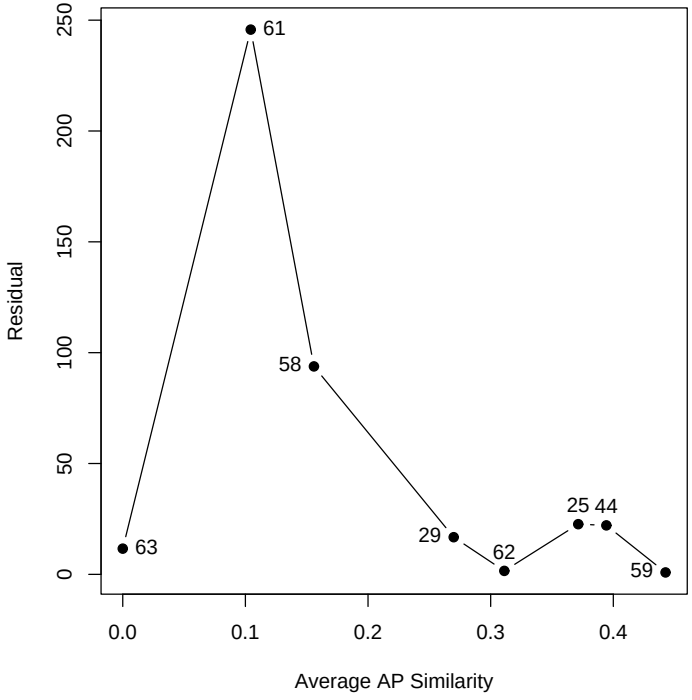
Sphere / AP - Correlation to SE of Prediction



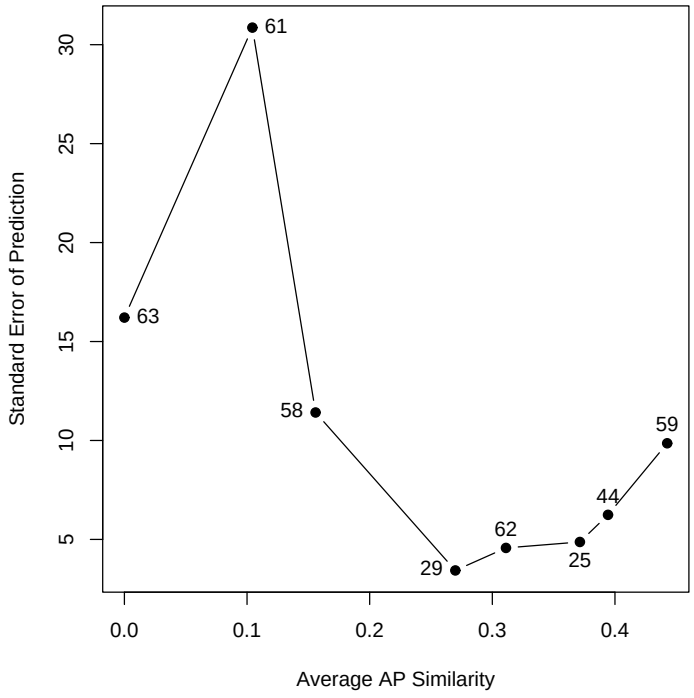
Sphere / AP - Removing Empty PSET Spheres

- Some of the PSET points had no TSET members around them
- Thus calculating average APS value was not possible and were set to zero
- However, it makes sense not to include those points

Plot of Average AP Similarity for Each PSET Sphere vs. Residuals for PSET Points

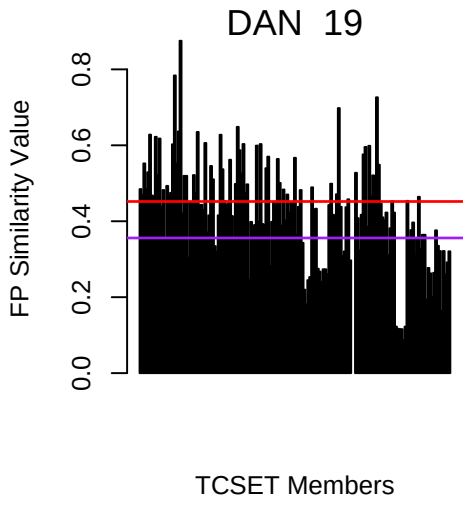


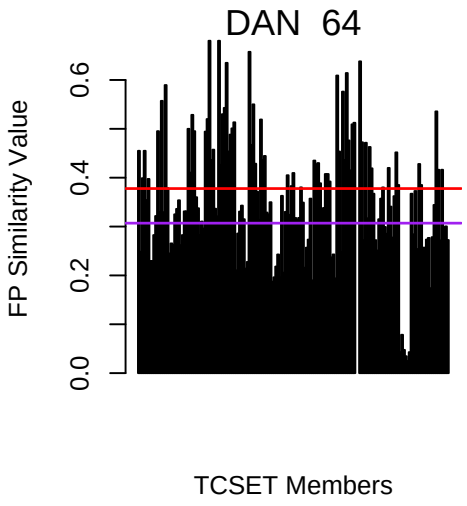
Plot of Average AP Similarity for Each PSET Sphere vs. Standard Error of Prediction for PSET Points

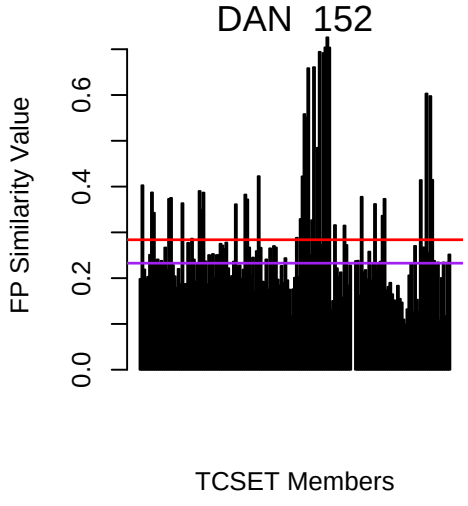


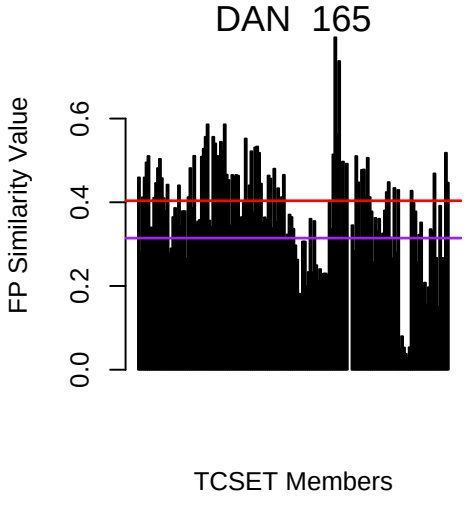
Sphere / AP - Artemisinin Dataset

Barplots of AP Similarity Values Between High Density PSET Points And TSET Members Inside And Outside Their Sphere



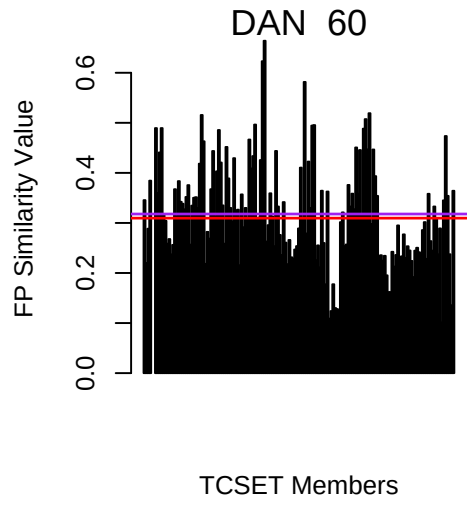


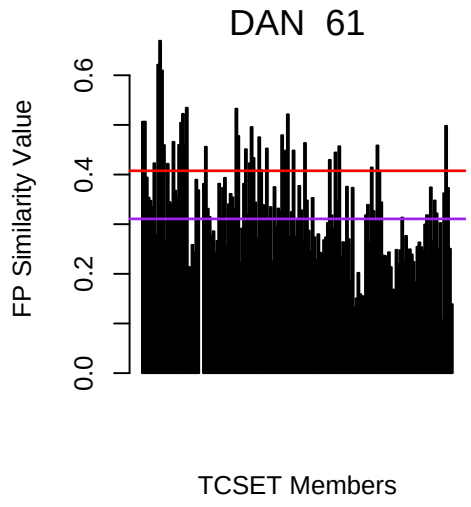


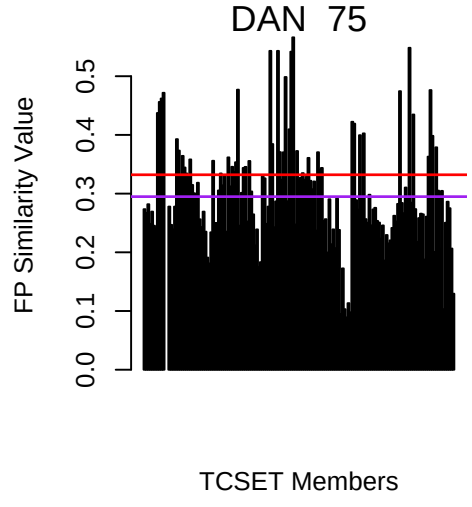


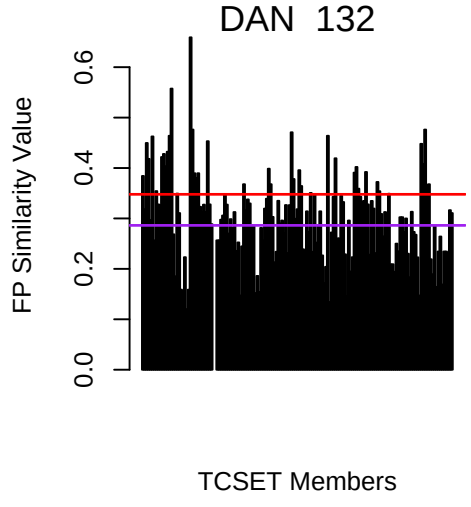
Sphere / AP - Artemisinin Dataset

Barplots of AP Similarity Values Between Low Density PSET Points And TSET Members Inside And Outside Their Sphere

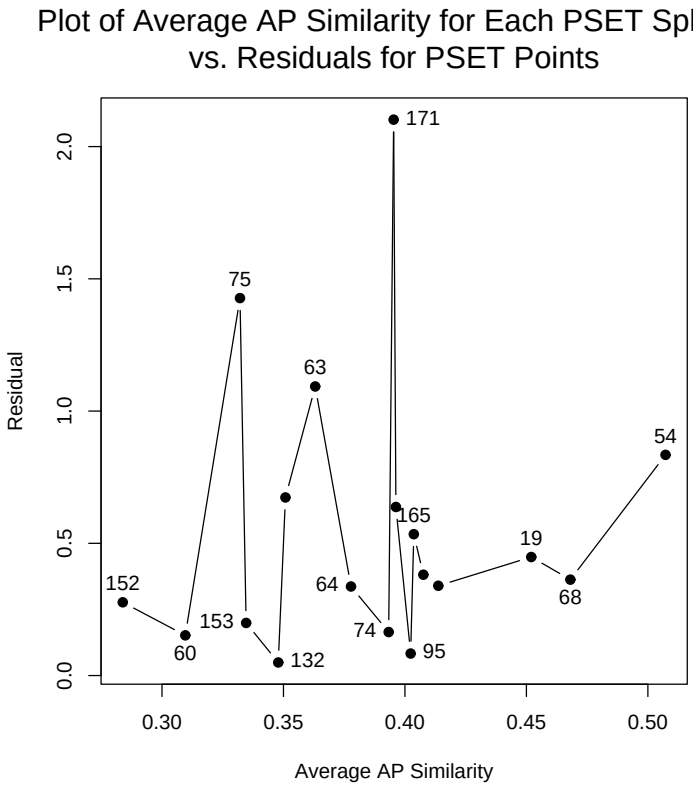


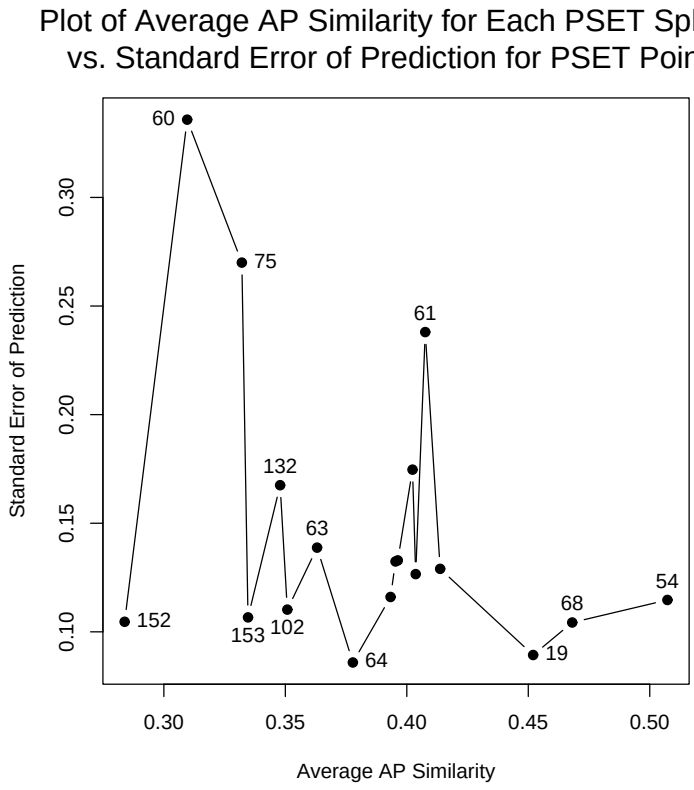






Sphere / AP - Artemisinin Residuals & SEoP





- The correlation to residuals is quite poor but there appears to be an inverse relation to the SEoP
- However, any trend in either plot is obscured by the several 'outliers'

Sphere / AP - Summary

- The APS values seem to be doing their job
- The problem with APS values is that for PSET points with empty spheres they are undefined
- The APS values appear to correlate better with SE of predictions rather than residuals
- In either case the trends are obscured
  - Could too many TSET points skew the average APS value for a PSET point?
- Overall, it seems that sphere densities appear to provide more direct & clear information about similarity and correlation to residuals/SEoP

Next Step

- See whether molecular fingerprints can help us
- Rather than consider all TSET points in the sphere, consider a subset - kNN style