Matching QSAR Sets

Using Non Parametric Statistics as QSAR Tools

Rajarshi Guha

Penn State University

What Are We Matching?

- The underlying goal is to see whether we can match PSET points to the TSET
- Ideally we would like to see whether a given PSET point is similar to the TSET in general
- Methods to achieve this include
  - Atom Pair Fragments
  - Daylight Fingerprints
  - 2D Holograms
  - SOM
  - *Statistical Methods?*

Statistical Methods

- The problem with statistical methods is that we cannot use single PSET points and make decisions
- These methods consider groups of points, i.e., distributions
- Thus these methods can decide whether 2 distributions are similar or whether a given distribution matches some assumed distribution with estimated parameters

Nonparametric Statistics - Overview

- Makes few assumptions about the model
- Essentially provides approximate probabilities to exact models
- Less computational work
- Ideally non parametric statistics are **distribution free**, but this is not always so.

Hypothesis Testing

- Hypotheses are stated in terms of the population
- A test statistic is selected
- A decision rule is created on the basis of the possible values of the statistic to decide whether to accept or reject the hypothesis
- The sample is used to calculate the test statistic and the decision rule is applied to accept or reject the hypothesis

χ^2 Goodness of Fit Test

- Data
  - Data consists of N independant observation
  - The data are binned into c classes
  - Each class has a frequency of $O_j,\ j = 1, 2, \ldots, c$
- Assumption
  - A random sample
  - Measurement scale is at least nominal
- Hypotheses:

$$
\begin{aligned}
H_0: &\quad F(x) = F^*(x) \quad \text{for all x} \\
H_1: &\quad F(x) \neq F^*(x) \quad \text{for at least one x}
\end{aligned}
$$

χ^2 Goodness of Fit Test

- Test Statistic
  - Assuming $F^*(x)$ is the distribution function, let p_j^* be the probability that a random observation falls in class j
  - Define E_j, the expected frequency of class j when H_0 is true as

$$E_j = p_j^* N, \; j = 1, 2, \ldots, c$$

  - The statistic T is given by

$$T = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}$$

χ^2 Goodness of Fit Test

- Decision Rule
  - The approximate distribution of T for large samples is the χ^2 distribution
  - Critical region corresponds to values of T greater than $x_{(1-\alpha)}$, where α is the level of significance.
  - The d.o.f is given by $c - k + 1$, where c is the number of non empty bins and k is the number of estimated parameters
  - Reject H_0 if $T > x_{(1-\alpha)}$

χ^2 Goodness of Fit Test

- Some features include
  - The statistic depends on the nature of binning
  - If a class frequency is less than 5 the class should be combined with an adjacent class
  - The test justifies the use of $F^*(x)$ as a good approximation to the true distribution by accepting H_0
  - Essentially it assumes a distribution for a set A and then indicates whether a set B matches that distribution

Kolmogorov - Smirnov Statistics

- This class of test statistics can check
  - whether a sample fits a certain distribution
  - whether two or more samples have similar distributions
- Though similar in intent to the χ^2 test, this class of statistics have higher **power**
- Example statistics include
  - Kolmogorov Goodness of Fit
  - Shapiro Wilk Test for Normality
  - Smirnov Test
  - Cramer von Mises Two Sample Test

Smirnov Test

- Also termed as the Kolmogorov Smirnov Two Sample Test
- Determines whether two samples have the same population distribution function
- Consistent against all types of differences between the two distribution functions

Smirnov Test

- Data
  - Two independant random samples of size n and m denoted by $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$
  - Unknown distribution functions denoted by $F(x)$ and $G(x)$
- Assumptions
  - Random samples
  - Independant samples
  - Measurement scale is ordinal
  - For the test to be exact the random variables should be continuous

Smirnov Test

- Hypotheses

$$\begin{aligned} H_0: &\quad F(x) = G(x) \quad -\infty < x < \infty \\ H_1: &\quad F(x) \neq G(x) \quad \text{for at least one x} \end{aligned}$$

- Test Statistic

$$T = \max_x |S_1(x) - S_2(x)|$$

where S_1 & S_2 are the empirical distribution functions

Smirnov Test

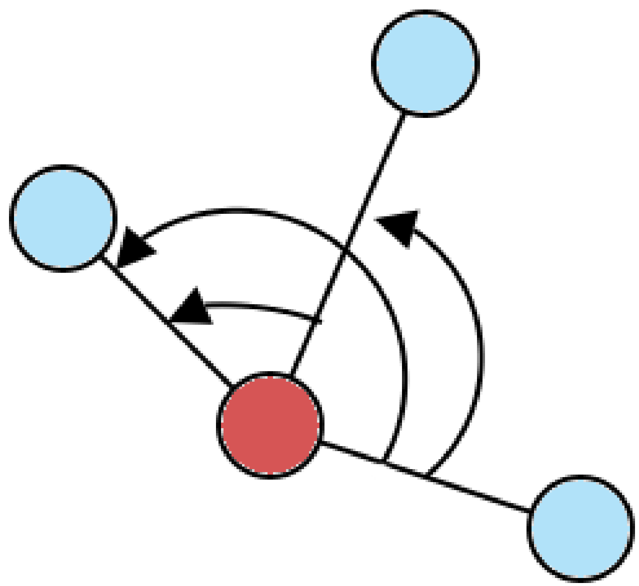
- Decision Rule
  - Reject H_0 at level of significance α if $T > q_{(m,n)}$.
  - Depending on whether m equals n and the level of significance, q can be evaluated from different large sample approximations.

Applying the Statistics

- Select a dataset.
- Perform a kNN calculation on the TSET and PSET.
- Rather than look at predicted values, look at **kNN distances** & **angles.**
- Investigate the statistics of the distances and angles of the TSET & PSET for a given dataset.
- Attempt to link the statistics to model performance

kNN Distance & Angles

- For each molecule in a given set, the sums and average of the distances to the n nearest neighbors were recorded.
- Sums and the average of the angles were also recorded.
- For angles, n was restricted to 3 - simplifies the number of angles to evaluate.



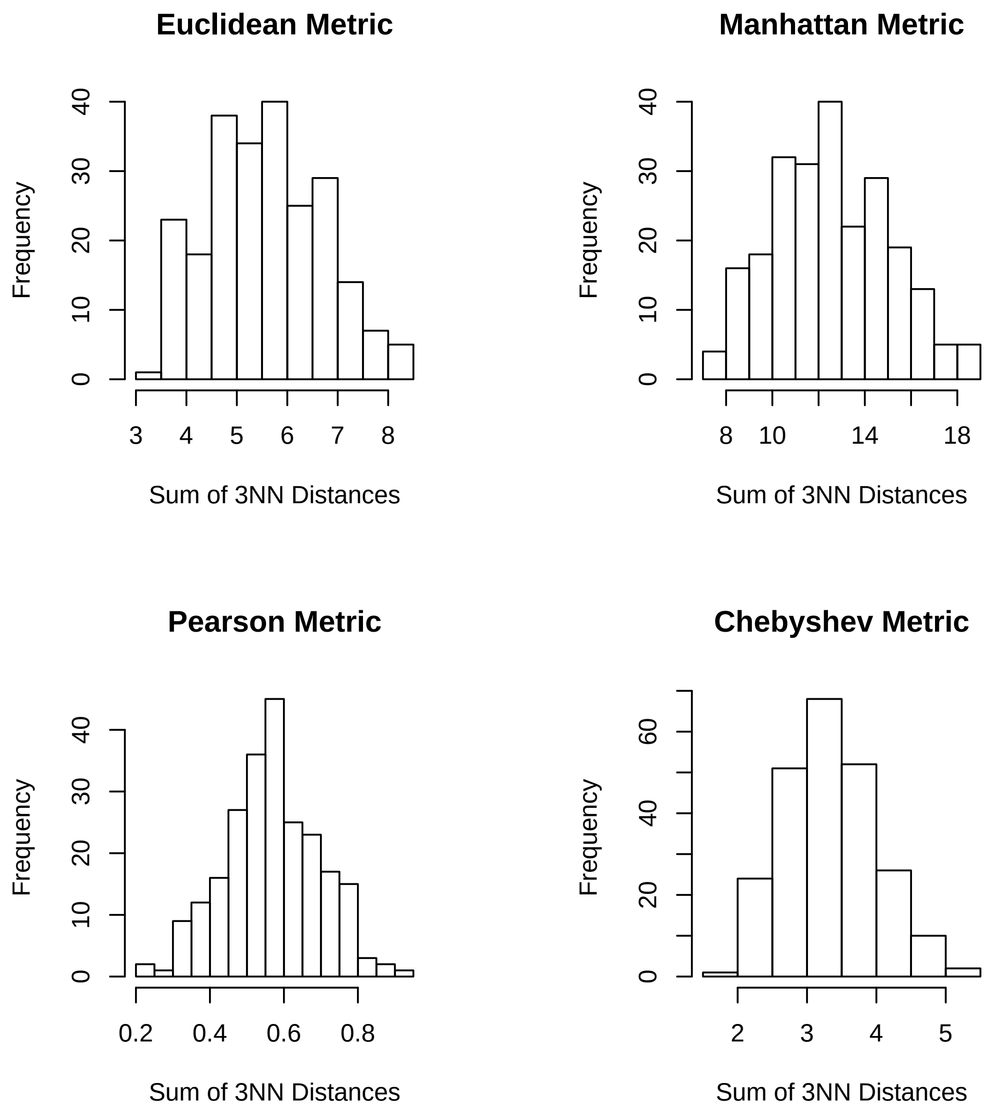
Results

Random Data

- A set of descriptor values and dependant variable values were randomly generated for 233 molecules
- A Gaussian distribution ($\sigma = 1.0, \mu = 0.0$) was used
- A descriptor length of 8 was used
- TSET, CVSET, PSET were generated by *setbin.py*
- Multiple 3NN runs were carried out with varying distance metrics

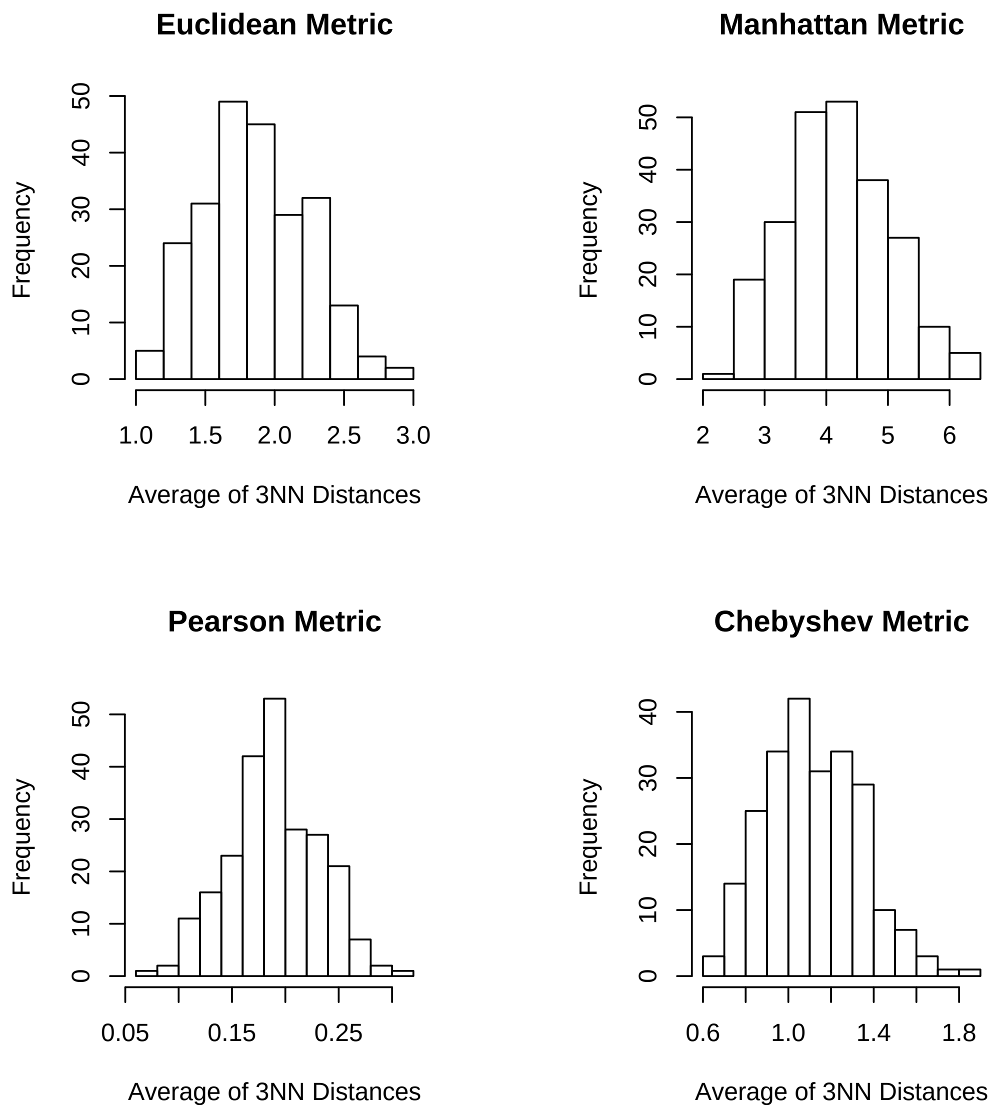
Random Data - Sums of Distances

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

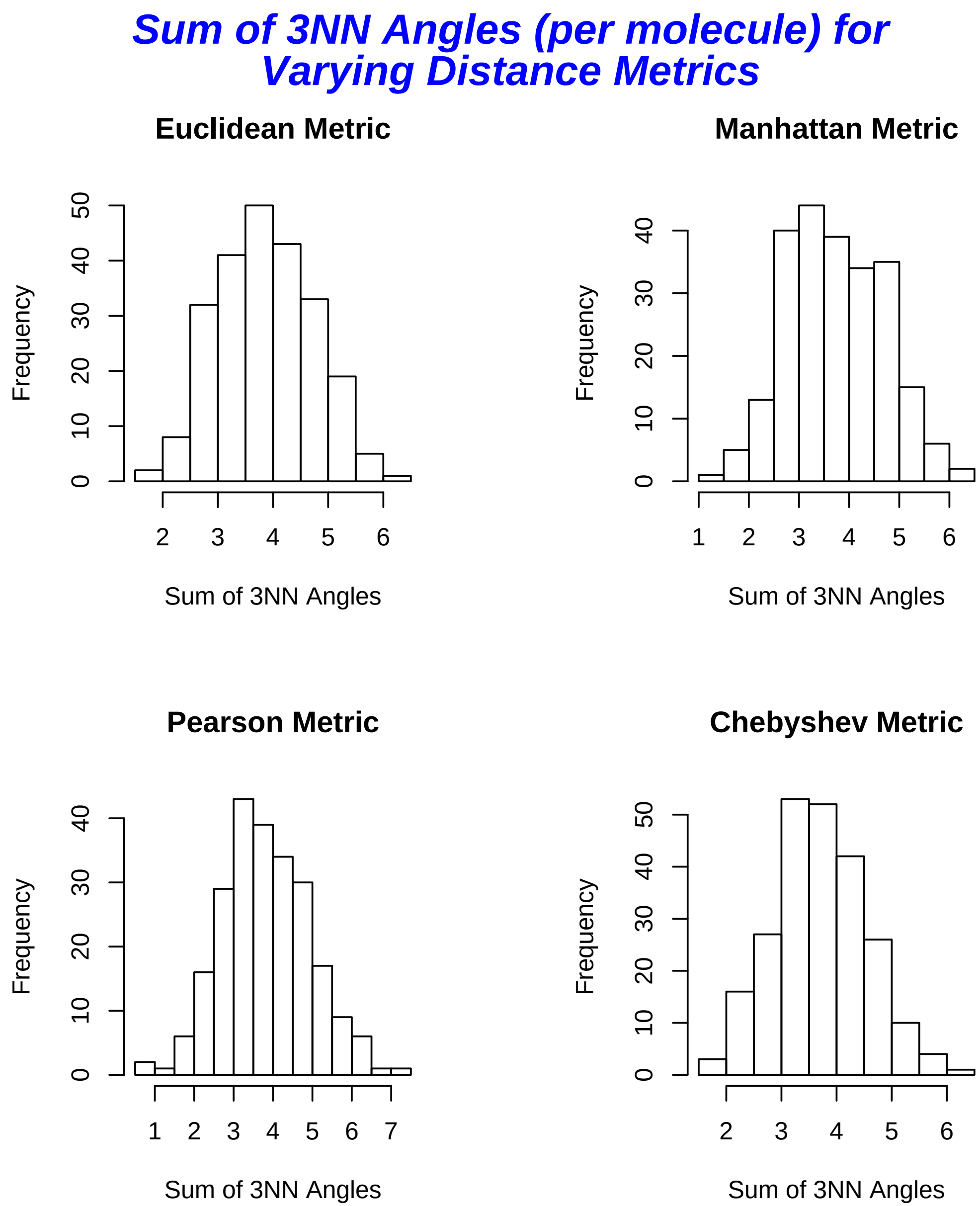


Random Data - Averages of Distances

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

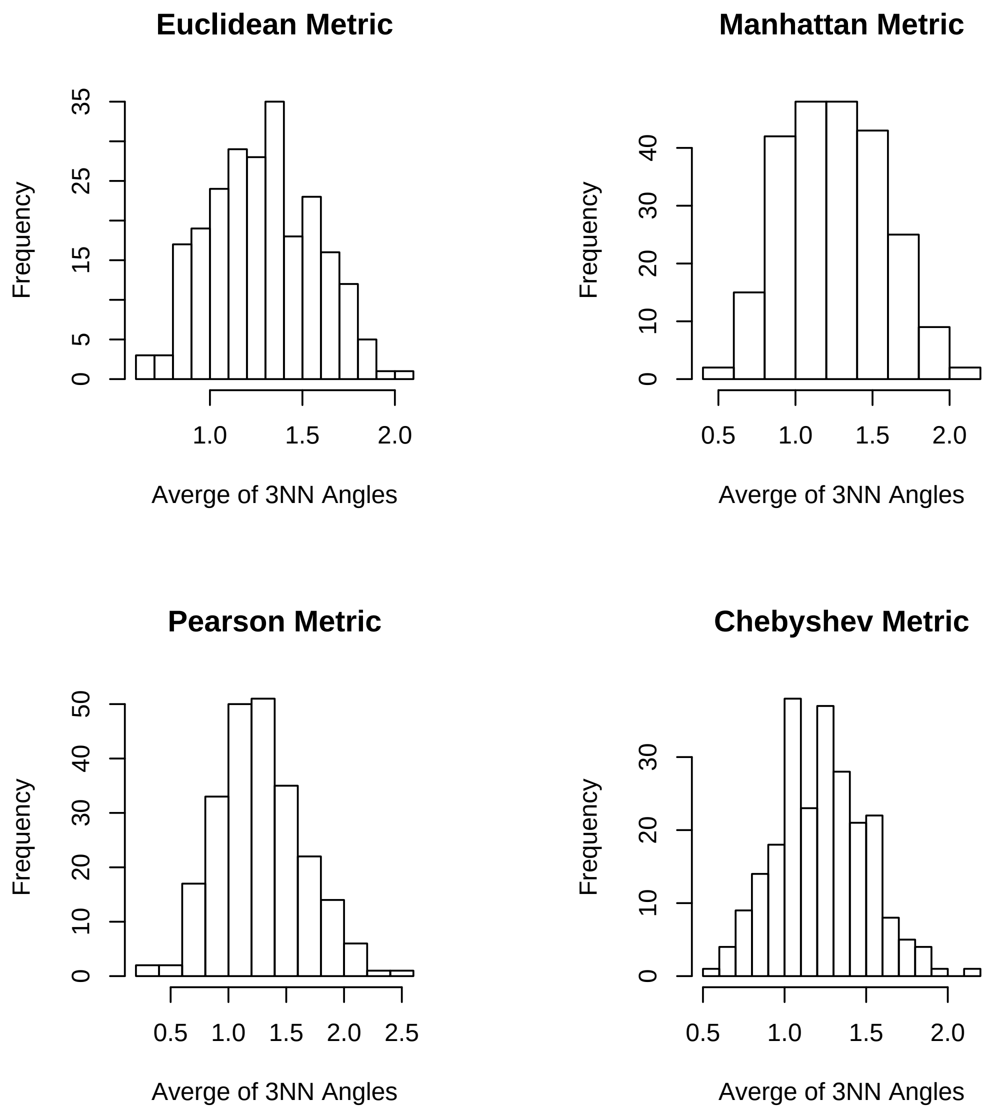


Random Data - Sums of Angles



Random Data - Averages of Angles

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

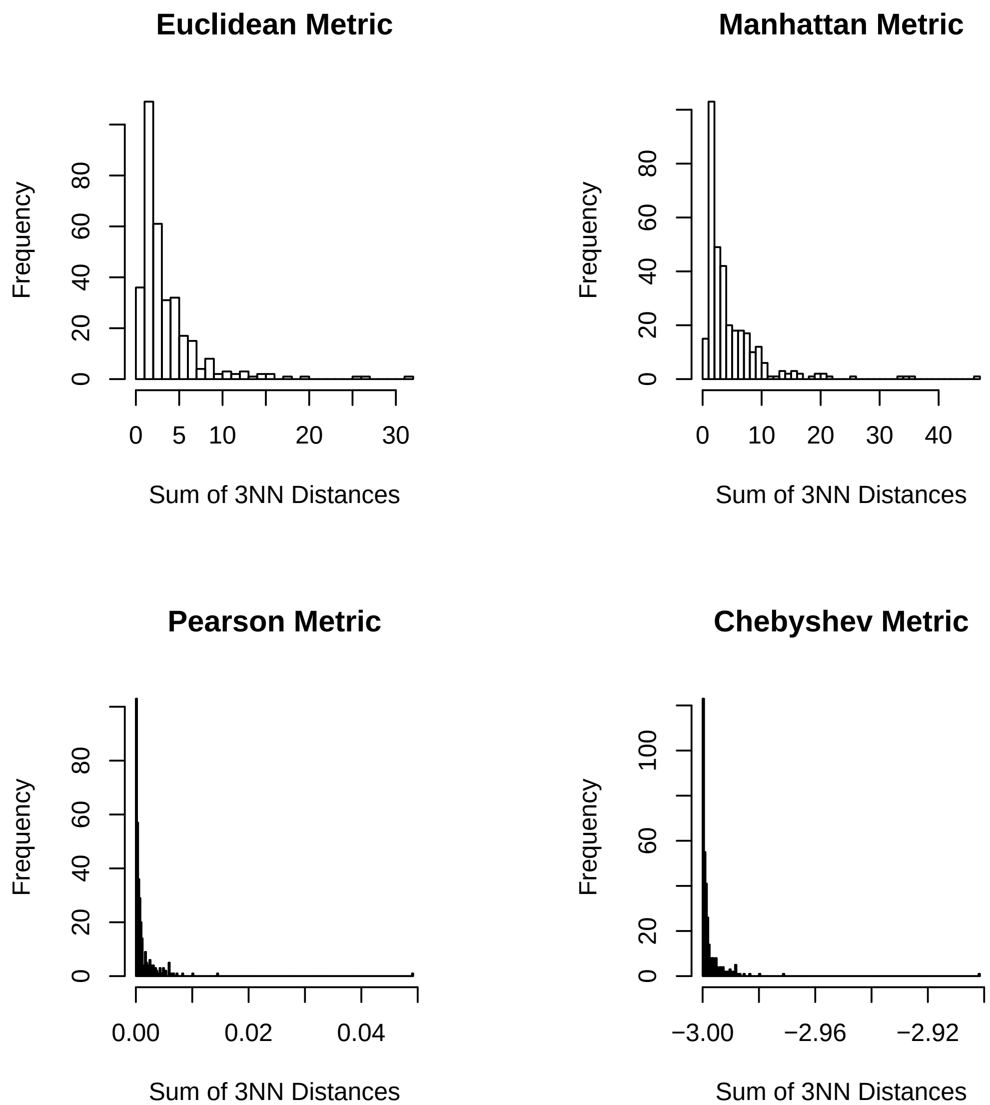


DHFR: BCUT - 2D Auto.

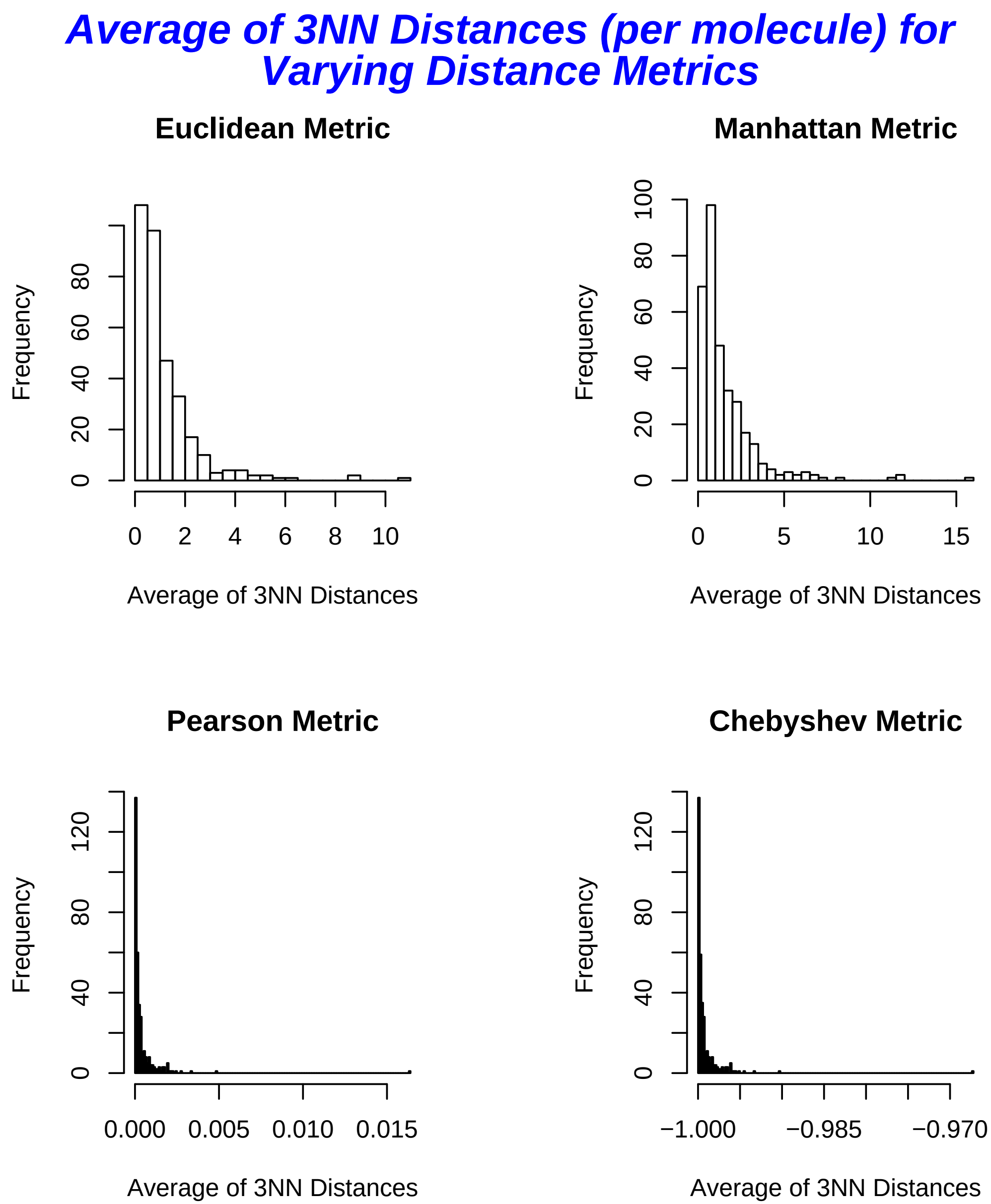
- Since a number of models exist the descriptors for the best model were chosen
- 5 descriptors chosen: N5CH, N7CH, NAB, WPSA, CHAA
- The model R^2 (TSET) was 0.45
- All the molecules were utilitzed by the kNN routine

DHFR - Sums of Distances

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

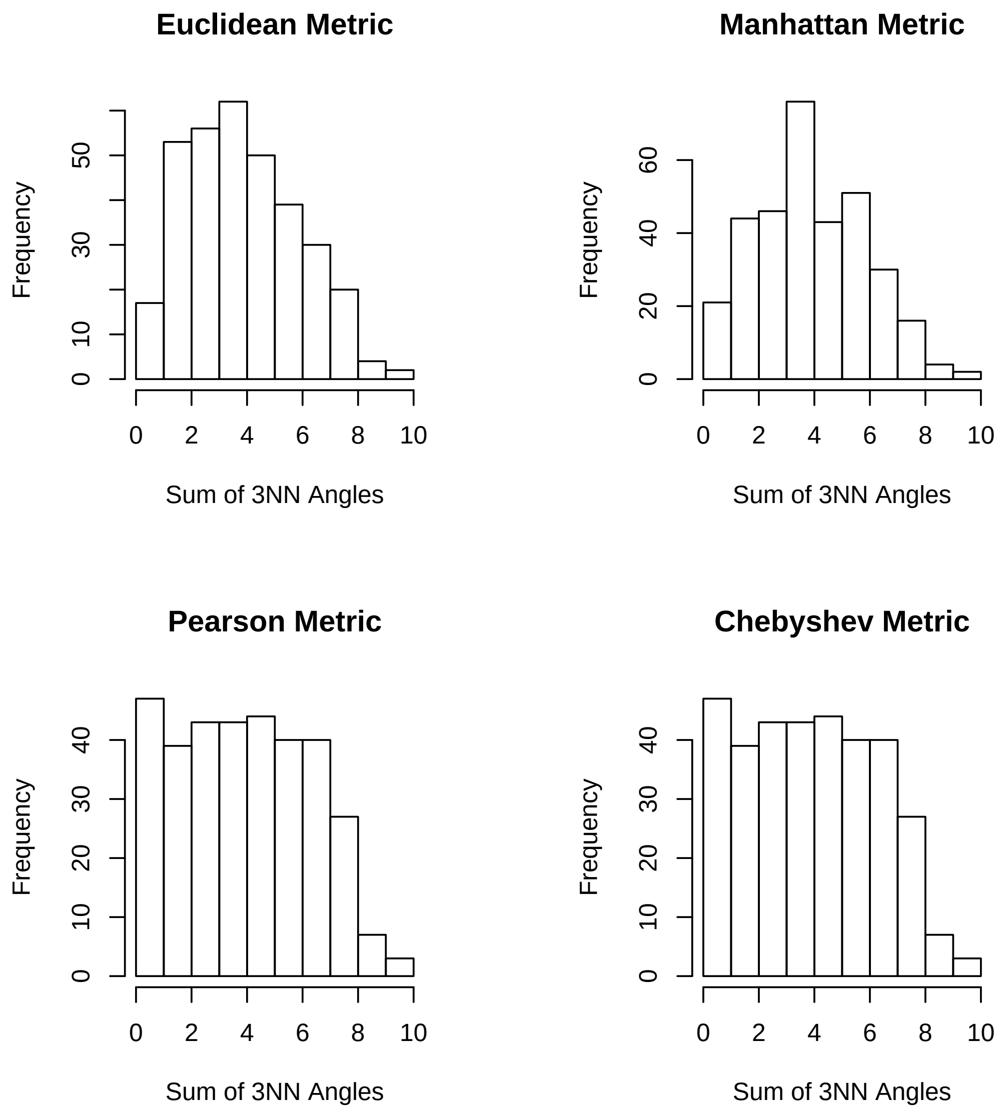


DHFR - Averages of Distances

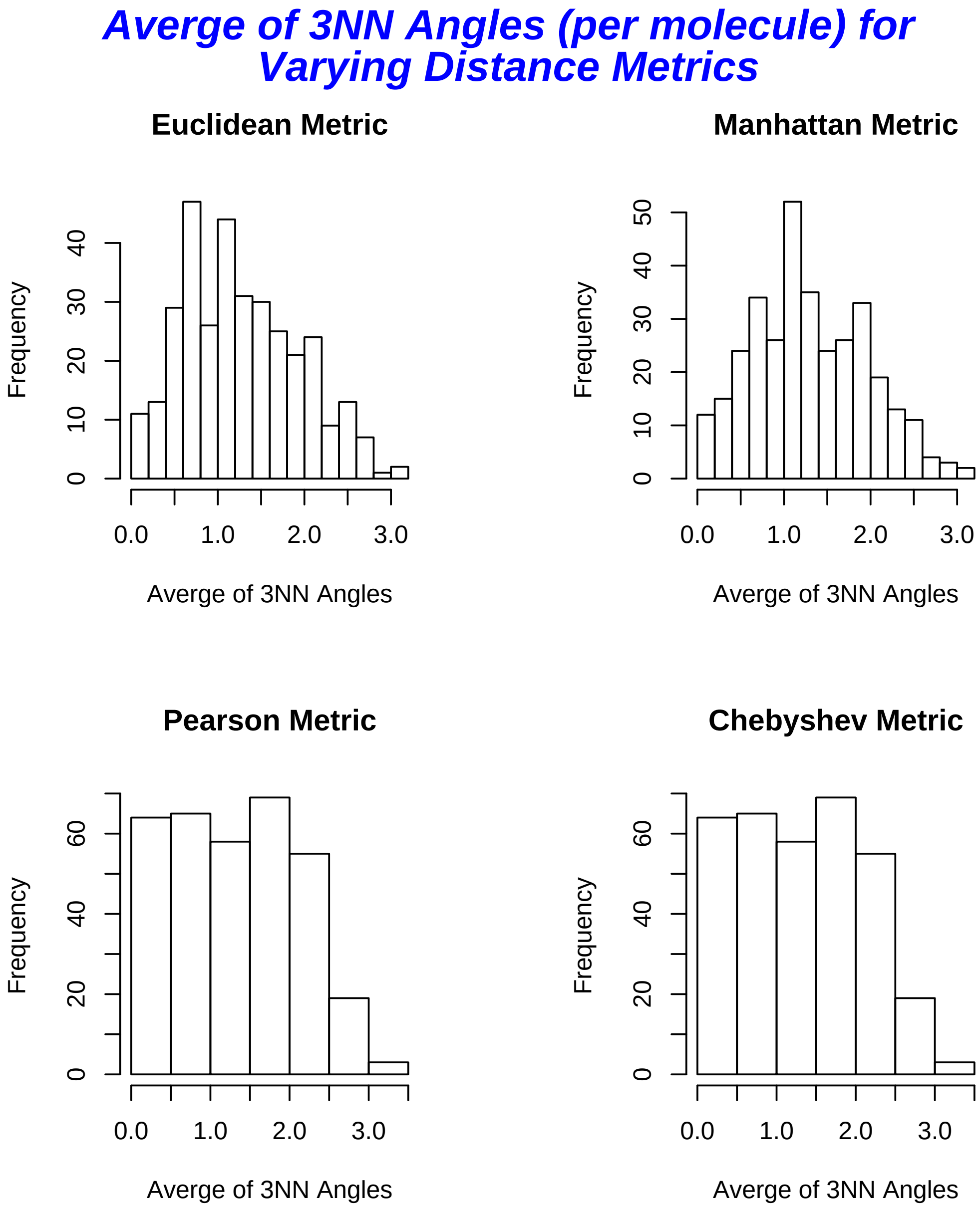


DHFR - Sums of Angles

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



DHFR - Averages of Angles

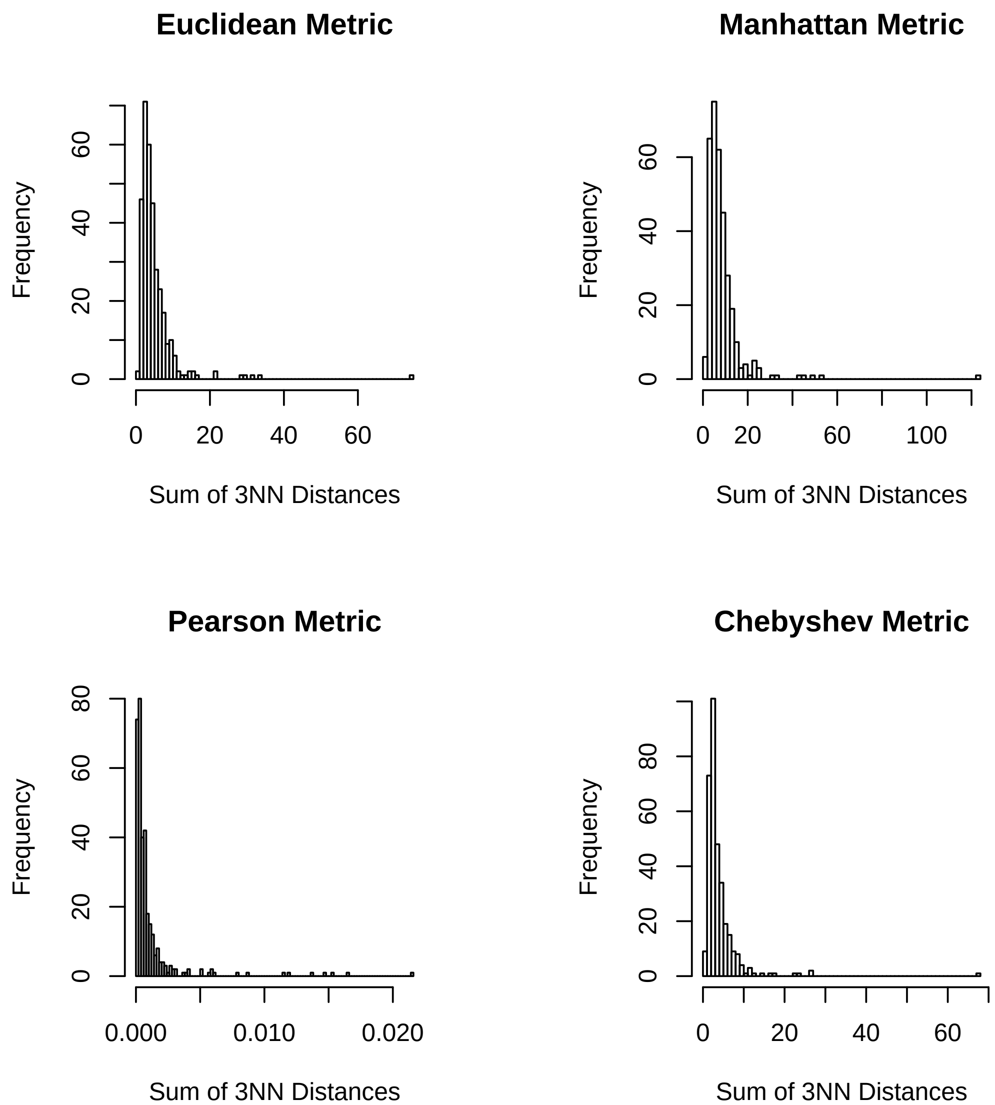


DHFR: GETAWAY

- Since a number of models exist the descriptors for the best model were chosen
- 10 descriptors chosen: N5CH, <OLC, NDB, WTPT, PND, elec, WNSA, CHAA2, CHAA3, SCAA
- The model R^2 (TSET) was 0.53
- All the molecules were utilitzed by the kNN routine

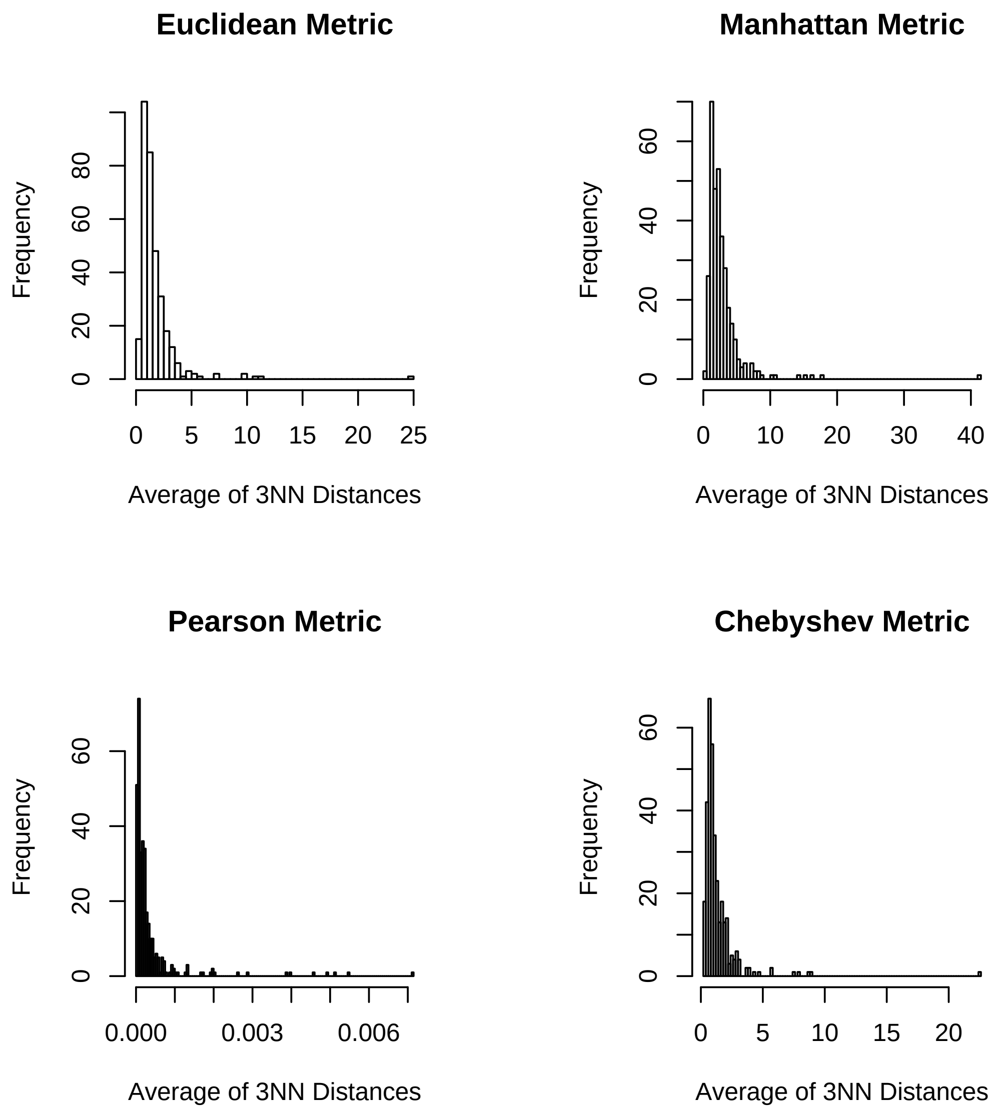
DHFR - Sums of Distances

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



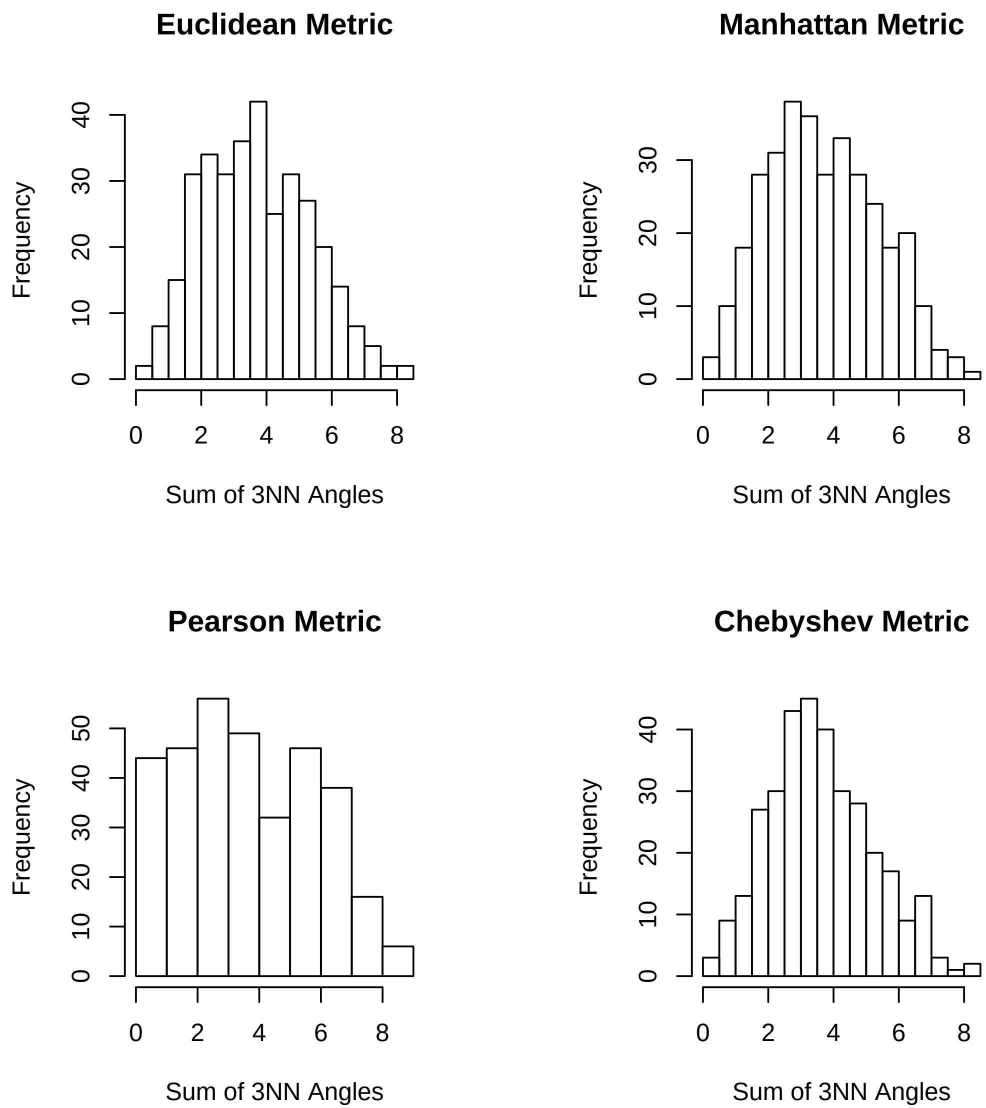
DHFR - Averages of Distances

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

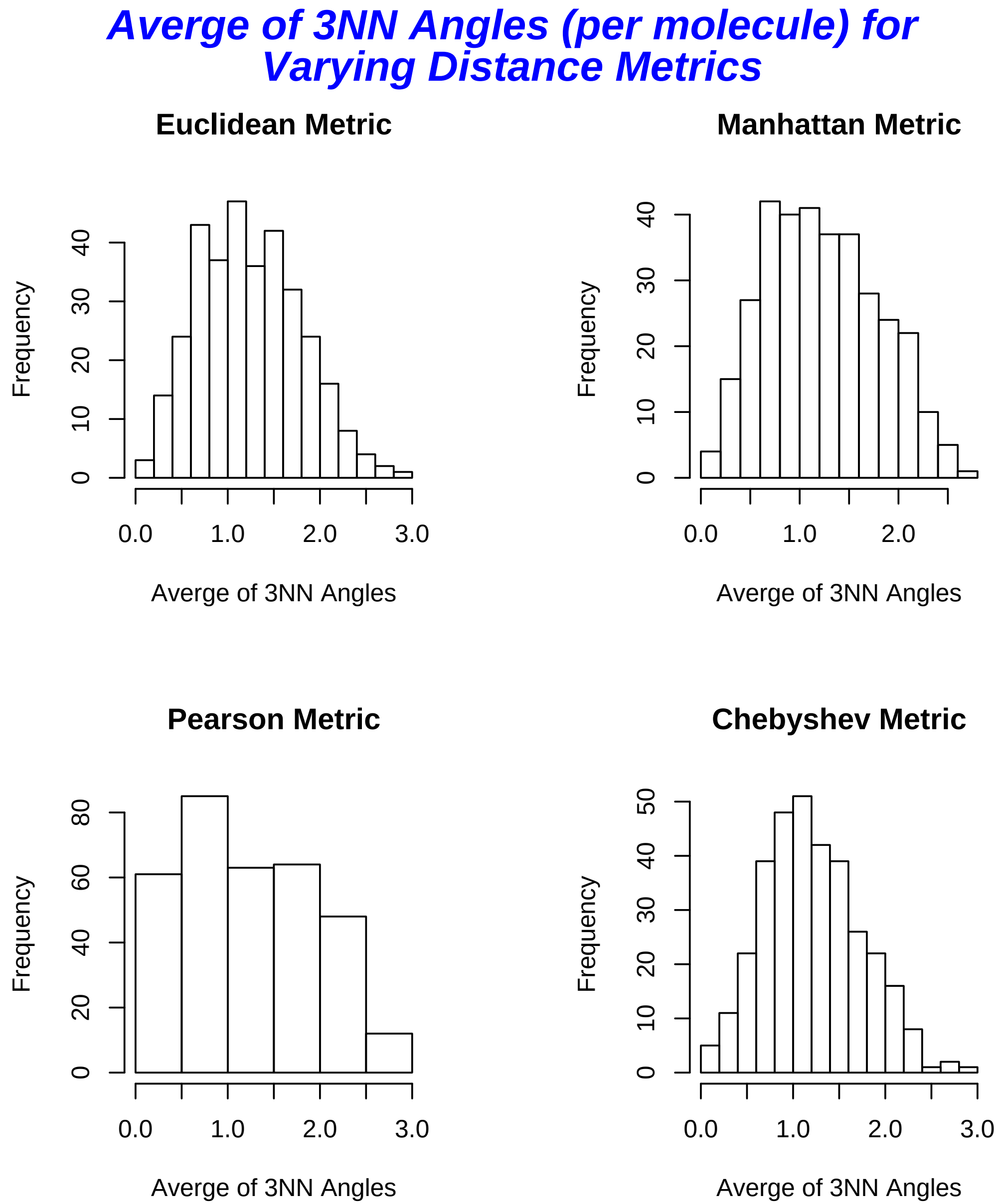


DHFR - Sums of Angles

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



DHFR - Averages of Angles



Smirnov Test Results - Random Data

- Dataset was Random and sums of 3NN distances were used

Metric	D(198,36)	Q(0.95)	H_0
Euclidean	0.6288	0.2464	Reject
Manhattan	0.5657	0.2464	Reject
Pearson	0.7652	0.2464	Reject
Chebyshev	0.6263	0.2464	Reject

- Clearly the original distribution of the descriptors in the two sets need not be carried over into subsequent calculations

Smirnov Test Results - Random Data

- Dataset was Random and sums of 3NN angles were used

Metric	D(198,36)	Q(0.95)	H_0
Euclidean	0.2399	0.2464	Accept
Manhattan	0.1717	0.2464	Accept
Pearson	0.2247	0.2464	Accept
Chebyshev	0.3056	0.2464	Reject

- Thus characterization of the distribution depends on which variable we are looking at (distance or angles) as well as type of metric used

Smirnov Test Results - DHFR Data

- Dataset was DHFR - BCUT/Auto and sums of 3NN distances were used

Metric	D(299,34)	Q(0.95)	H_0
Euclidean	0.6087	0.2461	Reject
Manhattan	0.6154	0.2461	Reject
Pearson	0.7458	0.2461	Reject
Chebyshev	0.6087	0.2461	Reject

Smirnov Test Results - DHFR Data

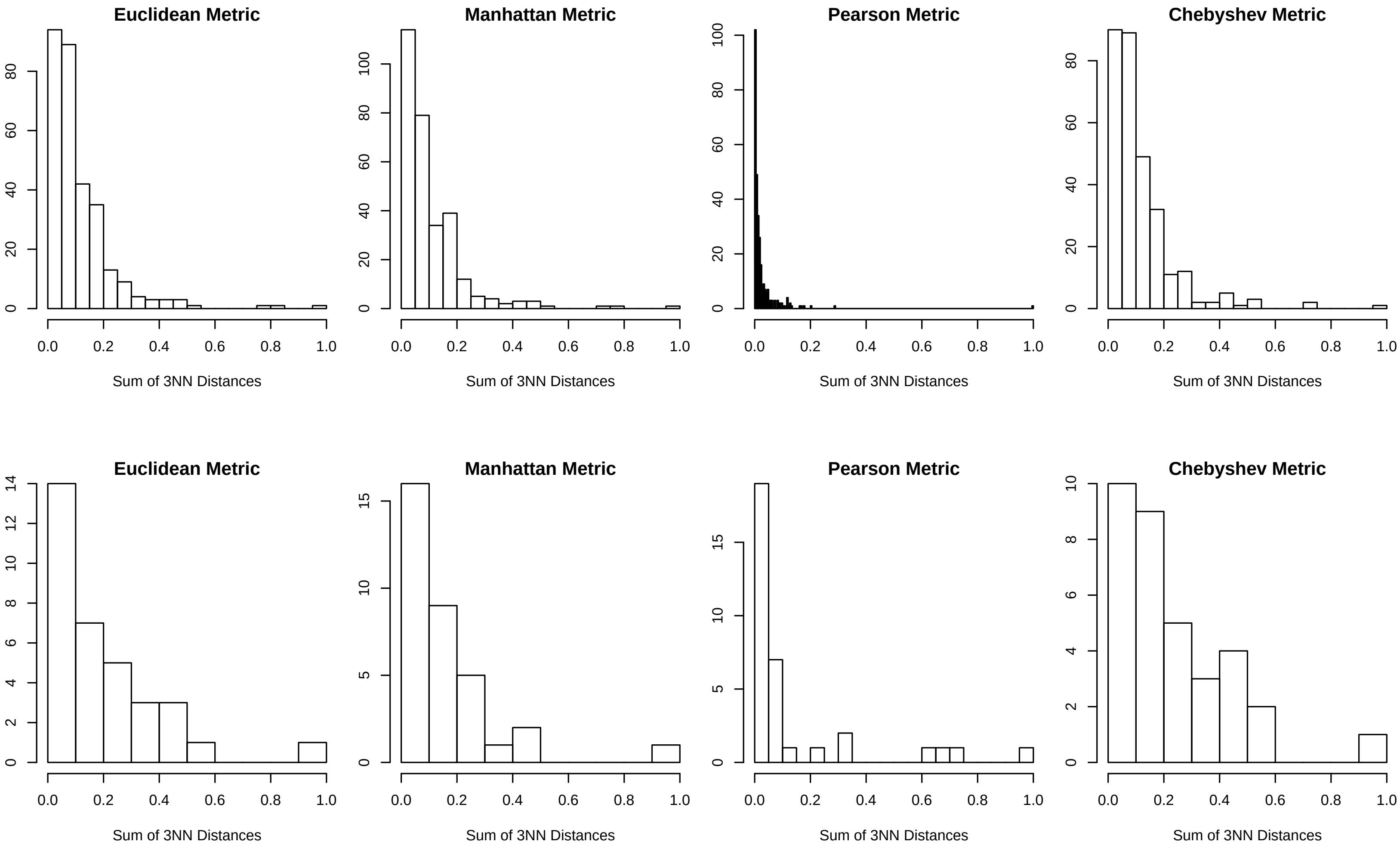
- Dataset was DHFR - BCUT/Auto and sums of 3NN angles were used

Metric	D(299,34)	Q(0.95)	H_0
Euclidean	0.1951	0.2461	Accept
Manhattan	0.1314	0.2461	Accept
Pearson	0.0913	0.2461	Accept
Chebyshev	0.1815	0.2461	Accept

DHFR - BCUT/Auto - Setwise Histograms

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

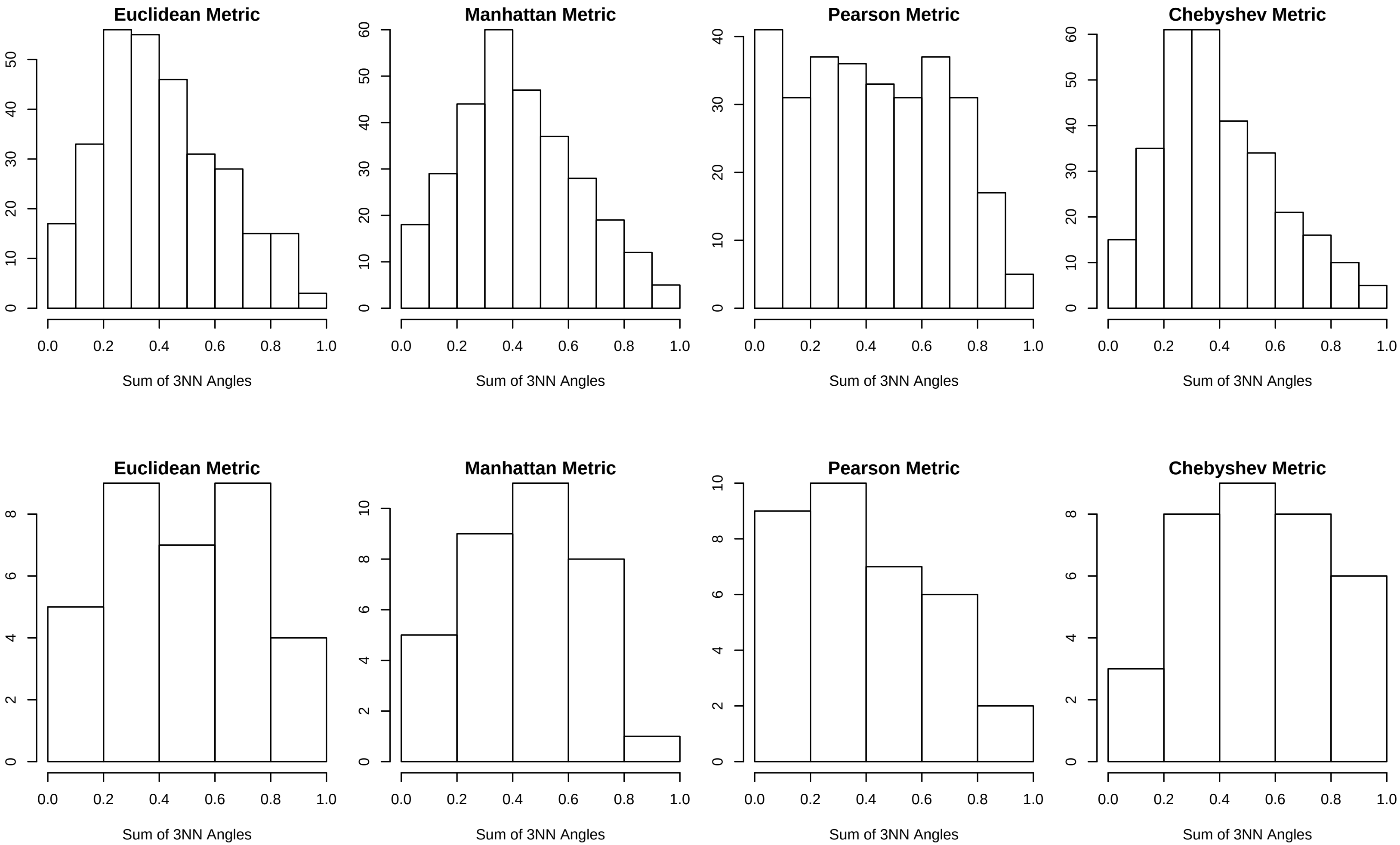
(Upper row is TSET and bottom row is PSET)



DHFR - BCUT/Auto - Setwise Histograms

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

(Upper row is TSET and bottom row is PSET)



Smirnov Test Results - Tutorial Data

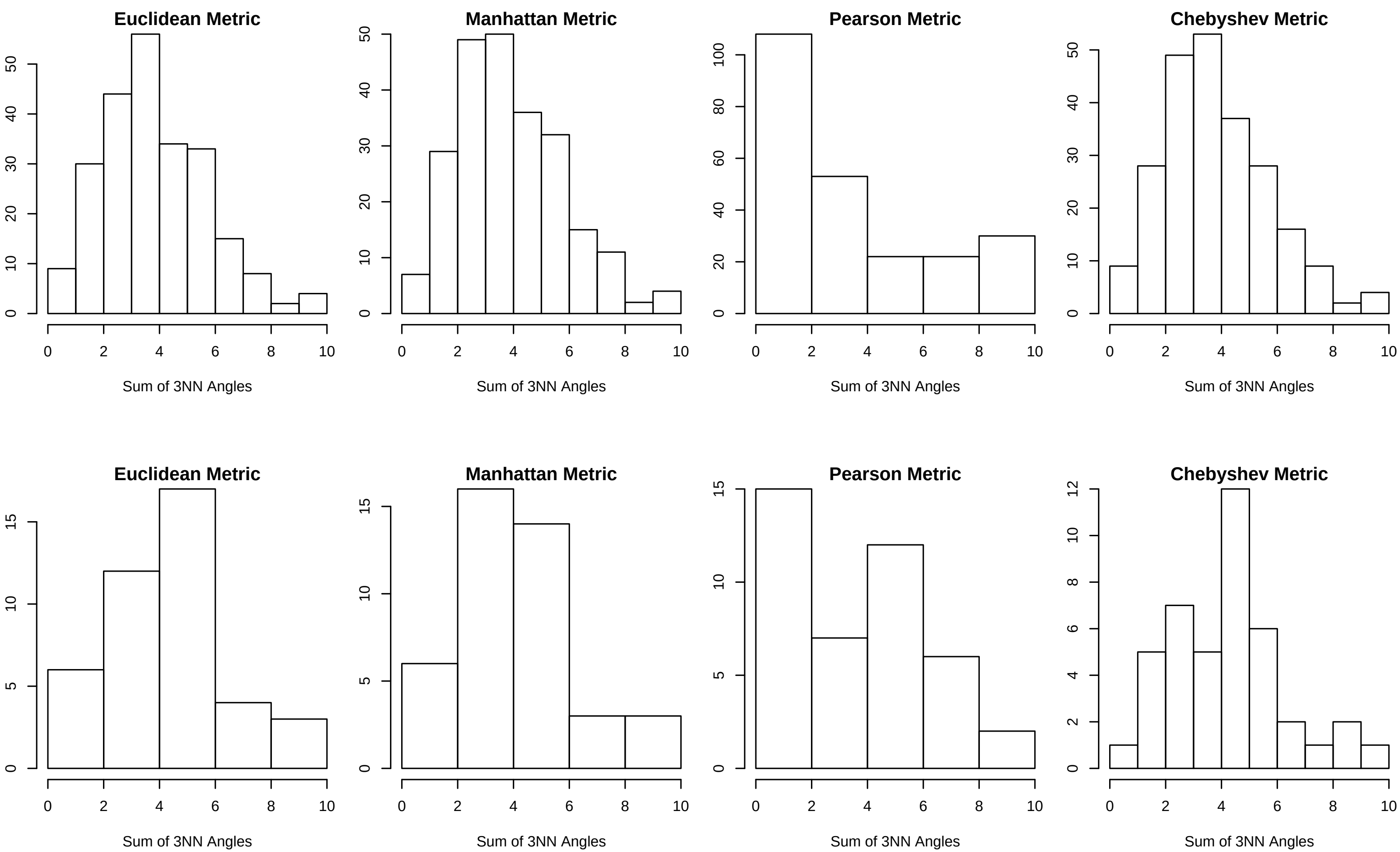
- Dataset was taken from the tutorial and sums of 3NN angles were used

Metric	D(235,42)	Q(0.95)	H_0
Euclidean	0.1629	0.2278	Accept
Manhattan	0.0892	0.2278	Accept
Pearson	0.1900	0.2278	Accept
Chebyshev	0.1672	0.2278	Accept

Tutorial - Setwise Histograms

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

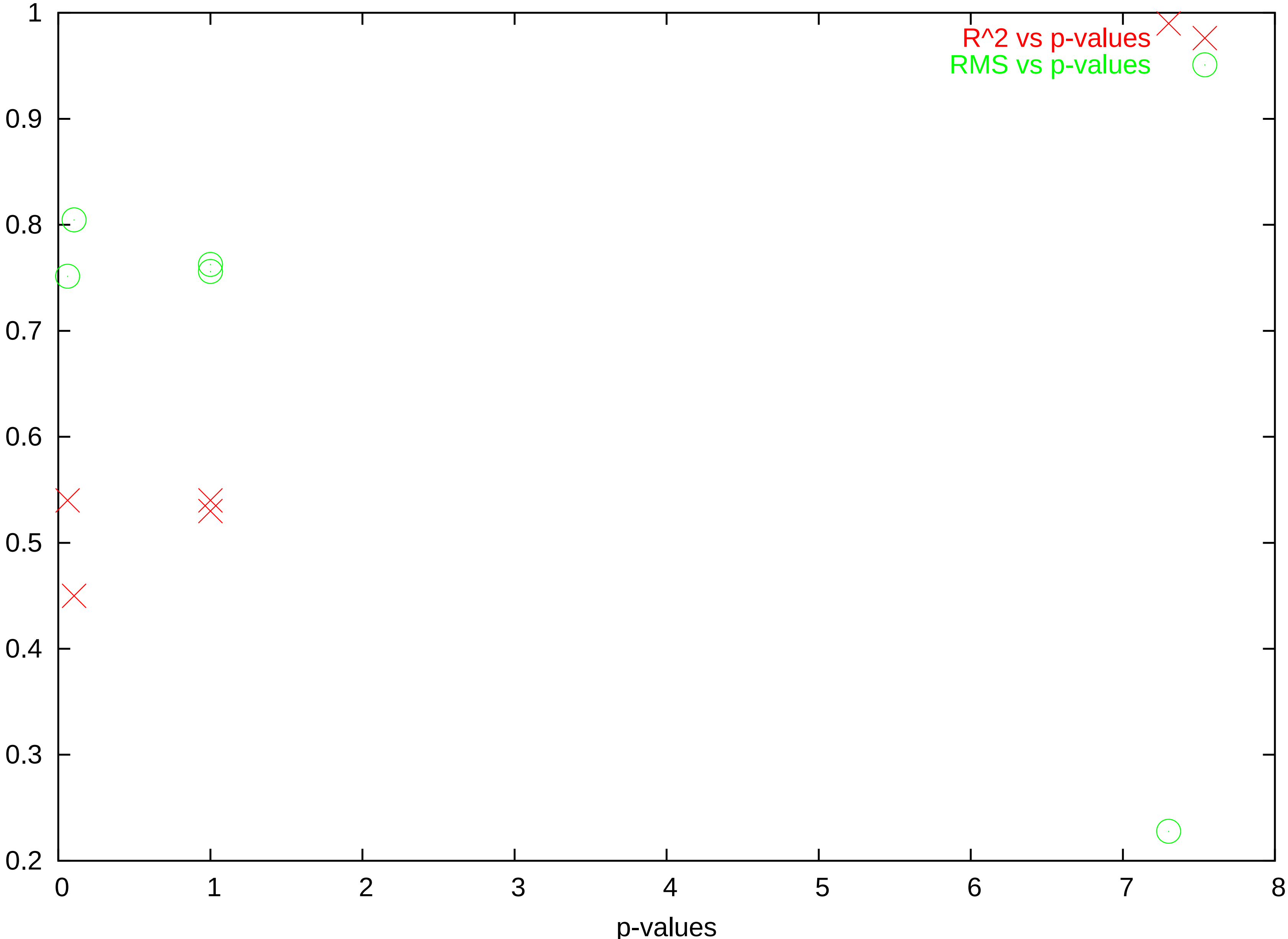
(Upper row is TSET and bottom row is PSET)



Correlating Distributions to Models

- We want to correlate the distribution statistics to the model performance
  - The value of the Smirnov test depends on the length of the two distributions.
  - Similarly for the 0.95 quantile value
  - The best distribution statistic to use is probably the p-value of the Smirnov test
  - Model features to correlate to include R^2 and RMSE
  - Another possible model feature(s) that might be correlated with are properties of the residuals such as distribution

Correlating Distributions to Models



Observations

- There does'nt seem to be much of a difference between sums and averages.
- It appears that angles are more evenly distributed than distances.
- The angle distribution appears to be more *normal* than the distance distributions
- It appears that sets having a more *normal* distribution work better in the KS test
- This statistical approach does'nt seem to help us reach the goal of looking at single points :(

General Plan

- Statistical measures would only be useful if we consider groups of PSET points rather than individual points
- Evaluate a similarity measure, A
  - Atom Pair
  - SOM
- Use A to calculate similarity between PSET point(s) & TSET
  - Reduce PSET - TSET similarities to one value?
  - Utilize the mutiple PSET - TSET similarity values?
- Link A to the performance of the model
  - Look at the trend of residual vs similarity value

Extra Information

χ^2 Goodness of Fit

- Calculation of expected value for a class

$$E_j = N\left[F(Y_u) - F(Y_l)\right]$$

  - F is the cumulative distribution function for the distribibution being tested
  - Y_u & Y_l are the upper and lower limits of the j'th class
  - N is the sample size