ABSTRACT

SINGULAR VALUE DECOMPOSITION OF ANALOGS OF GBR 12909

by
Anna Fiorentino

Analogs of GBR 12909 are drugs that could potentially be used to treat cocaine addiction. Singular Value Decomposition (SVD) is a multivariate analysis technique used to show relationships between the data and the variables associated with the data. The input data consists of the conformers of each analog (DM324, 728 conformers; TP250, 739 conformers) along with the eight torsional angles (A1, A2, B1-B6). A novel scaling technique was developed to address the problem of data circularity by subtracting the values of the torsional angles of the global energy minimum conformation from those of each conformer.

In SVD the original data matrix **X** of dimensions *r x c* is decomposed into three matrices, **U, S,** and **V** where $\mathbf{X}=\mathbf{USV}^{\mathbf{T}}$. The columns of **U** represent the principal component (PC) scores. The rows of $\mathbf{SV}^{\mathbf{T}}$ contain the PC loadings. Analysis of the score and loading plots shows that DM324 separates into three distinct groups along PC1 due to A1 and six groups due to A2. TP250 separates into three groups along PC7 (due to B4) and three groups along PC8 (due to B3) resulting in nine clusters. The significance of this work is that it is the first application of SVD to the clustering of very flexible molecules. In the future, representative conformations of these analogs will be used in pharmacophore modeling with the ultimate goal of designing a drug useful in the treatment of cocaine abuse.

SINGULAR VALUE DECOMPOSITION OF ANALOGS OF GBR 12909

by
Anna Fiorentino

A Master's Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Department of Computer Science

August 2004

APPROVAL PAGE

SINGULAR VALUE DECOMPOSITION OF ANALOGS OF GBR 12909

Anna Fiorentino

Dr. Carol A. Venanzi, Thesis Advisor — Date
Distinguished Professor of Chemistry, NJIT

Dr. Michael L. Recce, Committee Member — Date
Associate Professor of Computer and Information Science, NJIT

Dr. Rose Dios, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

BIOGRAPHICAL SKETCH

Author: Anna Fiorentino

Degree: Master of Science

Date: August 2004

Undergraduate and Graduate Education:

- Master of Science in Computational Biology
  New Jersey Institute of Technology, Newark, NJ, 2004

- Bachelor of Science in Biology/Bioinformatics
  College of Staten Island, Staten Island, NY, 2003

Major: Computational Biology

To my parents without whom I would not have a dedication to write

A i miei genitori sensa i qually non potevo fare questa dedica

ACKNOWLEDGMENT

I would like to express my gratitude to Dr. Carol A. Venanzi, for serving as my research supervisor. Special thanks are given to Dr. Michael Recce and Dr. Rose Dios for actively participating in my committee and assisting me in completing my thesis.

I would like to thank everyone in Dr. Venanzi's research lab for their constant help and insight into this project. I wish to thank Deepangi Pandit, Kathleen Gilbert, and Milind Misra of the Department of Chemistry and Environmental Science for all of the data and support they provided me. Additional thanks to the Computational Biology program for their encouragement.

A special thanks is given to Dr. Ron Wehrens of the Department of Chemometrics at the University of Nijmegen for his technical support.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this thesis is to apply a multivariate analysis technique known as singular value decomposition (SVD) to classify the conformations of two analogs of the dopamine reuptake inhibitor, GBR 12909. The analogs, DM324 and TP250, differ only by a small change in the central heterocyclic ring; yet have different binding and selectivity characteristics as well as a different distribution of conformer populations among molecular shapes. The purpose of this project is to explore the potential usefulness of SVD in uncovering the relationships of torsional angles to subtle differences in conformations of GBR 12909 analogs.

SVD analysis was carried out on the eight torsional angles which connect the ring systems of the analogs and which determine the overall shape of the molecules. The reason for using this approach is to see if there are any distinctive differences in the range of torsional angles available to the conformations of these analogs that could be related to the differences in their biological activity. The long-range goal is to see if these subtle differences can be related to differences in biological activity with the goal of designing a more effective drug useful in the treatment of cocaine abuse.

This analysis was accomplished by calculating the score and loading plots generated from SVD using the conformation data generated by the Random Search function of the molecular modeling program SYBYL. A novel type of scaling that addresses the circular nature of the data proved to be critical for visualizing clusters

1.2 Background Information

1.2.1 The Dopamine Transporter and Cocaine

The dopamine transporter (Figure 1.1) is a widely-studied protein. It consists of 12 membrane-spanning segments. Though the tertiary structure is unknown it is evident that when certain molecules bind to the dopamine transporter they affect the way in which the neurotransmitter dopamine is released into the body. The dopamine transporter is located on the plasma membrane of nerve terminals and functions to transport dopamine across the membrane. Dopamine plays a very important role in the neurotransmission of the central nervous system. The dopamine transporter is a major target for drugs of abuse, mainly psychostimulants such as cocaine. The drugs can affect the way dopamine is distributed through the body by blocking the transport of dopamine[1].

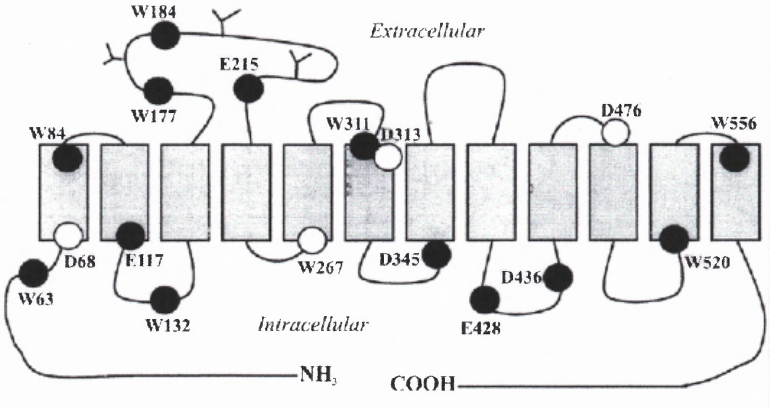


Figure 1.1 Dopamine transporter.

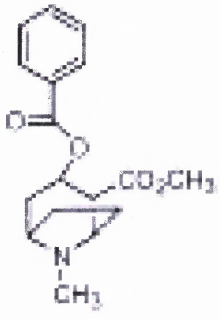


Figure 1.2 Cocaine.

Dopamine plays an important role in the control of movement, cognitive functions, and neuroendocrine systems. The dopamine transporter is a membrane-bound protein that functions to release dopamine into presynaptic terminals. The dopamine transporter works closely with the norepinephrine and serotonin transporters. All of these transporters are dependent on the presence of Na^+ and Cl^- in the extracellular fluid. Though substances such as cocaine and amphetamine can inhibit all three transporters, it is thought that the dopamine transporter is responsible for the locomotor stimulatory effects of these drugs. However, it is not clearly understood how dopamine or cocaine interact with the dopamine transporter [2].

Cocaine (Figure 1.2) blocks the normal role of the dopamine transporter in terminating dopamine signaling [3]. Structure-activity relationships have suggested the effects of cationic and aromatic interactions among dopamine, cocaine and the protein itself [4]. Studies have shown that the phenyl ring of cocaine is necessary for normal cocaine recognition by the dopamine transporter. Selective blockade of cocaine recognition in the brain reward pathway of cocaine has importance for anticocaine medications [5].

Cocaine-induced euphoria appears to result from dopamine reuptake inhibition that increases extracellular dopamine concentration in the mesolimbic and mesocorticol pathways in the transporter cocaine exerts a conformational change in the protein [7]. Cocaine appears to brain [6]. Studies have shown that upon binding to the dopamine bind to the external face of the dopamine transporter [8]. In order to develop an antagonist for drugs of abuse the relationship between inhibitor binding sites must be determined. The

challenge is to find a drug that will block the action of a drug of abuse while still allowing the dopamine transporter to function properly to transport dopamine.

Cocaine is a member of a class of compounds known as phenyltropanes. The cocaine-like compounds all share a tropane-like structure which contains a quaternary nitrogen and a phenyl ring that is attached by an ester link. These compounds have a high affinity for the dopamine transporter. However, other ligands that are structurally dissimilar to cocaine, such as the GBR 12909 series, have been shown to also have a high binding affinity for the dopamine transporter and to be highly potent inhibitors of dopamine reuptake [9]. These compounds will be discussed in the next section.

1.2.2 The Analogs of GBR 12909

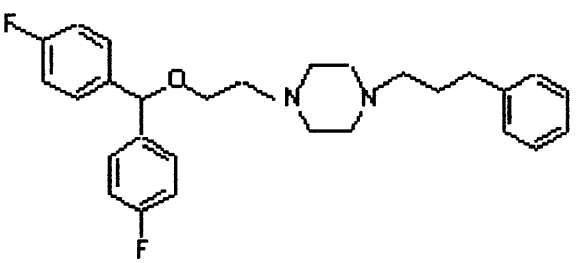


Figure 1.3 Structure of GBR 12909.

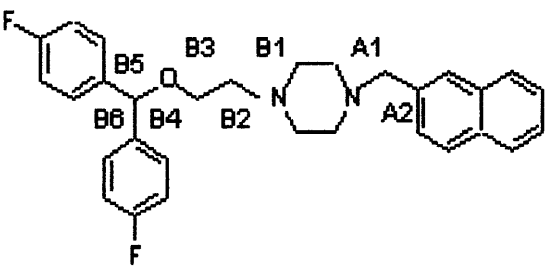


Figure 1.4 Structure of DM324.

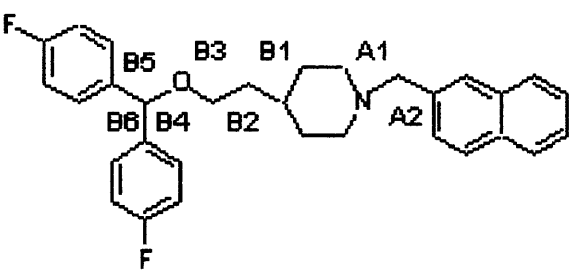


Figure 1.5 Structure of TP250.

Many compounds are under investigation as potential treatments for cocaine addiction. One of the most promising candidates is GBR 12909 shown in Figure 1.3. GBR 12909 is a dopamine reuptake inhibitor. Notice the structure of GBR 12909 is much

different than that of cocaine and yet both molecules contain the important aromatic (phenyl) ring and quaternary nitrogen. GBR 12909 was found to be competitively interactive at the cocaine-binding site. It is also important to note that GBR 12909 is highly selective for the dopamine transporter and not the serotonin transporter [10]. It has been shown that only one of the two nitrogens in the central piperazine ring is required for activity at the dopamine transporter [11]. Structures DM324 (piperazine, Figure 1.4) and TP250 (piperadine, Figure 1.5) are analogs of GBR 12909 which differ only in their heterocyclic ring system. They were chosen for analysis because they are somewhat less flexible than GBR 12909 and therefore are easier to deal with computationally. They have fewer rotatable bonds than GBR 12909 on the "A" (naphthyl) side of the molecule, while the "B" (bisphenyl) side is exactly the same as GBR 12909. Table 1.1 shows the difference in binding and selectivity for GBR 12909, DM324 and TP250 at the dopamine transporter (DAT) and serotonin transporter (SERT). The piperidine TP250 has the highest dopamine transporter binding affinity (lowest IC_{50}) and significantly better selectivity than the piperazines GBR 12909 and DM324 [12]. Since DM324 and TP250 have different biological activities, a set of conformers of each analog was analyzed in the present study to see if there are any differences in the relationship of the torsional angles to the molecular shapes.

Table 1.1 Binding Affinity of Selected GBR Analogs at the DAT and SERT Labeled With [^{125}I] RTI-55

Analog	DAT[a]	SERT[a]	SERT/DAT[b]
GBR12909 [13]	3.7 (±0.4)	126 (±5)	34
DM324 [13]	8.0 (±0.3)	312 (±15)	39
TP250 [12]	0.71 (±0.6)	229 (±21)	323
[a]IC_{50} in nanomolar concentration (nM), standard deviation in parentheses.			
[b]SERT/DAT = ratio of DAT binding affinity to SERT binding affinity.			

The eight rotatable torsional angles of DM324 and TP250 (shown in Figures 1.4. and 1.5, respectively) are the key to understanding the molecular shape. A centroid is defined here as the average position of the atoms of a ring. Deepangi Pandit of the Venanzi group classified each set of DM324 or TP250 conformations into shapes based on the distance between each "A"-side and each "B"-side centroid. The lowest of these four values was selected and used to classify the conformers into shapes. Conformers having the lowest minimum distance between centroids were classified as the C (cup) shape, followed by the I (intermediate between C and V), V (open cup), and E (extended) shapes. DM324 and TP250 were found to have a slightly different distribution of conformations among the shapes. It is for this reason that multivariate analysis is being used to elucidate the underlying relationships between the torsional angles and the molecular shapes for the DM324 and TP250 analogs.

The long-range goal is to see if these subtle differences can be related to differences in biological activity with the goal of designing a more effective treatment for cocaine abuse.

1.2.3 Principal Component Analysis and Singular Value Decomposition

Clustering is a method that uses no prior information about the class variable assumed, and the objective is to find the groups in the data. Principal component analysis (PCA) constructs a set of uncorrelated directions that are ordered by their variance. In many cases, directions with the most variance are the most relevant to the clustering. PCA works by filtering out the features with the lowest correlation between the leading principal components (PC). However, if all of the principal components are correlated to each other, then this will result with those specific components having a relatively high variance. The PC's are orthogonal directions that can be defined as the leading eigenvectors of the covariance matrix. The eigenvalue associated with each vector is the variance in that direction. Therefore the first PC explains the most variance [14].

The concept of SVD goes as far back as 1884 and is a more general form of PCA. SVD can tackle certain problems that standard PCA cannot [15]. SVD decomposes the data matrix into a score matrix and a loading matrix. Class separation is obtained via the score plots, whereas relations between variables are visualized through the loading plots. The application of SVD to the classification of molecular structures has been tested by application to DNA. SVD has been shown to correctly classify DNA X-ray structures into four well-known molecular shapes (A, B_I, B_{II}, and crankshaft) based on the nine torsional angles that define the monophosphate backbone [1]. The same general type of approach was applied here. However, the GBR analogs are far more flexible than DNA.

The work of Deepangi Pandit seems to indicate that the conformations of DM324 and TP250 take on a continuous range of shapes rather than the well-defined and distinctively different shapes of DNA. However the purpose of this project is to explore the potential usefulness of using SVD to uncover the relationships of torsional angles to subtle differences in conformations of GBR 12909 analogs.

CHAPTER 2

METHODS

2.1 Problem Statement

The SVD method is sensitive to how the data is presented to the program. The data must first be scaled before it can be decomposed. The "svd" command of MATLAB 6.0 was used to decompose the original data matrix into the three matrices, **U**, **S**, and **V**. The output was then plotted for every combination of two PC's. Any separation of conformers shows that those specific PC's are responsible for that separation. The variables with the highest correlation coefficient to the PC's that separate the conformers are considered to be the major contributors. The input data for SVD are the conformations of the GBR 12909 analogs generated by the random conformational search function of SYBYL. The rows of the input data are the conformers while the columns are the eight torsional angles.

2.2 Random Conformational Search

Conformational analysis was carried out by Milind Misra of the Venanzi group using the Random Search (RS) option in the SYBYL molecular modeling program (available from Tripos, Inc.). The algorithm is designed to locate the local minima on the conformational potential energy surface. The RS algorithm randomly alters the values of chosen torsional angles and then optimizes the geometry by minimizing the energy of the molecule at each new conformation. All of the eight non-ring torsional angles of DM324 and TP250 were allowed to vary. The torsional angles are numbered consecutively,

starting from the central heterocyclic ring: A1 andA2 on the "A" side, with A1 the closer to the central ring; B1 through B6 on the "B" side, with B1 closest to the central ring. All the angles are shown in Figure 1.4 and 1.5. The rings were held fixed as aggregates. The conformational energy of the analogs was calculated using the Tripos force field and Gasteiger-Hückel atomic point charges. The RS was allowed to run for 1,000 cycles and an energy cutoff of 20 kcal/mol was used to ensure that the RS algorithm collected all conformers with energies within 20 kcal/mol of the conformation it found to be the global minimum energy (GEM). The relative energy of each conformer was calculated by subtracting the absolute energy of the GEM from that of each conformer.

2.3 Data Pre-treatment

The input data consisted of the conformations of each analog (DM324: 728 conformations; TP250: 739 conformations) along with the eight torsional angles (A1, A2, B1 – B6). The DM324 and TP250 data were analyzed separately. Then they were combined into one 1467 x 8 matrix and analyzed together. The torsional angles of the conformers in the original RS data ranged in value from -180° to 180°.

In the case of circular data such as torsion angle data, SVD can produce erroneous results because distances between two circular data objects are defined differently than the case for linear data [16]. This is illustrated in Figure 2.1 where the distance between two data points is visualized for both linear and circular examples.

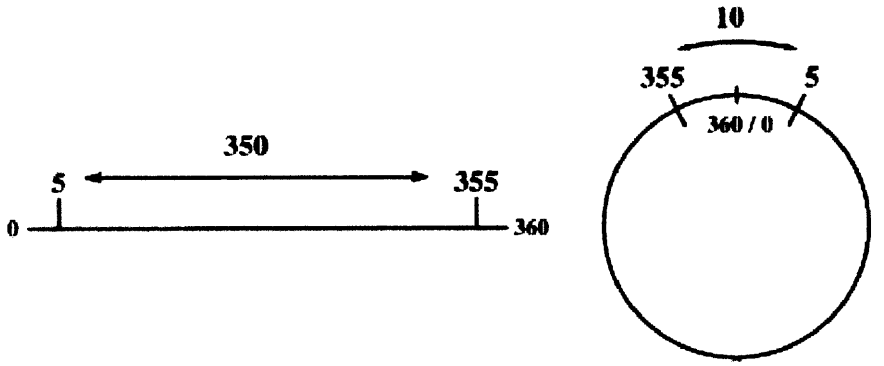


Figure 2.1 Linear data (left) shows a very large difference between the two data points. However, for circular data (right) the difference between the two data points is small.

Two approaches were used to deal with the circular data. First, the angle range was changed to be from 0° to 360° as opposed to -180° to 180° because the former range has proven to be more accurate for representing circular data [16]. The data were changed by taking all of the negative values and adding 360° to them, while leaving the positive values untouched. Second, it was suggested by Kathleen Gilbert of the Venanzi group that the data should be "GEM-scaled" by taking the values of the torsional angles of the GEM conformer of each data set and subtracting them from the corresponding torsional angle values of each conformer in the data set. For example if the GEM value has an A1 angle of 60° and another conformer has an A1 angle of 70° then that conformer has an angle value of 10° relative to the GEM angle. If another conformer has an A1 value of 50° then this new circular value is -10° relative to the GEM angle. Furthermore, to ensure that no difference between any angle and the GEM angle is either greater than 180° or less than -180°, the smallest difference between those two angles was taken. For example, if the A1 GEM value is 60° and another conformer has a value of 300° for A1, then to obtain the new, scaled A1 value for that conformer, the GEM value is subtracted from 300° (giving 240°) and that result is then subtracted from 360° to attain a value of -120°. In other words, that conformer has an A1 value that is -120° (rather than 240°) relative to

the GEM. In the opposite case where the GEM is 300° and the other conformer is 60°, then that conformer angle value is subtracted from the GEM value to give -240° and 360° must be added to this value to attain the correct value of 120°. In other words the conformer has angle 120° relative to that of the GEM.

When analyzed separately, the DM324 and TP250 data were GEM-scaled to their respective GEM conformers. When they were analyzed together, two separate calculations were carried out: one in which the data were GEM-scaled to the DM324 GEM and one in which they were GEM-scaled to the TP250 GEM.

Originally the data was median-scaled (see Appendices A and B) because some of the variables did not have a normal distribution [1]. However this presented a problem because it led to errors because the data is circular. Box plots of the data were constructed to see energy outliers. The reason for detecting outliers is to see if any data points should be discarded from the data set because they are too far away from the median value. The outliers found were only mild outliers and therefore were not removed.

2.4 Singular Value Decomposition

Singular value decomposition (SVD) decomposes the original data matrix X of dimensions $r \times c$ into three matrices, **U**, **S**, and **V** where $\mathbf{X}=\mathbf{USV}^{\mathbf{T}}$ [17]. Each row of the matrix represents a separate conformer with eight torsion angle variables contained in the columns. **U** represents a unitary matrix of dimensions $r \times r$, **V** is also a unitary matrix with dimensions $c \times c$, and **S** represents a diagonal matrix of singular values with the same dimensions as the original data matrix. The columns of **U** are left singular vectors of **XX'**, where as the rows of **V** are right singular vectors of **X'X**. **S** contains the square

roots of the eigenvalues ordered from highest to lowest [15]. Therefore the first principal component has the highest eigenvalue and consequently the largest amount of variance. These eigenvalues are the singular values.

Important analysis features of SVD are the scores and the loadings. The scores show the relationships between the principal components and each conformer while the loadings show the contribution and correlation of each angle. The columns of **U** contain the principal component scores; score values were plotted for every possible combination of two principal components. The rows of $\mathbf{SV}^{\mathbf{T}}$ contain the principal component loadings and these were plotted for every possible combination of two principal components [1]. In the loading plots the relative contribution of each torsional angle variable to each set of principal components is given by the placement of that variable from the origin, with those that contribute the most being furthest away.

2.5 Variance Explained by Each Principal Component

The matrix **S** contains the square roots of the eigenvalues. Therefore by squaring these values the eigenvalue of each principal component is obtained. Consequently, the variance explained by each principal component is simply the sum of all the eigenvalues divided by the eigenvalue of that corresponding principal component [1].

2.6 Correlation Coefficients

The correlation coefficients between the variables and the principal components are obtained by using the MATLAB function corrcoef(x,y) where x is the column of the corresponding **U** matrix and y is the column of the variable from the data matrix **X** [1].

For instance to find the correlation coefficient between PC1 and angle A1, x is the first column of matrix **U**, and y is the first column of the data matrix **X**. The correlation coefficient is related to the covariance matrix. If **C** is the covariance matrix, then the correlation coefficient matrix is the matrix whose (i,j)th element is

$$\mathbf{C}(i,j)/\mathrm{SQRT}(\mathbf{C}(i,i)*\mathbf{C}(j,j)).$$

The correlation coefficient between the variables and the principal components indicates which variables contribute to each principal component. High values with either negative or positive signs indicate major contributors [1].

The correlation coefficients between the variables themselves are obtained by taking the correlation coefficient matrix of the entire data set. Therefore the correlation coefficient matrix has dimensions $c \times c$.

2.7 Software

Singular value decomposition including the correlation matrices and plots was performed with MATLAB for Windows version 6.0 by The Mathworks, Inc. Detection of outliers was performed using Microsoft Excel Add-In SSC Stat 4.0 box plot option.

CHAPTER 3

RESULTS

3.1 DM324

3.1.1 Random Conformational Search

The random conformational search of DM324 produced 728 conformers with energy ranging from that of the GEM conformer at 11.2 kcal/mol to a maximum of 28.9 kcal/mol. This is a range of 17.7 kcal/mol relative to the energy of the GEM taken as 0.0 kcal/mol. The torsional angles A1, A2, and B1-B6 of the GEM conformer are 63.8°, 88.6°, 262.7°, 310.1°, 183.4°, 64.4°, 115.1°, and 339.8°, respectively. These were used to carry out the GEM scaling described in the Methods section.

3.1.2 Box Plots

Figure 3.1 shows the box plot of the energies for DM324. The box plot was used as a means to search for any outliers. The outliers shown are only mild outliers and were subsequently not removed from the data set before performing SVD analysis.

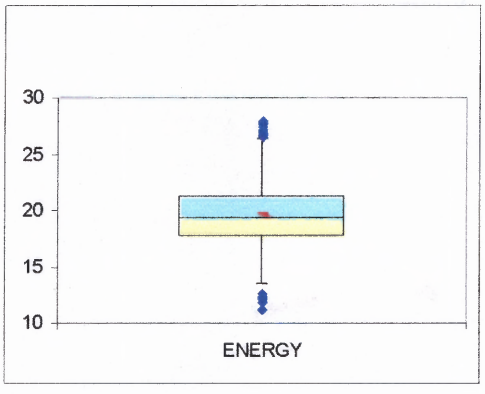


Figure 3.1 Box plot of DM324 based on energy. Outliers are shown in blue; the median is represented by the red line.

3.1.3 Singular Value Decomposition

Score plots were plotted for every possible combination of two principal components (see Appendix C). Those that did not involve PC1 did not show separation of the data.

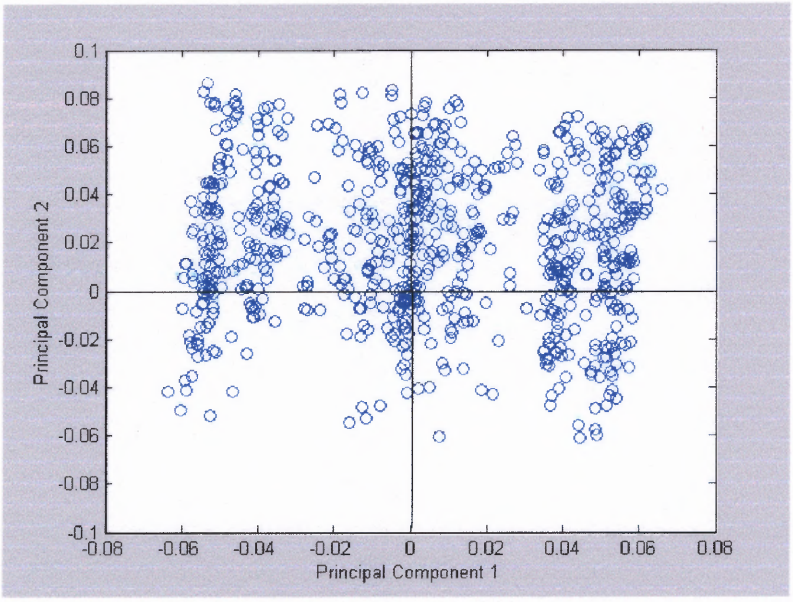


Figure 3.2 Score plot of DM324 GEM-scaled data for PC1 vs PC2.

Figure 3.2 shows the score plot of all 728 conformers of DM324 for PC1 vs PC2. Because the data are GEM-scaled the points are given relative to the GEM conformer located at the origin. It is clear from the figure that the data separate into three distinct groups along the PC1-axis. Furthermore each of the three groups appears to be subdivided into two groups (or possibly three, in the case of the middle cluster). The separation of data seen in this plot is typical of the score plots of *any* PC with PC1. The PC1 vs. PC2 plot was selected for presentation because it shows the best separation of the data and corresponds to the components which explain the highest percentage of the variance.

Figure 3.3 shows the loading plot of all eight torsional angles for PC1 vs. PC2. The variables with the largest contribution to each principal component are furthest from the origin. The loading plot is a way to visualize the correlation coefficients between the angles and the PC's (given in Table 3.2 below). The loading vector for each angle has a component along the PC1 (or PC2) axis proportional to its correlation coefficient to PC1 (or PC2). As is evident from Figure 3.3, angles A1 and A2 are the highest contributors to PC1 because they are the furthest away from the origin along the PC1 axis. Therefore the angles A1 and A2 are responsible for separating the data in Figure 3.2. Since the A1 and A2 loading vectors are found along the negative PC1 axis, these variables have a large negative correlation with PC1 and a small positive correlation to PC2.

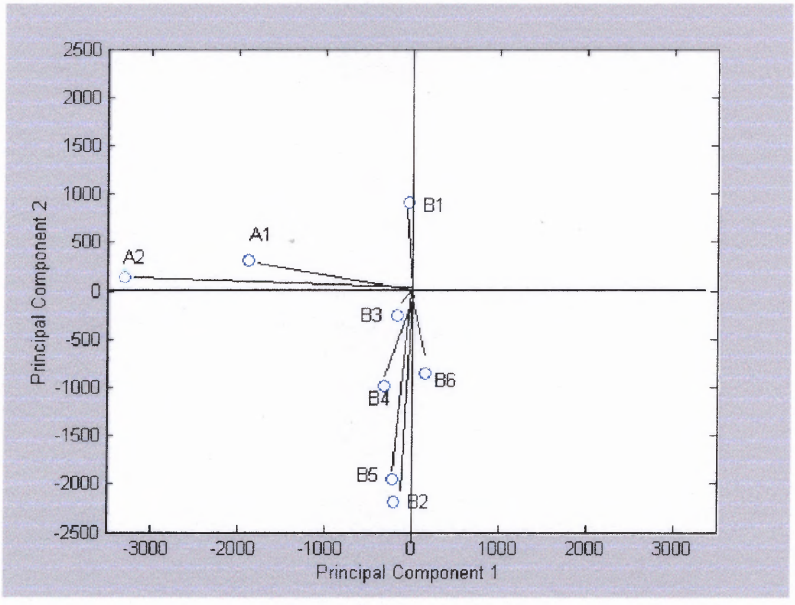


Figure 3.3 The loadings plot of the variables on PC1 vs. PC2.

The terms "high", "middle", and "low" will be used throughout the next sections to describe the values of the (A1, A2, B1-B6) torsional angles of each conformer relative to those of the GEM conformer. Each term is associated with a range of values of the

principal component along which the data separates and can be related to a range of values of those torsional angles which are the chief contributors to that principal component. If those torsional angles have a *negative* correlation to the principal component (as seen in the loading plot of PC1 vs. PC2 for DM324 in Fig. 3.3), then "high" will be defined in terms of *negative* values of that principal component as described below. If, however, those torsional angles which are the chief contributors have a *positive* correlation to the principal component (as is the case of TP250, Figures 3.8 and 3.10), then "high" will be defined in terms of *positive* values of that principal component as described in the TP250 section below.

The data points in Figure 3.2 can roughly be divided into three clusters based on their values along the PC1 axis. The left-hand cluster has a PC1 value approximately equal to -0.02 to -0.06. Conformers in that cluster are defined to have "high" values of A1 and A2 relative to the values of A1 and A2 in the GEM conformer. The right-hand cluster has PC1 values equal to 0.02 to 0.06. Conformers in that cluster are defined to have "low" values of A1 and A2 relative to the values of A1 and A2 in the GEM conformer. Data points in the middle cluster have PC1 values between -0.02 to 0.02. Conformers in this cluster have similar (or "middle" between high and low) values of A1 and A2 relative to the values of A1 and A2 in the GEM conformer.

The concept of "high", "middle", and "low" can be illustrated by plotting angles of representative conformers from each of the three groups on the clock face shown in Figure 3.4.

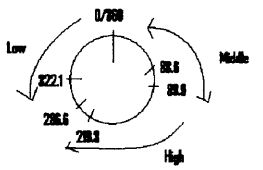


Figure 3.4 Clock face showing low, middle, and high values of angle A2 relative to GEM A2 of DM324.

Points on the clock face designate torsional angle values between 0° and 360°. Those data points in the left-hand cluster of Figure 3.2 will be distributed along the clockface going in a clockwise direction from the GEM value, such that those that correspond to the more negative value of PC1 will be found on the clock face closer to the GEM value plus 180°. These are said to have "high" values relative to the corresponding GEM angle. Similarly, those data points in the right-hand cluster of Figure 3.2 will be distributed along the clockface going in a counter-clockwise direction from the GEM value, such that those that correspond to the more positive values of PC1 will be found on the clock face closer to the GEM value plus 180°. These are said to have "low" values relative to the corresponding GEM angle. Those data points in the middle cluster of Figure 3.2 will be distributed around the GEM values in both a clockwise and counter-clockwise fashion, depending on their corresponding negative or positive value of PC1, respectively. These angles are said to have "middle" values with respect to the corresponding angle in the GEM conformer.

Figure 3.4 plots the A2 values of some DM324 conformers relative to the DM324 GEM value of A2. Torsional angle A2 of the GEM conformer (conformer number 683)

has a value of 88.6°. Conformers with "high" values of A2 are found distributed clockwise along the clock face closer to 88.6° + 180° = 268.6°. Conformer values with "low" values of A2 are found distributed counter-clockwise along the clock face closer to 268.6°. Conformers with "middle" values of A2 are found distributed around the GEM value at 88.6°. For example conformer 490 has a value of -0.04 for PC1 and is found in the left-hand cluster of Figure 3.2. It has an A2 value of 219.3° and is found on the clock face by moving in a clockwise direction from the GEM A2 value of 88.6°. Its value is relatively close to 268.6°, so it is said to be "high" relative to the GEM value of 88.6°. Similarly conformer 408 has a PC1 value of 0.04 and is found in the right-hand cluster of Fig. 3.2. It has an A2 value of 322.1° and is found on the clock face by moving in a counter-clockwise direction from the GEM. It is relatively close to 268.6° and is said to have a "low" value relative to the GEM A2 value. Finally, conformer 682 has a PC1 value of 0.0006 and is found in the center cluster of Figure 3.2. It has an A2 value of 89.9° which is close to the GEM value. So this angle has a "middle" value compared to the GEM A2 value.

3.1.4 Variance Explained by Each Principal Component

Each principal component has a specific variance associated with it. The sum of the variances of each PC is equal to 100. The percentage of the variance explained by each PC is shown in the Table 3.1. Table 3.1 shows that for DM324 no one PC explains a large part of the variance. In fact the first three PC's explain only 53.55% of the variance.

Table 3.1 Percentage of Variance Explained by Each PC After SVD of GEM-Scaled DM324 Data

Principal	Eigenvalue^2* 10^6	Variance	Σ Variance(%)
1	1.4686	23.20	23.20
2	1.1266	17.80	41.00
3	0.8580	13.55	53.55
4	0.7547	11.92	65.47
5	0.6624	10.47	75.94
6	0.5624	8.89	84.83
7	0.5485	8.67	93.50
8	0.3487	5.51	99.01

3.1.5 Correlation Coefficients

Table 3.2 shows the correlation coefficients between the PC's and each variable. The values in red indicate the major contributors to each PC. This table shows a quantitative view of Figure 3.3. PC1 and A1 have the second highest negative correlation (-0.7), while PC2 and A1 have a small positive correlation (0.1). That is why the A1 loading vector appears along the negative PC1 axis slightly tilted towards the positive PC1 axis in Figure 3.3. Similarly, PC1 and A2 have the highest negative correlation (-0.9), while PC2 and A2 have a small positive correlation (0.1). That is why the A2 loading appears farthest from the origin along the negative PC1 axis, tilted slightly towards the positive PC2 axis. On the other hand, B2 and B5 have negative correlations to PC2 (-0.6) but a very small negative correlation (-0.05) to PC1, so their loading vectors appear to fall along the negative PC2 axis, slightly tilted towards the negative PC1 axis. With this in mind, it can be said that if a loading plot for a variable appears very close to the origin this means that it does not contribute to either PC. This is the case of B3, which has a value of -0.1 correlation coefficient with both PC1 and PC2.

Table 3.2 Correlation Coefficients Between the Angles and the PC's for GEM-Scaled DM324

Angle	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A1	-0.68	0.11	-0.04	0.07	-0.14	-0.62	0.35	-0.03
A2	-0.94	0.11	-0.04	0.00	0.02	0.30	-0.15	0.02
B1	-0.03	0.24	0.82	0.08	-0.46	0.03	-0.15	0.05
B2	-0.03	-0.64	0.38	0.56	0.18	-0.05	0.09	0.15
B3	-0.08	-0.10	0.35	0.07	0.30	-0.05	-0.11	-0.87
B4	-0.11	-0.37	0.14	-0.50	0.20	-0.43	-0.60	0.12
B5	-0.05	-0.65	0.10	-0.60	-0.37	0.16	0.32	-0.07
B6	0.08	-0.27	-0.48	0.34	-0.65	-0.10	-0.36	-0.14

Table 3.3 Correlation Coefficients Between all Eight Angles for DM324 GEM-Scaled Data

	A1	A2	B1	B2	B3	B4	B5	B6
A1	1	0.40	0.01	0.01	0.00	0.03	-0.00	-0.00
A2	0.40	1	0.01	-0.02	0.04	0.04	-0.00	-0.05
B1	0.01	0.01	1	0.07	0.09	-0.02	-0.00	-0.09
B2	0.01	-0.02	0.07	1	0.16	0.03	0.05	0.02
B3	0.00	0.04	0.09	0.16	1	0.10	-0.02	-0.14
B4	0.03	0.04	-0.02	0.03	0.10	1	0.17	-0.04
B5	-0.00	-0.00	-0.00	0.05	-0.02	0.17	1	0.02
B6	-0.00	-0.05	-0.09	0.02	-0.14	-0.04	0.02	1

Table 3.3 shows the correlation coefficients between all eight angles. These correlation coefficients are not partial coefficients but full coefficients. Table 3.3 shows that angles A1 and A2 have a high positive correlation (0.39), which is the highest

correlation out of all the possible combinations of variables. This is twice the magnitude of the next largest correlation coefficients for B2/B3 (0.17), B3/B6 (-0.15) and B4/B5 (0.18). This is important because A1 and A2 are responsible for separating the data and contain the highest variances of any of the PC's.

3.2 TP250

3.2.1 Random Conformational Search

The random conformational search of TP250 produced 739 conformers with energy ranging from that of the GEM conformer at 15.8 kcal/mol to a maximum of 29.1 kcal/mol. This is a range of 13.3 kcal/mol relative to the energy of the GEM taken as 0.0 kcal/mol. The torsional angles A1, A2, and B1-B6 for the GEM conformer are 298.7°, 291.8°, 51.5°, 58.5°, 189.6°, 306.7°, 67.9°, and 6.5°, respectively.

3.2.2 Box Plots

Figure 3.5 shows the box plot of energies of TP250. The box plot shows only mild outliers and therefore they were not removed from the data set before performing SVD analysis.

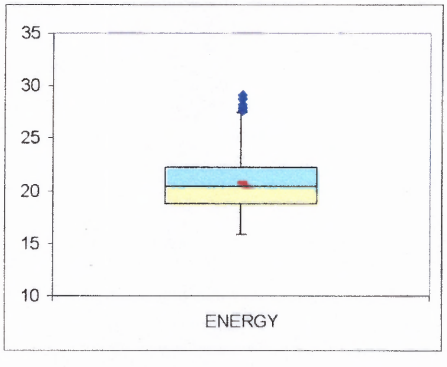


Figure 3.5 Box plot of energy for TP250. Mild outliers are shown in blue; the median is represented by the red line.

3.2.3 Singular Value Decomposition

Score plots were constructed for all possible combinations of PC's (see Appendix D) but the data did not separate on any of the PC's from PC1-PC6. For example, Figure 3.6 shows that the data does not separate on either PC1 or PC2. This means that the variables responsible for those PC's do not separate the data into clusters.

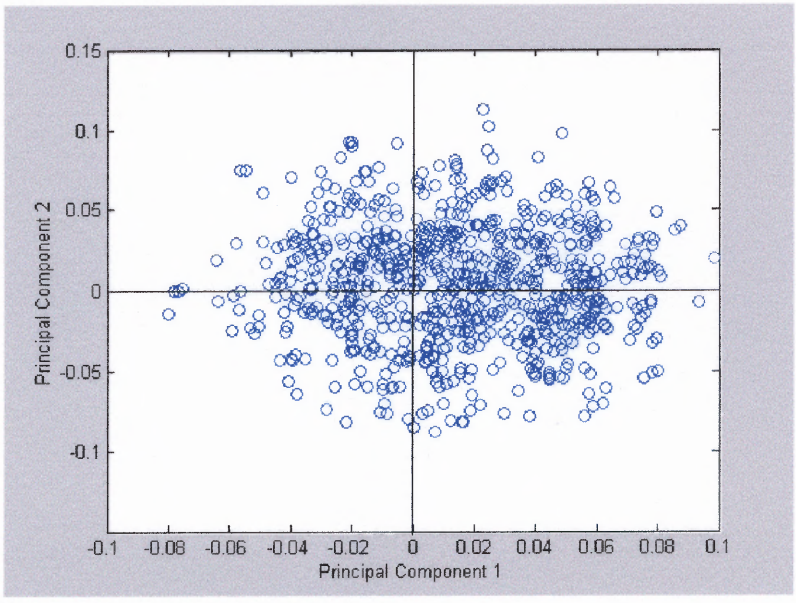


Figure 3.6 Score plot of all 739 conformers of TP250 GEM-scaled data for PC1 vs. PC2.

However, the score plot of PC1 vs PC7 (Figure 3.7) shows that the data separates into three groups along the PC7 axis. This is typical of the score plot of *any* PC with PC7. Figure 3.8 shows the loadings for PC1 vs PC7. It is evident that the major contributor to PC7 is angle B4. Since angle B4 is the furthest away from the origin on the positive PC7 axis, it has a large positive correlation to PC7. Since the data do not separate on PC1, the angles (A2 and B1) that are the major contributors to that PC are not as important as B4. Therefore it is angle B4 that is responsible for separating the data into those three groups.

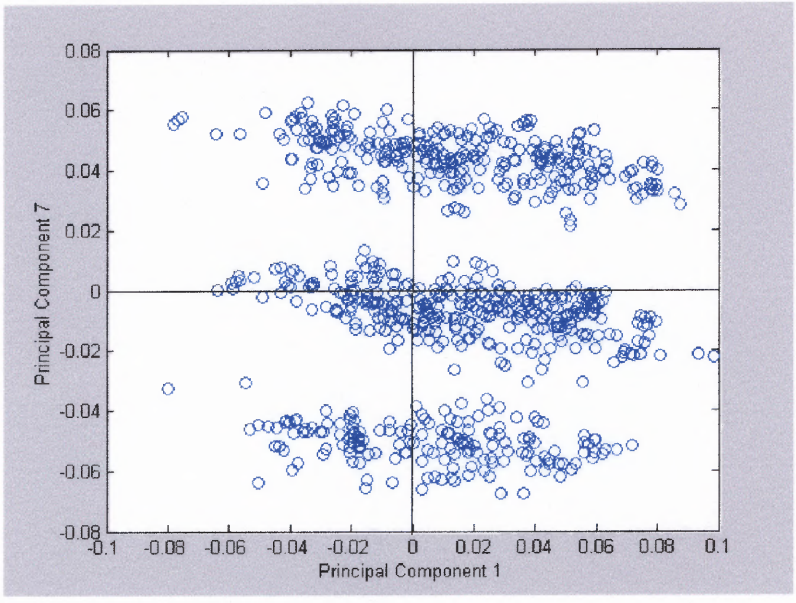


Figure 3.7 Score plot of all 739 conformers of TP250 GEM-scaled data for PC1 vs. PC7.

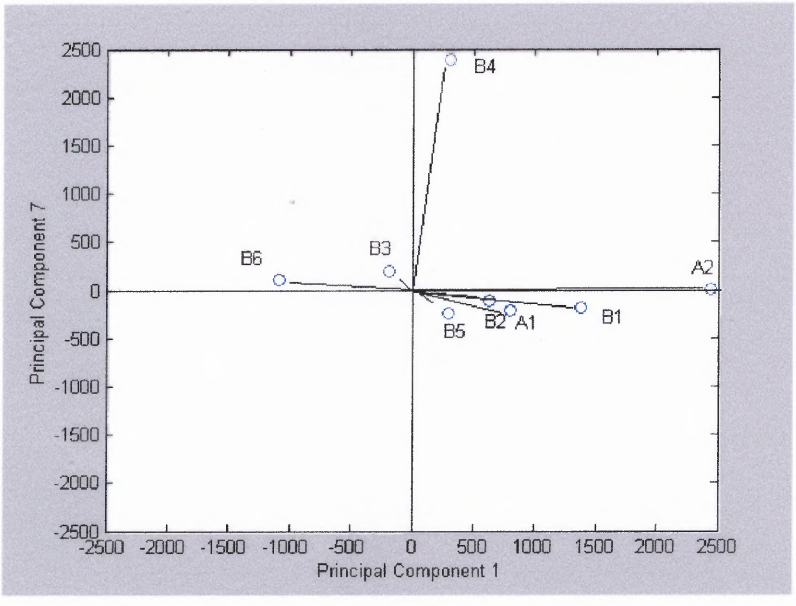


Figure 3.8 Loading plot of eight variables of TP250 GEM-scaled data for PC1 vs. PC7.

The three groups in Figure 3.7 contain conformers that have values which are either "high", "low", or "middle" relative to the B4 value of the GEM conformer, 306.7°, based on their range of PC7 values. Since angle B4 has a large *positive* correlation to PC7, "high" is defined in terms of *positive* values of PC7 (in contrast to the definition of "high" for DM324). Conformers that have PC7 values in the range of 0.04 to 0.06 have "high" values of B4 relative to the GEM B4. Their B4 values would be distributed on a clock face in a clockwise fashion from the GEM value of 306.7° to 126.7° (i.e. 306.7° + 180 ° – 360°). Those with more positive PC7 values would be found closer to 126.7°. Conformers that have PC7 values in the range of -0.04 to -0.06 have "low" values of B4 relative to the GEM conformer's B4. Their B4 values would be distributed on a clock face in a counter-clockwise fashion from the GEM value of 306.7° to 126.7°. Those angles with more negative PC7 values would be found closer to 126.7°. Conformers with PC7 values between -0.02 and 0.02 are defined as "middle" relative to the B4 GEM. Their values would be found on either side of the B4 GEM on the clock face.

Representative conformers were chosen from each of the three groups in Figure 3.7 in order to illustrate the definitions of "high, "middle, and "low". Conformer number 135 is the GEM conformer of TP250. Conformer number 9 has PC7 value equal to 0.05 and was chosen from the "high" cluster of conformers with PC7 between 0.04 and 0.06. This conformer has a value for B4 of 55.7°, which is found by moving clockwise from 306.7° and is considered "high" relative to the GEM B4. Conformer number 737 was taken from the "low" cluster with PC7 value of -.04. It has B4 value of 189.3° which is found by moving in a counter-clockwise direction from 306.7° and is considered "low"

relative to the GEM. Conformer number 84 was taken from the "middle" cluster and is very close to the origin. It has a B4 value of 308.8° which is very similar to the GEM.

The data also separates into three groups along the PC8-axis as shown in Figure 3.9. This is typical of the score plots of *any* PC with PC8. The loading plot for PC1 vs PC8 (Figure 3.10) shows that the major contributor to PC8 is angle B3, as B3 is the furthest from the origin along the positive PC8 axis. Since B3 has a large positive correlation to PC8, the three clusters can be divided into groups of angles with values that are "high" (conformers with PC8 between 0.04 and 0.06), "middle" (conformers with PC8 between -0.02 and 0.02), and "low" (conformers with PC8 between -0.04 and -0.06) relative to the GEM B3.

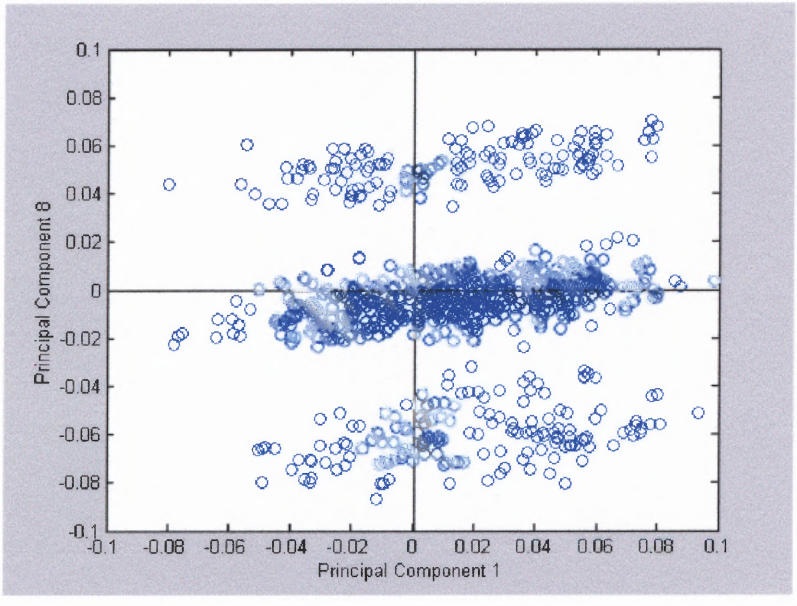


Figure 3.9 Score plot of all 739 conformers of TP250 GEM-scaled data for PC1 vs. PC8.

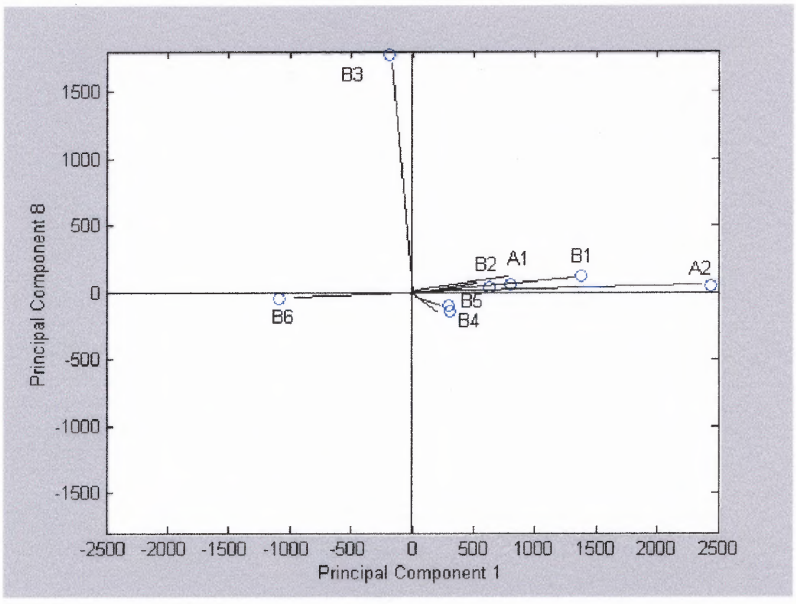


Figure 3.10 Loading plot of variables of TP250 GEM-scaled data for PC1 vs. PC8.

However, since both angles B3 and B4 are responsible for separating the TP250 data into groups, the data was also plotted on PC7 vs PC8. Nine separate groups are formed as shown in Figure 3.11.

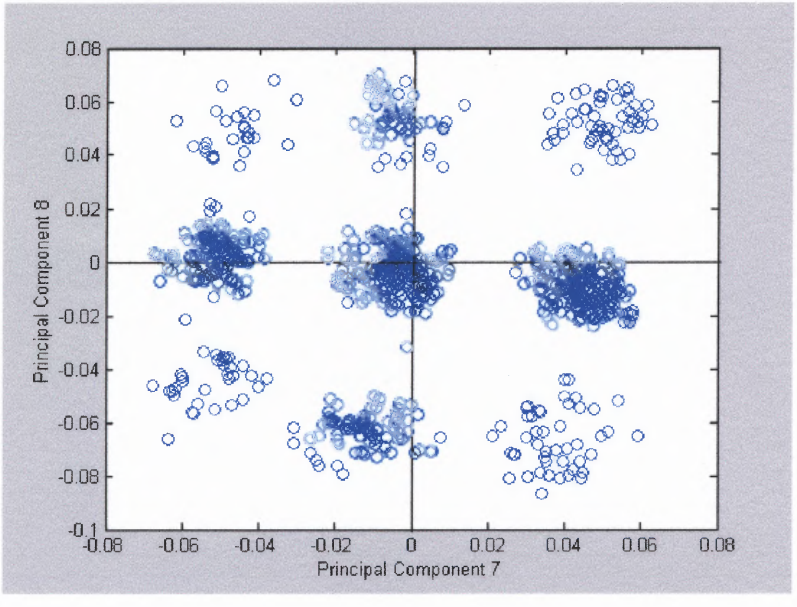


Figure 3.11 Score plots of all 739 conformers of TP250 GEM-scaled data for PC7 vs. PC8.

The same representative conformers from the PC7 plots were analyzed for their respective B3 values. The GEM conformer, number 135, has a B3 value of 189.6°. Conformer number 9 is from the "high" cluster in Fig. 3.9 and has a PC8 value of 0.06. It has a B3 value of 306.7° which is located by moving clockwise from the GEM by 117.1°. This angle is relatively far from the GEM and is therefore considered "high" compared to the GEM. Conformer number 737 is from the "low" cluster in Figure 3.9. It has a PC8 value of -0.06. It has a B3 value of 89.3°, which is located by moving 100.3° in a counter-clockwise direction from the GEM value. It is relatively far from the GEM and is considered "low" compared to the GEM value. Conformer number 84 is from the "middle" cluster in Figure 3.9 and is close to the origin. It has a B3 value of 192.2° which is very similar to that of the GEM conformer.

Since conformers 9 and 737 both have positive values of PC7 and PC8 in the range of 0.4 to 0.6, they are found in the group of conformers (Group 3, see below) that are in the upper right quadrant in Figure 3.11. This means that these particular conformers all have B3 and B4 values that are high relative to those of the GEM. Since these conformers have high PC7 and PC8 values, there exists a high correlation with B4 and B3, respectively.

Each of the nine groups in Fig. 3.11 cluster by PC7/PC8 values as well as B3/B4 angles. Group 1 in the upper left-hand corner has low PC7 values (-0.04 to -0.06) but high PC8 values (0.04 to 0.06). Therefore these conformers have low B4 and high B3 values with respect to the GEM's B4 and B3 values. Group 2, the upper middle group in Figure 3.11, has middle PC7 values (-0.02 to 0.02) and high PC8 values. These conformers in turn have B4 values similar to that of the GEM and B3 values higher than

that of the GEM. Group 3 conformers (in the upper right-hand corner) have high PC7 and PC8 values and therefore have high B4 and B3 values with respect to the GEM. Conformers 283 and 737, used as examples above, are found in this group. Group 4, the middle left group, has low PC7 and middle PC8 values. These conformers have low B4 values and similar B3 values with respect to the GEM. Group 5, the middle group, has values for both PC7 and PC8 near the origin. Therefore conformers in this group, such as conformer 84, have B4 and B3 values similar to those of the GEM. Group 6, the middle right group, has high PC7 values with PC8 values near the origin. Therefore these conformers have high B4 values and similar B3 values relative to the GEM. Group 7 in the lower left hand corner of Figure 3.11 has conformers with both low PC7 and PC8 values which, in turn, have low B3 and B4 values with respect to the GEM. Group 8 is the lower middle group and these conformers have PC7 values near the origin and low PC8 values. Therefore these conformers have similar B4 values and low PC8 values with respect to the GEM. Group 9 has high PC7 values and low PC8 values and contains conformers that have high B4 values and low PC8 values with respect to those of the GEM.

The major contributors to each PC can also be found by calculating the correlation coefficients of each variable with each PC. The results of the correlation coefficients are shown in Section 3.5 and agree with the loading plots of each data set.

3.2.4 Variance Explained by Each Principal Component

The variance explained by each principal component of the TP250 GEM-scaled data is shown in Table 3.4. Although PC7 and PC8 explain the smallest percentage of the variance (10.42% and 5.65%, respectively), it is not that small compared to the variance of PC1 (18.13%) or PC2 (15.01%). For instance the sum of the variances of PC7 and PC8 equals the variance of PC2.

Table 3.4 Percentage of Variance Explained by Each PC After SVD of GEM-Scaled TP250 Data

PC	Eigenvalue^2* 10^6	Variance explained	Σ Variance(%)
1	1.0263	18.13	18.13
2	0.8496	15.01	33.04
3	0.8143	14.38	47.42
4	0.7970	14.08	61.50
5	0.6659	11.76	73.26
6	0.5979	10.56	83.82
7	0.5899	10.42	94.24
8	0.3201	5.65	99.89

As with DM 324, Table 3.4 shows that for TP250 no one PC explains a large part of the variance. The first three PC's taken together explain only 47.42% of the variance.

3.2.5 Correlation Coefficients

The correlation coefficients between the angles and each of the PC's for the TP250 GEM-scaled data set are given in Table 3.5. This is the data used to produce the qualitative pictures shown by the loading plots in Figures 3.8 and 3.10. Table 3.5 shows that A2 has a high positive correlation (0.79) with PC1 and a very small negative correlation (-0.02) with PC7. That is why, in the PC1 vs. PC7 loading plots shown in

Figure 3.8, the A2 loading vector appears far from the origin along the positive PC1 axis. Similarly, B4 has a high positive correlation (0.98) with PC7 and a small positive correlation (0.08) with PC1. That is why the B4 loading vector is found far from the origin along the positive PC7 axis, slightly tilted towards the positive PC1 axis, in Figure 3.8. B3 has a high positive correlation (0.98) with PC8 but a very small negative correlation (-0.03) with PC1. That is why, for the PC1 vs. PC8 loading plot in Figure 3.10, the B3 loading vector is found far from the origin along the positive PC8 axis, slightly tilted towards the negative PC1 axis. Although A2 is not responsible for separating the TP250 data, Table 3.5 shows that A2 has a large positive correlation (0.79) with PC1 and a small positive correlation (0.07) with PC8. For this reason, in Figure 3.10 the A2 loading vector is found far from the origin along the positive PC1 axis, tilted slightly towards the positive PC8 axis.

Table 3.5 Correlation Coefficients Between the Angles and the PC's for TP250 GEM-Scaled Data

Angle	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A1	0.31	0.33	-0.10	0.10	0.01	-0.88	-0.09	0.04
A2	0.79	0.07	0.30	-0.33	-0.41	0.13	-0.02	0.07
B1	0.45	-0.33	-0.57	-0.29	0.51	0.05	-0.08	0.08
B2	0.14	0.64	0.27	0.37	0.50	0.24	-0.05	0.05
B3	-0.03	0.08	-0.04	0.04	-0.07	0.05	0.13	0.98
B4	0.08	0.03	-0.12	0.07	0.02	-0.05	0.98	-0.04
B5	0.10	0.38	-0.72	0.33	-0.41	0.21	-0.09	-0.03
B6	-0.38	0.57	-0.12	-0.73	0.05	-0.02	0.04	-0.03

Table 3.6 Correlation Coefficients Between all Eight Angles for TP250 GEM-Scaled Data

	A1	A2	B1	B2	B3	B4	B5	B6
A1	1	0.07	0.02	0.07	-0.00	0.01	0.07	-0.00
A2	0.07	1	0.06	-0.00	0.00	-0.00	-0.03	-0.04
B1	0.02	0.06	1	-.0.01	-0.02	0.01	0.04	-0.03
B2	0.07	-0.00	-0.10	1	0.03	-0.00	0.03	0.01
B3	-0.00	0.00	-0.02	0.03	1	0.06	0.05	0.01
B4	0.01	-0.00	0.01	-0.00	0.06	1	0.03	-0.01
B5	0.07	-0.03	0.04	0.03	0.05	0.03	1	0.01
B6	-0.00	-0.04	-0.03	0.01	0.01	-0.01	0.01	1

Table 3.6 gives the correlation coefficients between all the eight angles for the TP250 GEM-scaled data. These coefficients represent full, not partial, coefficients. The table shows that, in contrast to DM324, there is no large correlation between any of the variables. Although angles B4 and B3 separate the data along PC7 and PC8, respectively, they have only a very small, positive correlation (0.07) because this separation occurred along two different PC's. This is in contrast to the DM324 case, where A1 and A2 have a correlation of 0.40 (Table 3.3) and both are responsible for separating along the *same* PC (PC1). Therefore angles B3 and B4 in the TP250 data set do not behave similar to angles A1 and A2 in the DM324 data set. The angles B1 and B2 have the largest correlation (-.10) in the table. In summary, the DM324 data separate along PC1 due to A1 and A2. The TP250 data separate along PC7 due to B4 and PC8 due to B3.

3.3 DM324 and TP250 Together

Two different approaches were used to analyze the combined data sets. The first approach used the GEM conformer of DM324 to scale the data. The second approach used the GEM conformer of TP250 to scale the data. Only the score plots that showed the best separation of the data are given below. The other score plots are given in the appendix (Appendix E: combined data scaled to DM324 GEM; Appendix F: combined data scaled to TP250 GEM). In the score plots the DM324 conformers are indicated by circles and the TP250 conformers by plus (+) signs.

3.3.1 Results of Scaling to the DM324 GEM Conformer

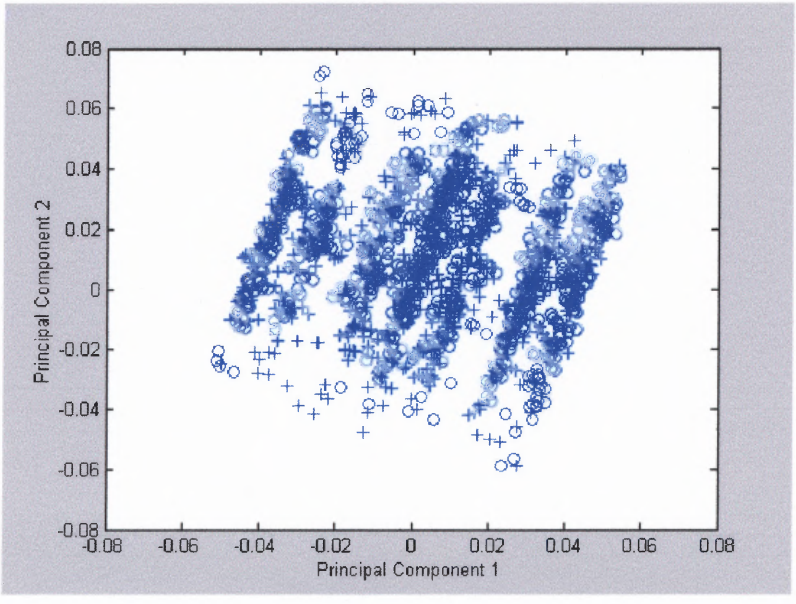


Figure 3.12 Score plot of DM324 and TP250 data GEM-scaled to DM324 GEM for PC1 vs. PC2.

The score plot in Figure 3.12 is somewhat similar to that in Figure 3.2 in which the DM324 data is analyzed separately. In both figures there are three major groups with two clear subdivisions in the right- and left-handed groups. Figure 3.12 shows three subdivisions in the middle cluster of conformers. This is less obvious in Figure 3.2. The

score plots differ in the orientation of the three clusters with respect to the PC1 and PC2 axes. In Figure 3.2 the three groups are oriented parallel to the PC2 axis and clearly separate along the PC1 axis. In Figure 3.12, they are slightly "skewed" with respect to the two axes, almost as if the three clusters had been rotated clockwise around the origin.

In the case of the DM324 data, Figure 3.2 is typical of the score plot of any PC with PC1. This is not true for Figure 3.12. For the combined data, the score plots of higher PC's with PC1 give data that is progressively more skewed so that data separation decreases as the PC number increases except for PC1 vs. PC8 (see Appendix E).

Figure 3.12 shows that the analogs do not separate from one another in the PC1 vs PC2 score plot. Conformers of DM324 and TP250 are found throughout each of the groups. This is typical of all the score plots in Appendices E and F and shows how similar the analogs are in comparison to one another. This is not surprising based on the similarity of their molecular structures (Figures 1.4 and 1.5). The two analogs differ only by the replacement of a nitrogen (lone pair) by a carbon (hydrogen) in the central ring system.

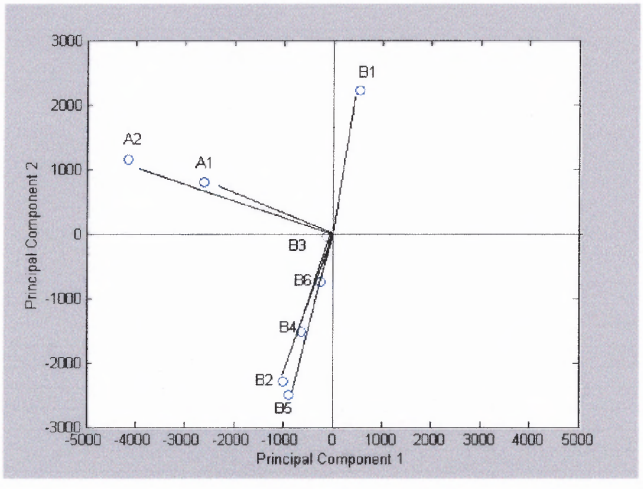


Figure 3.13 Loading plot of DM324 and TP250 data GEM-scaled to the DM324 GEM for PC1 vs. PC2.

The loading plot of the combined data GEM-scaled to the DM324 GEM (Figure 3.13) is similar to that of the DM324 data (Figure 3.3) in that A1, A2, B2, and B5 are the major contributors. As with the DM324 data, A1 and A2 have large negative correlations to PC1 and are responsible for separating the data along PC1. For both data sets, B2 and B5 have large negative correlations to PC2. The loading plots differ in the fact that, for the combined data, B1 has a larger positive correlation to PC1 than for the DM324 data and is therefore a major contributor for the combined data set. The loading plot in Figure 3.13 appears to be rotated clockwise from that in Figure 3.3. This is due to the fact that, in the combined data, A1 and A2 have larger positive correlations to PC2, B2 and B5 have larger negative correlations to PC1, and B1 has a larger positive correlation to PC1 than in the DM324 data. This causes the skewing of the data so that the clusters do not lie parallel to the PC2 axis and makes it difficult to define "high", "middle", and "low" values of the angles relative to the GEM angles because the data do not separately cleanly along PC1.

In contrast, the TP250 data, when analyzed alone, do not separate along PC1 as shown in the PC1 vs. PC2 score plot (Figure 3.6). Table 3.5 shows that A2 and B1 are the chief contributors to PC1 and B2 and B6 are the chief contributors to PC2 for this data set. B1 also has a negative correlation to PC2 (-0.33) and B6 has a negative correlation to PC1 (-0.38). This gives a very different loading plot (not shown) for the TP250 data in Figure 3.6 and explains, in part, why the TP250 score plot for PC1 vs. PC2 is so different from that of the combined data set. It appears that scaling the TP250 data to the DM325 GEM angle values causes the TP250 data set to take on some of the characteristics of the DM324 data set.

Figure 3.14, the score plot of the combined data for PC7 vs PC8, shows a separation into three groups along the PC8 axis. This plot is typical of the score plots of the combined data for *any* PC with PC8 for this data set. This figure is similar to Figure 3.9 where TP250, when analyzed alone for PC1 vs. PC8, separated into three groups along the PC8 axis. Similarly, the loading plot for the combined data (Figure 3.15) shows that the major contributor to PC8 is B3, as in the TP250 case (Figure 3.10).

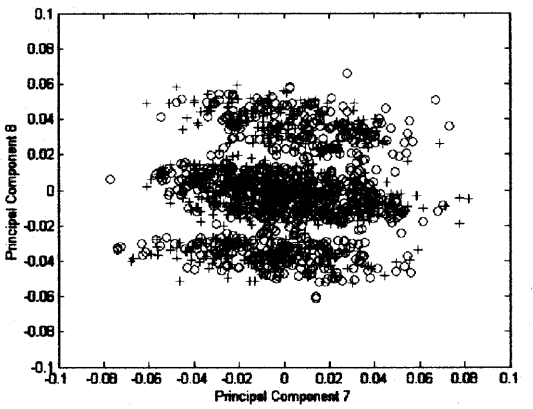


Figure 3.14 Score plot of DM324 and TP250 GEM-scaled to DM324 GEM for PC7 vs. PC8.

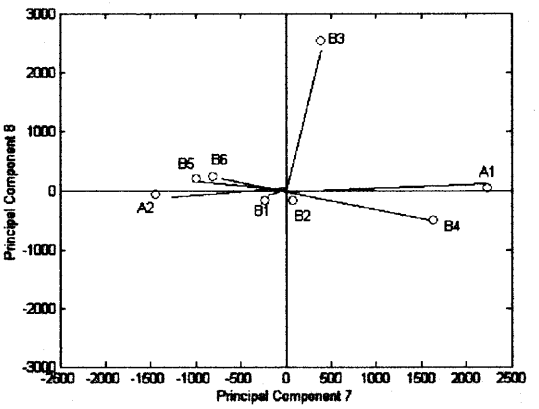


Figure 3.15 Loading plot of DM324 and TP250 GEM-scaled to DM324 GEM for PC7 vs. PC8.

Table 3.7 Percentage of Variance Explained by Each PC After SVD of DM324 and TP250 Data Combined and GEM-Scaled to the DM324 GEM

PC	Eigenvalue^2* 10^6	Variance explained	Σ Variance(%)
1	2.661	21.56	21.56
2	2.1281	17.21	38.77
3	1.6264	13.15	51.92
4	1.4999	12.13	64.05
5	1.4049	11.36	75.41
6	1.1980	9.69	85.10
7	1.1555	9.34	94.44
8	0.6863	5.55	99.99

Table 3.8 Correlation Coefficients Between the Angles and the PC's for DM324 and TP250 Data Combined and GEM-Scaled to the DM324 GEM

Angle	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A1	-0.66	0.23	-0.05	-0.14	-0.09	-0.43	0.55	0.01
A2	-0.89	0.30	0.07	0.02	0.06	0.22	-0.31	-0.01
B1	-0.08	0.49	-0.24	0.71	-0.38	-0.10	-0.05	-0.02
B2	-0.19	-0.47	-0.60	0.40	0.37	-0.09	0.01	-0.07
B3	-0.03	0.02	0.00	0.20	0.02	0.20	0.14	0.94
B4	-0.14	0.36	0.22	0.24	-0.39	0.58	0.45	-0.15
B5	-0.18	-0.60	0.52	0.24	-0.27	-0.41	-0.25	0.03
B6	-0.03	-0.12	-0.56	-0.41	-0.65	-0.00	-0.21	0.05

Table 3.7 gives the percentage of the variance explained by each of the PC's. As in the case of the DM324 data (Table 3.1) and TP250 data (Table 3.4), no one PC explains a large amount of the variance and the first three PC's explain about 50% of the variance.

Table 3.8 gives the correlation coefficients between the angles and the PC's and explains the loading plots (Figures 3.13 and 3.15). Angles A1 and A2 have a large negative correlation with PC1 (-0.66 and -0.89, respectively) and are major contributors to PC1. Angles B1 (0.49), B2 (-0.47), and B5 (-0.60) are all contributors to PC2. Angle B3 (0.94) is the only major contributor to PC8, whereas A1 (0.55) and B4 (0.45) are contributors to PC7.

Comparison of Table 3.8 (combined data) to Table 3.5 (TP250 data) explains why Figure 3.14 is distinctly different than Figure 3.11. Both figures show the score plots for PC7 vs PC8. The TP250 data (Figure 3.11) separate into nine clusters: three along PC7 and three along PC8, whereas the combined data (Figure 3.14) separate into only three clusters along PC8. Table 3.5 shows that B4 has a very large positive correlation to PC7 for the TP250 data, whereas Table 3.8 shows that B4 and A1 have only a moderate correlation to PC7 for the combined data. Appendix E shows that the combined data does not separate along PC7. But both data sets show a very large positive correlation between B3 and PC8 and both separate into three groups along PC8. Although a large correlation ($\pm$0.9) between an angle and a PC does not guarantee that the data will separate along that PC, it is interesting to note that in all cases in which the various data sets are found to separate along a certain PC, each of those PC's has a $\pm$0.9 correlation coefficient with a particular angle.

Table 3.9 gives the correlation coefficients between the angles. As was seen in the DM324 results (Table 3.3), angles A1 and A2 have a high correlation (0.34). Table 3.9 shows that angles B4 and B5 have the second largest correlation (0.16). This is also

typical of the DM324 data in Table 3.3. This is in contrast to the TP250 results (Table 3.6) where no angle pairs have a large correlation (the largest being B1 and B2 (-0.10)).

Table 3.9 Correlation Coefficients Between all Eight Angles for DM324 and TP250 Data Combined and GEM-Scaled to the DM324 GEM

	A1	A2	B1	B2	B3	B4	B5	B6
A1	1.00	0.34	-0.00	0.02	-0.00	0.01	0.015	0.03
A2	0.34	1.00	0.00	0.03	0.03	0.03	0.03	-0.01
B1	-0.00	0.00	1.00	0.01	0.07	-0.01	-0.09	0.01
B2	0.02	0.03	0.01	1.00	0.02	0.00	0.03	0.01
B3	-0.00	0.03	0.07	0.02	1.00	0.09	-0.03	-0.07
B4	0.01	0.03	-0.01	0.00	0.09	1.00	0.15	-0.00
B5	0.01	0.03	-0.09	0.03	-0.03	0.15	1.00	-0.05
B6	0.03	-0.01	0.01	0.01	-0.07	-0.00	-0.05	1.00

In order to see if the results were independent of what conformer was used as the reference for the scaling procedure, the data was also scaled relative to the TP250 GEM angle values. The results are discussed in the next section.

3.3.2 Results of Scaling to the TP250 GEM Conformer

Figure 3.16 gives the score plot of the combined data scaled to the TP250 GEM for PC1 vs. PC2. The figure is similar to the DM324 data (Figure 3.2) and the combined data scaled to the DM324 GEM (Figure 3.12) in that there are three main clusters that separate along PC1. As in Figure 3.12, the data are skewed relative to the axes and do not cleanly separate along PC1. Similar behavior is seen in the PC1 vs. PC3 and PC1 vs. PC5 score

plots (Appendix F). The three clusters do not separate into the subdivisions seen in Figures 3.2 and 3.12.

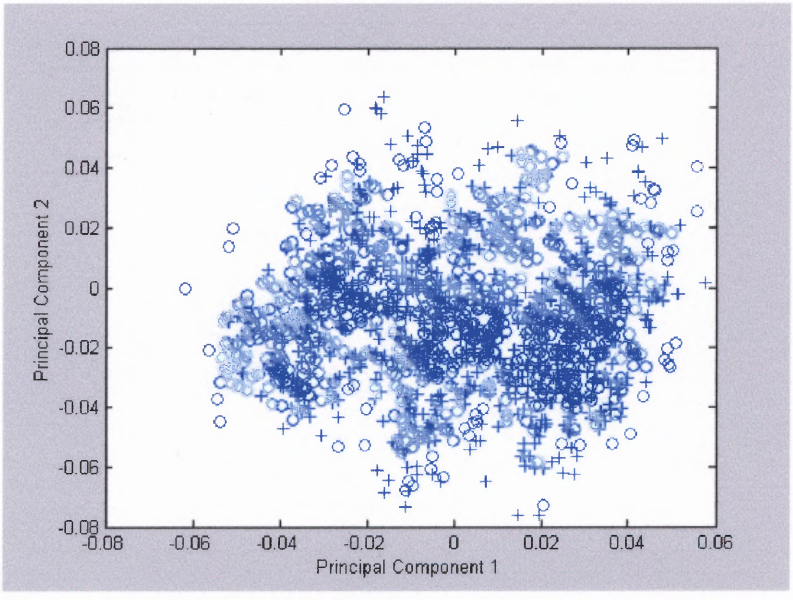


Figure 3.16 Score plot of DM324 and TP250 GEM-scaled to TP250 GEM for PC1 vs. PC2.

Although the score plots appear to be somewhat similar, the loading plot for the combined data scaled to the TP250 GEM (Figure 3.17) is dramatically different from that of DM324 (Figure 3.3) and the combined data scaled to the DM324 GEM (Figure 3.13). (It is also different than the loading plot of the TP250 data (not shown) that corresponds to the PC1 vs. PC2 score plot in Figure 3.6.) Comparison of Figure 3.17 to Figures 3.3 and 3.13 shows that B1 has a large positive correlation to PC1 for the combined data scaled to the TP250 GEM, but a positive correlation to PC2 for the DM324 and combined data scaled to the DM324 GEM. Also, A1 and A2 have large negative correlations to PC2 in the combined data scaled to the TP250 GEM, but large negative correlations to PC1 for the other two data sets. B5 is not a major contributor in the combined data scaled to the TP250 GEM, but has a large negative correlation to PC2 in the other two data sets.

For all three data sets, B2 has a large negative correlation to PC2. So even though all three data sets separate along PC1, the separation is due to B1 in the case of the combined data scaled to the TP250 GEM and due to A1 and A2 for the other two data sets.

In addition, Figure 3.17 shows that B1 has a moderate negative correlation to PC2, while A2, B2 and B6 have small correlations to PC1. This causes a slight skewing of the data so that it does not lie parallel to the PC2 axis and makes it difficult to define "high", "middle", and "low" values of the angles relative to the GEM angles because the data do not separately cleanly along PC1.

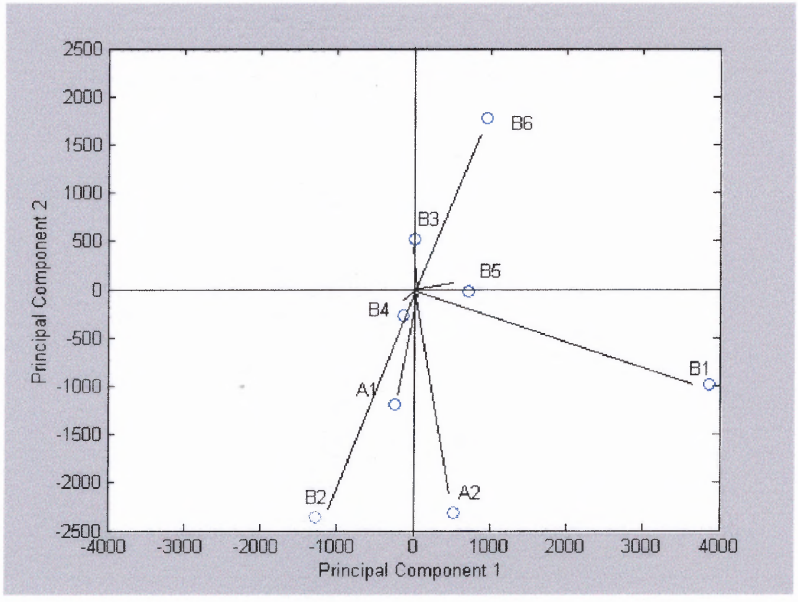


Figure 3.17 Loading plot of DM324 and TP250 combined GEM-scaled to the TP250 GEM.

Figure 3.18 gives the score plot of the combined data scaled to the TP250 GEM for PC7 vs. PC8. The data separates into three groups along the PC8 axis. This plot is typical of the score plot of *any* PC with PC8 for this data set (see Appendix F) in that the data always separates along the PC8 axis. The corresponding loading plot (Figure 3.19) shows that angle B3 is the major contributor to PC8. This is similar to the separation seen

along PC8 due to B3 in the TP250 data (Figures 3.9 and 3.10), as well as the combined data scaled to the DM324 GEM (Figures 3.14 and 3.15).

Comparison of Figure 3.19 (combined data scaled to TP250 GEM) to Figure 3.15 (combined data scaled to DM324 GEM) shows that in both cases B4 is the major contributor to PC7 and B3 is the major contributor to PC8. However, B3 and B4 have positive correlations to PC8 and PC7 for the combined data scaled to the DM324 GEM, but negative correlations for the other combined data set. Also A1 has a large positive correlation to PC7 for the combined data scaled to the DM324 GEM, but not for the other data set.

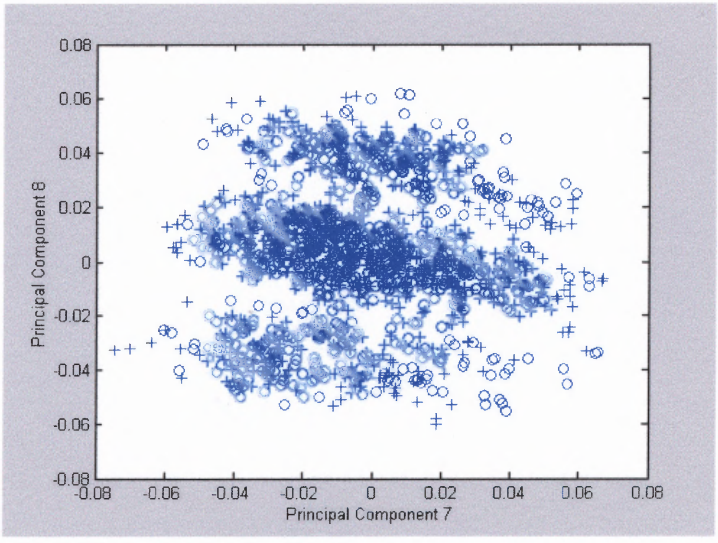


Figure 3.18 Score plot of DM324 and TP250 GEM-scaled to the TP250 GEM for PC7 vs. PC8.

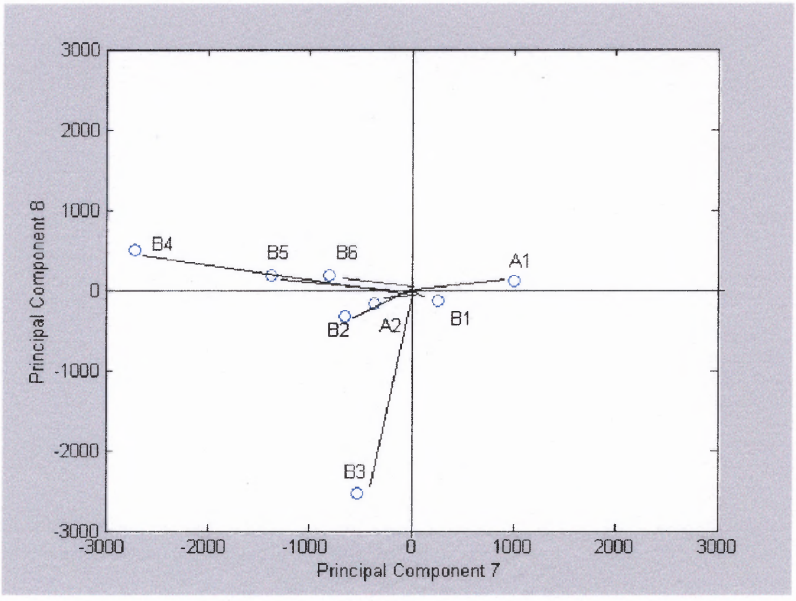


Figure 3.19 Loading plot of DM324 and TP250 data GEM-scaled to the TP250 GEM for PC7 vs. PC8.

Table 3.10 Percentage of Variance Explained by Each PC after SVD of DM324 and TP250 Data Combined and GEM-Scaled to the TP250 GEM

PC	Eigenvalue^2* 10^6	Variance explained	Σ Variance(%)
1	1.8415	16.34	16.34
2	1.6865	14.96	31.30
3	1.5867	14.08	45.38
4	1.5573	13.82	59.20
5	1.4338	12.72	71.92
6	1.2866	11.42	83.34
7	1.1851	10.52	93.86
8	0.6927	6.15	100.01

Table 3.11 Correlation Coefficients Between the Angles and the PC's for DM324 and TP250 Data Combined and GEM-Scaled to the TP250 GEM

Angle	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A1	-0.07	-0.33	-0.07	0.10	0.39	0.80	0.27	0.02
A2	0.13	-0.55	-0.05	0.47	0.50	-0.40	-0.06	-0.07
B1	0.92	-0.21	-0.00	-0.08	-0.28	0.10	0.07	-0.04
B2	-0.34	-0.54	0.58	-0.26	-0.31	0.08	-0.13	-0.11
B3	0.01	0.14	-0.12	0.06	0.06	0.14	-0.22	-0.93
B4	-0.04	-0.04	-0.40	0.29	-0.24	0.30	-0.76	0.12
B5	0.18	-0.01	-0.01	-0.73	0.54	-0.02	-0.36	0.04
B6	0.24	0.46	0.72	0.34	0.21	0.13	-0.20	0.04

Table 3.10 gives the percentage of the variance explained by each of the PC's. As in the case of the DM324 data set, TP250 data set, and the combined data scaled to the DM324 GEM, no one PC explains a large amount of the variance. The first three PC's explain only 45.38% of the variance--a result similar to that of the other data sets.

Table 3.11 gives the correlation coefficients between the angles and the PC's for the combined data scaled to the TP250 GEM. The table explains the loading plots in Figure 3.17 (PC1 vs. PC2) and Figure 3.19 (PC7 vs. PC8). For example, although B1 has a large positive correlation to PC1 (0.92) it lies off the PC1 axis in Figure 3.17, tilted towards the negative PC2 axis because B1 has a negative correlation coefficient (-0.21) with PC2. Table 3.11 shows that B1 is the major contributor to PC1 with correlation coefficient 0.92; A2 (-0.55), B2 (-0.54), and B6 (0.80) contribute to PC2; B4 (-0.76) contributes to PC7 and B3 (-0.93) contributes to PC8. Even though B4 has a large negative correlation to PC7, the data do not separate along PC7 (see Figure 3.18 and

Appendix F). This is in contrast to the TP250 results in Table 3.5 which show that B3 and B4 both have large positive correlations (0.9) to PC8 and PC7, respectively.

Table 3.12 shows that there is a slight positive correlation between angles B3 and B4. This is not seen in the other data sets.

Table 3.12 Correlation Coefficients Between all Eight Angles for DM324 and TP250 Data Combined and GEM-Scaled to the TP250 GEM

	A1	A2	B1	B2	B3	B4	B5	B6
A1	1.00	0.06	-0.01	0.02	0.02	0.01	0.01	-0.02
A2	0.06	1.00	0.02	-0.05	-0.00	-0.00	-0.02	-0.00
B1	-0.01	0.02	1.00	-0.07	-0.01	-0.00	0.05	0.03
B2	0.02	-0.05	-0.07	1.00	-0.06	-0.06	-0.00	-0.01
B3	0.02	-0.00	-0.01	-0.06	1.00	0.12	0.02	0.03
B4	0.01	-0.00	-0.00	-0.06	0.12	1.00	-0.07	-0.06
B5	0.01	-0.02	0.05	-0.00	0.02	-0.07	1.00	-0.02
B6	-0.02	-0.00	0.03	-0.01	0.03	-0.06	-0.02	1.00

In summary, when analyzed together using the DM324 GEM angles as the GEM, the combined data set separates along PC1 due to A1 and A2, as occurred when the DM324 data set was analyzed alone. The TP250 data set, when analyzed alone, does not separate along PC1. The combined data set also separates along PC8 due to B3. This is behavior typical of the TP250 data set, but not of the DM324 data set.

When the analogs were analyzed together using the TP250 GEM angles as the GEM, the combined data set separates along PC1 due to B1 and along PC8 due to B3. Some behavior of the TP250 data set when analyzed alone (separation along PC7 due to

B4) was not observed in the combined data set for either DM324 or TP250 GEM scaling. This indicates that the results of SVD are sensitive to the conformer chosen as the GEM for scaling and suggests that it may be more accurate to analyze the data separately, scaled to each analog's GEM, rather than together, scaled to the GEM of either analog. The application of SVD to the combined data set (using either GEM scaling) showed the similarity of the molecular conformations because the DM324 and TP250 conformers occupy relatively the same regions in the score plots.

Comparison of the score plots and correlation coefficients between angles and PC's for all four data sets shows that every time data separation occurs along a particular PC, there is one angle which has a very large (+0.9) correlation coefficient with that PC. For all four data sets, no one PC explains a significant amount of the variance. Even the first five PC's explain only 75% of the variance. This is why PC7 and PC8 are important to the data analysis; they make up 15% of the data.

CHAPTER 4

DISCUSSION

4.1 The Problem of Circular Data

The objective of this project was to see if SVD could be useful in uncovering the relationship of torsional angles to subtle differences in the conformations of the GBR 12909 analogs, DM324 and TP250. Data separation (i.e. separation of conformers into groups) was obtained by using a novel scaling technique based on defining the torsional angles of each conformer relative to the corresponding angles of the GEM conformer. This was in contrast to median scaling, which failed to lead to data separation. Therefore, the classification of these conformers is very sensitive to the way in which the data is scaled.

The original approach was to median-scale the data since this technique was useful in the analysis of DNA data [1]. Median scaling is obtained by subtracting the median of each angle (A1, A2, B1-B6) from the value of that angle in each conformer of the data set. However, after median scaling the data did not separate along any principal component. The results of median scaling are shown in Appendix A for DM324 and Appendix B for TP250.

GEM scaling was then hypothesized as a better way to address the issue of data circularity. It was applied to the DM324 and TP250 data sets separately by scaling the data to their individual GEM angle values. The data was GEM scaled so that no difference between any angle and the GEM value of that angle was greater than 180° or less than -180°, as was explained in Chapter 2. This was shown to be important to

accurately represent the data because this procedure resulted in a clear separation of the data along certain principal components. The data separation was attributed to the angles that have the largest correlation to those principal components.

This procedure allows one to backtrack to find the actual torsional angle values of a conformer given the conformer's value for any PC, as described in Chapter 3. For instance if a conformer has a value for PC1, the corresponding angle values can be obtained by finding the **X** value through the equation of SVD **X=USV'**, and can then be descaled. Descaling takes into account the fact that all angles are given relative to the GEM.

In order to test the effect of GEM versus median scaling on a well-known data set, a form of GEM scaling was performed on the DNA data (kindly provided by Dr. Ron Wehrens of the University of Nijmegen) to see if it produced results similar to those of median scaling. Since the energies of the DNA conformers were not known, the first conformer in the DNA data set was arbitrarily chosen to be the "GEM". All the angles in the other conformers were scaled relative to the angles in this conformer using the procedure described in Chapter 2. The results of GEM scaling for DNA (see Appendix G) produced better results than the median scaling approach. The data still separate in the same way as with median scaling, but the clusters are easier to distinguish [1]. Since the DNA data consists of four different types of structures (A, B_I, B_{II}, and crankshaft) this suggested that it might be useful to analyze the DM324 and TP250 data *together* by scaling the data relative to an arbitrarily chosen conformer, such as the GEM of DM324 or the GEM of TP250. This procedure was carried out, but the combined data sets did not separate as cleanly as the individual data sets nor as well as the DNA data set. This is not

surprising since DNA can take on four distinctly different shapes, whereas the GBR 12909 analogs can take on a continuum of related shapes. SVD was carried out on the phosphate backbone torsional angles of DNA which are locked into fairly rigid orientations due to the restrictions of the double helix. There are few restrictions on the range of the torsional angles in the GBR analogs. Each torsional angle can take on a range of values between 0° and 360°, limited only by the fact that certain values are not allowed because of poor steric interactions (i.e. interactions of high energy) between the functional groups of the torsional angle. For example it is known that "eclipsed" conformers (where functional groups on either side of the angle are lined up in close proximity) are of higher energy than "staggered" conformers (where functional groups are set as far away from each other as possible, staggered between positions of high energy). So, for example, rotation around a C (sp^3) - C(sp^3) bond will result in three low energy conformers spaced 120° apart. Rotation of the molecule around this bond has three-fold "symmetry". Since the Random Search procedure that produced the GBR analog data sets finds only conformers of low energy, only those conformers constitute the data set. However, since the GBR 12909 analogs are not constrained to a particular molecular framework, such as a double helix, there is a continuum of related conformers available to them.

In summary, the idea of GEM scaling seems appropriate on a chemical level because the torsional angles of the GEM conformer produce the conformer of least energy. It also makes sense on a statistical level because other types of scaling do not take the chemical aspect of the data into account. This novel scaling procedure allows the GEM conformer to act as a circular median.

4.2 Comparison to Fuzzy and Hierarchical Clustering Results

In order to validate the SVD procedure for the analysis of molecular conformations, the results of SVD analysis were compared to the results of fuzzy and hierarchical clustering of the DM324 data set. The DM324 data set was shown in Chapter 3 to separate along PC1 whose major contributors are A1 and A2. Fig. 3.2 shows three major groups, each of which is divided into two parts. Fuzzy clustering of the same DM324 data set was carried out by Milind Misra of the Venanzi group and Amit Banerjee of the Davé group at New Jersey Institute of Technology using the Davé k-means fuzzy clustering algorithm and software. They defined a unique feature vector to analyze the conformers which were first superimposed by four atoms of the central piperazine ring. The results show three main clusters (Figure 4.1), each of which is subdivided into two clusters (Figure 4.2). These results are quite similar to the SVD results (Figure 4.3). In order to be able to visually compare the SVD results to the fuzzy clustering results, the conformers in the three figures were superimposed in the same way as above (by four atoms of the piperazine ring). However, it should be noted that since the SVD technique uses only the values of the torsional angles, the results are independent of how the conformers are superimposed. This is in contrast to the fuzzy and hierarchical clustering techniques which give results that are sensitive to the way in which the conformers are superimposed

From the structure of DM324 (Figure 1.4) it can be seen that the three clusters are due to the approximate three-fold rotational symmetry around the C (sp^3) - C(sp^3) bond of the A1 torsional angle. Similarly, each large cluster is divided in two due to the approximate six-fold rotational symmetry around C (sp^3) - C(sp^2) bond of the A2 torsional angle. Figure 4.3 is a representation of the six clusters obtained with SVD.

Although the figure is not identical to Figure 4.2 the results are quite similar. There are three major groups with two subdivisions.

These results also correlate very well with the hierarchical clustering of the DM324 conformers performed by Kathleen Gilbert of the Venanzi group using the Xcluster module of the Macromodel program (available from Schrödinger, Inc., New York, NY). Appendix H compares the cluster memberships found by all three techniques. The major similarities are the following. Fuzzy clustering produced three main groups containing 229, 270, and 229 conformers each. Hierarchical clustering also found three main groups of sizes 221, 235, and 262. All of the conformers in the 262-member hierarchical group are also found in the 270-member fuzzy group. The members of the SVD groups were chosen by inspection of the score plot in Figure 3.2. Conformers were assigned to three groups based on the following PC1 ranges: -0.08 to -0.03, -0.03 to 0.03 and 0.03 to 0.08. Some arbitrariness was involved in the assignment of conformers with PC1 values of 0.03 and -0.03. Nonetheless, SVD found three groups of sizes 182, 234 and 312. In the first group of 182 members, 142 conformers are common to both the 262-membered hierarchical group and the 270-membered fuzzy group. Furthermore, SVD found 620 conformers in total agreement with both the fuzzy and hierarchical clustering group assignments. The other 108 conformers are either in common with the fuzzy *or* the hierarchical groups. However, these discrepancies could be due to the somewhat arbitrary cutoffs that were applied to the PC1 values in order to classify the conformers into three distinct groups.

Figure 4.4 is a representation of the nine well-defined clusters found for TP250 (Figure 3.11). Members of each cluster were defined by inspection of Figure 3.11. The

nine groups were defined by the ranges of PC7 and PC8 as defined in Chapter 3. Specifically, the (PC7,PC8) ranges are: (-0.08 to -0.03, 0.03 to 0.08), (-0.03 to 0.03, 0.03 to 0.08), (0.03 to 0.08, 0.03 to 0.08), (-0.08 to -0.03, -0.02 to 0.03), (-0.03 to 0.03, -0.02 to 0.03), (0.03 to 0.08, -0.02 to 0.03), (-0.08 to -0.03, -0.08 to -0.03), (-0.03 to 0.03, -0.08 to -0.03), and (0.02 to 0.08, -0.08 to -0.03). In order to be able to visualize the clustering, the conformers were superimposed by the atoms that define angles B3 and B4. In viewing Figure 4.4 (where the A side of the molecule is not shown for clarity) it would help to keep in mind that a "cluster" member consists of the whole bis-phenyl moiety. So each cluster will occupy a region such that one phenyl ring is in, say, position X and the other phenyl ring is in position Y. While two clusters might have overlapping X positions, their Y positions are different and will be clearly observable. This is true for all combinations of two clusters in that figure. For example, consider the BLUE and ORANGE clusters; it is clear that BLUE and ORANGE overlap in one region but not in the other. Since no fuzzy or hierarchical clustering has yet been carried out on the TP250 data set, no comparison can be made. The nine clusters are due to the approximate three-fold rotational symmetry around the C (sp^3) - C(sp^3) bond in the B3 torsional angle combined with the approximate three-fold rotational symmetry around the C (sp^3) - C(sp^3) bond in B4.

Figure 4.1. Fuzzy clustering of DM324 on the "A" side shows three groups (figure provided by Milind Misra). Conformers are superimposed by central piperazine ring. Only atoms involved in the A1 and A2 torsional angles are shown. The view is looking down the A1 torsional angle towards the central piperazine ring.

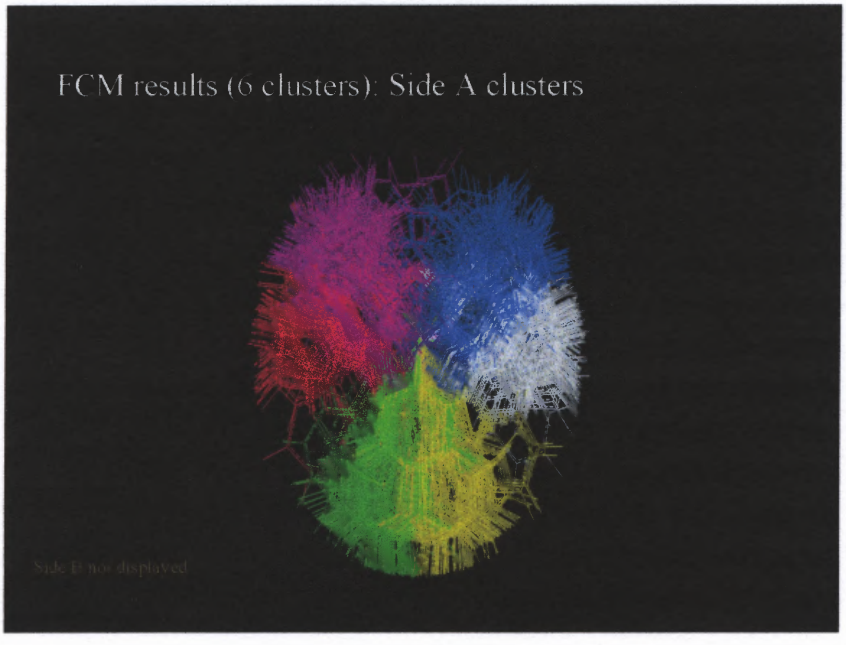


Figure 4.2 Fuzzy clustering of DM324 on the "A" side shows three main groups each subdivided into two (figure provided by Milind Misra). Same view as Figure 4.1

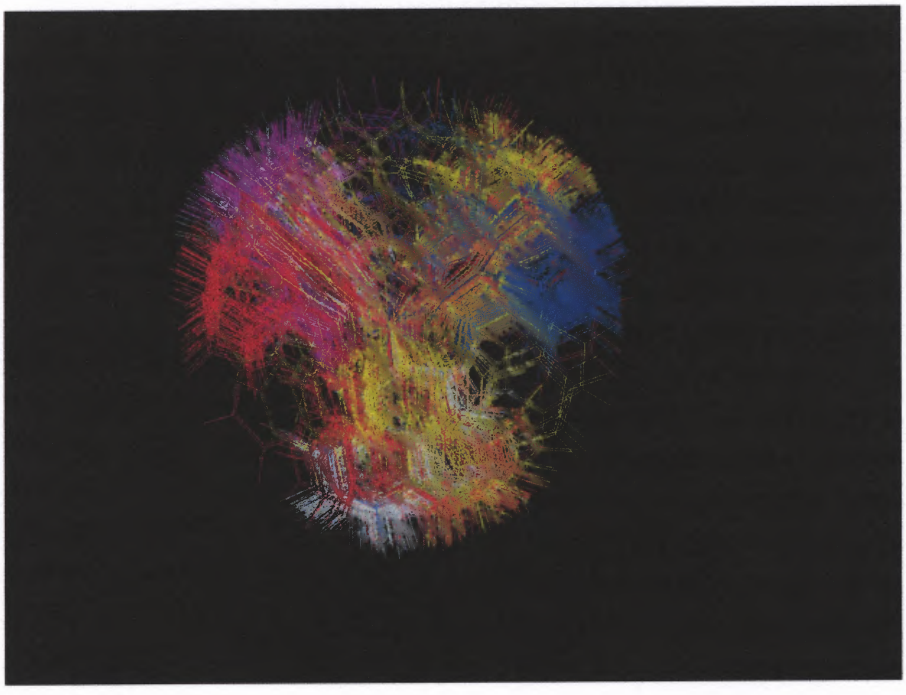


Figure 4.3 Singular value decomposition of DM324 shows three major groups with two subdivisions. Same view as Figure 4.1.

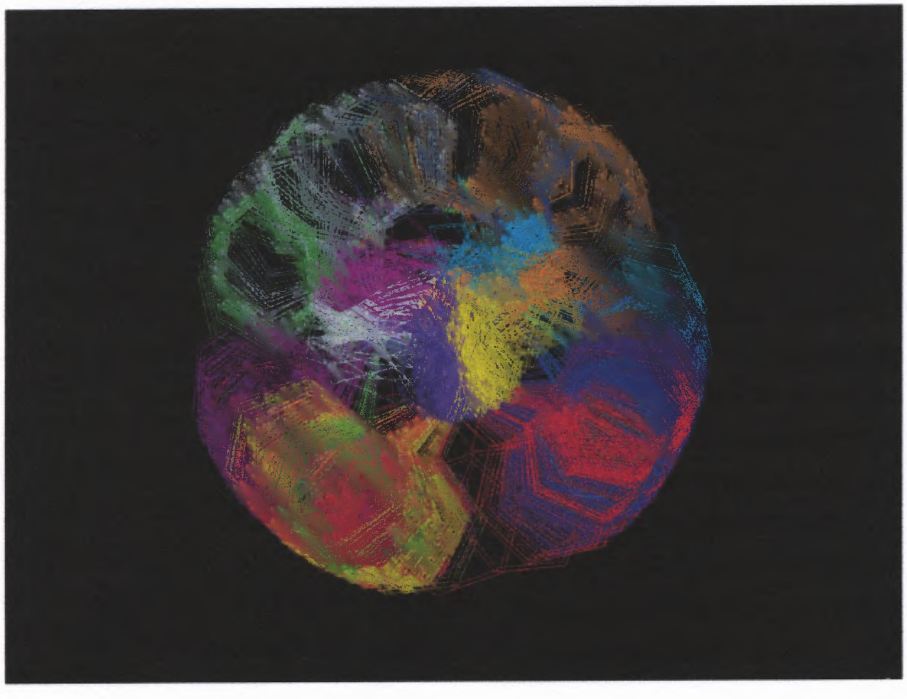


Figure 4.4 Singular value decomposition of TP250 shows nine clusters. Only atoms of the B side are shown. Conformers are superimposed on B3 and B4.

4.3 Comparison of Score and Angle Plots

In order to validate the SVD results, the score plots were compared to plots of the "raw" data (with angles given between a 0° and 360°) and the GEM-scaled data (with angles scaled relative to the GEM conformer of each data set) in Figures 4.5-4.7. Neither the raw data nor the GEM-scaled data plots are identical to the score plots because they show the data plotted in two-dimensional angle space. The score plots, in contrast, show the data plotted in two-dimensional principal component space, where each principal component consists of contributions from all eight angles. However, since the data separate along particular principal components and only certain angles contribute significantly to these PC's, some similarities can be seen in the raw data, GEM-scaled data, and score plots.

To show that SVD reproduces the known behavior of the data for DM324, the SVD score plot of PC1 vs. PC2 (Figure 3.2) was compared to the raw data plot of A2 vs. A1 (Figure 4.5). Figure 4.5 shows three main divisions of conformers along the A1 axis for A1 approximately equal to 60°, 180°, and 300°. This illustrates the approximate three-fold rotational symmetry around A1. The data is much more spread out along the A2 axis. The intermediate region with A1 equal to 100°-150° and 300°-360° and A2 equal to 0°-360° is mainly unoccupied and correspond to regions of high energy due to steric hindrance. The SVD score plot (Figure 3.2) shows regions of space that are also empty. These regions represent those same angles. This was proven by looking at the regions that are unoccupied in PC1 vs. PC2 and using MATLAB to find the corresponding A1 and A2 values. The regions that are not occupied by any DM324 conformer have PC1 values around -0.03 and 0.03. These correspond to A1 values of 100°-150° and 300°-360° values with A2 equal to 0°-360°. It is also evident from Figure

3.2 that there are some conformers that are in these areas, but for the most part these regions are unoccupied. Therefore SVD reproduces known behavior of the DM324 data.

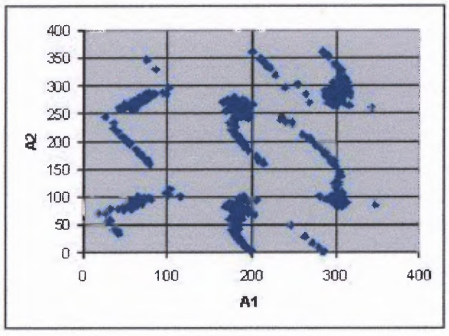


Figure 4.5 Raw data plot of DM324 plotted against A1 and A2.

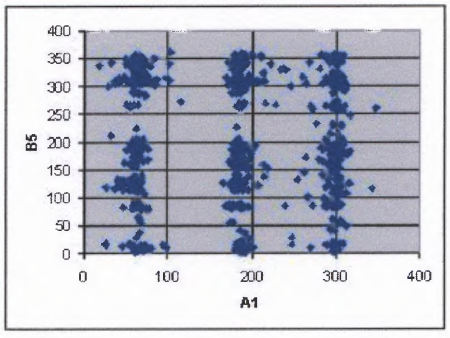


Figure 4.6 Raw data plot of DM324 plotted against A1 and B5.

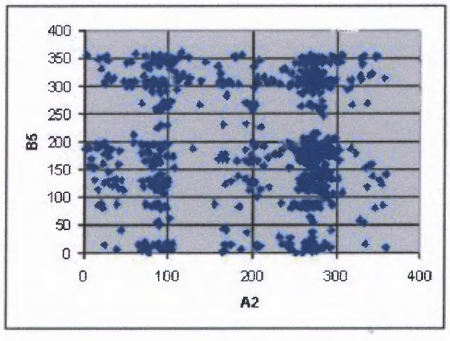


Figure 4.7 Raw data plot of DM324 plotted against A2 and B5.

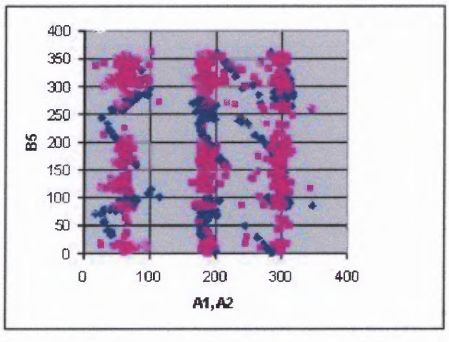


Figure 4.8 Raw data plot of DM324 plotted against A1,B5 and A2,B5.

Since the separation of the DM324 data along PC1 is due to both A1 and A2, a type of angle plot which better represents the relationships in the score plot (Figure 3.2) is given in Figures 4.6-4.8. Here PC2 is represented by B5, the angle which has the largest correlation coefficient with PC2 (Table 3.2). PC1 is represented by A1 (Figure 4.6), A2 (Figure 4.7), or both (Figure (4.8). Figure 4.6 shows that the data clearly separate along A1 at approximately 60°, 180°, and 300°, again illustrating repeating pattern of conformational minima which determine the approximate the three-fold symmetry around A1. Figure 4.7 shows that the data separate along A2, approximately at A2 equal

to 40°, 80°, 160°, 200°, 280°, and 320°. The conformations of low energy are found in three groups in the range 0°-120°, 120°-240°, and 240°-360°. These three regions are subdivided in two groups each at approximately 40° and 80°, 160° and 200°, and 280° and 320°. This illustrates the sterically-allowed values of A2 and corresponds to a very rough six-fold rotational symmetry around A2. Conformers of high energy would be found at A2 values of approximately 0°, 120°, and 240°. These are eclipsed conformers that are of high energy due to the poor steric interactions. Since the Random Search procedure locates energy minima, few conformers are found in this region. Figure 4.8 combines the plots from Figures 4.6 and 4.7. The x-axis represents either A1 or A2. The subtle division into six groups noted in Figure 4.7 along the A2 axis is overlayed by the three major groups along the A1 axis that are centered between the regions 40°-80°, 160°-200°, and 280°-320°, obscuring the six clusters. This is similar to the way the score plot of Figure 3.2 represents the data. The same pattern is seen in the GEM-scaled plots in Figures 4.9 through 4.12. These plots are given relative to the A1, A2, and B5 angles taken as zero.

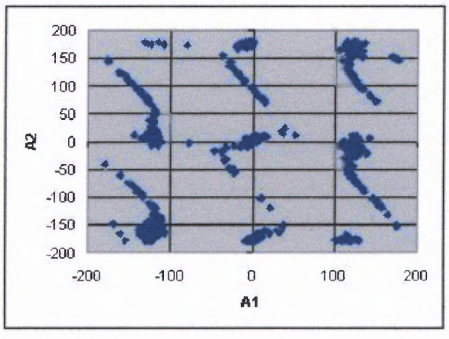


Figure 4.9 GEM-scaled data plot of DM324 plotted against A1 and A2.

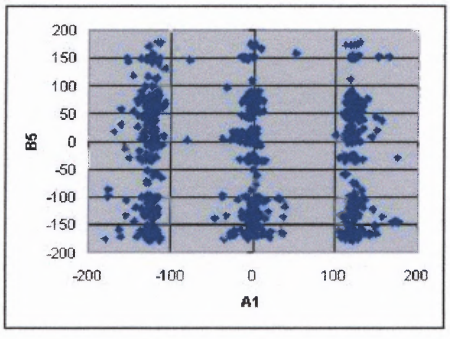


Figure 4.10 GEM-scaled data plot of DM324 plotted against A1 and B5.

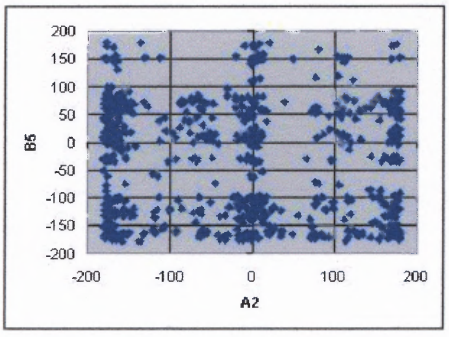


Figure 4.11 GEM-scaled data plot of DM324 plotted against A2 and B5.

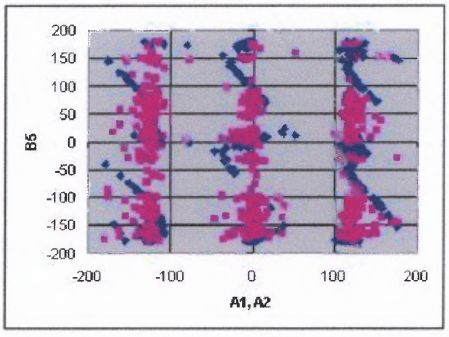


Figure 4.12 GEM-scaled data plot of DM324 plotted against A1, A2 and B5.

Figures 4.13 and 4.14 plots the TP250 raw (Figure 4.13) and GEM-scaled (Figure 4.14) data that correspond to the score plot in Figure 3.11. Since the TP250 data separate along PC7 (due to B4) and PC8 (due to B3) and since both angles have large correlation coefficients to their respective PC's (Table 3.5), plotting the data in (B3, B4) space gives a fairly direct comparison to the score plot in (PC8, PC7) space.

The score plot for PC7 vs. PC8 (Figure 3.11) shows nine clusters. The raw data plot for TP250 (Figure 4.13) also shows nine clusters centered at (B3, B4) values of approximately (60^{o}, 60^{o}), (60^{o}, 180^{o}), (60^{o}, 300^{o}), (180^{o}, 60^{o}), (180^{o}, 180^{o}), (180^{o}, 300^{o}), (300^{o}, 60^{o}), (300^{o}, 180^{o}), and (300^{o}, 300^{o}), and illustrate the three-fold rotational symmetry around both B3 and B4. The empty spaces represent combinations of torsional value of B3 and B4 not found in any of the low energy conformers. They correspond to regions of high energy due to poor steric interactions. A few stray conformers fill some unoccupied spaces but mainly these spaces can be considered "forbidden" to the conformer.

Figure 4.14 presents the GEM-scaled TP250 data plotted in (B3, B4) space. The clusters represent combinations of torsional values of B3 and B4 relative to their GEM values taken as zero. Figure 4.14 is similar to Figure 4.13.

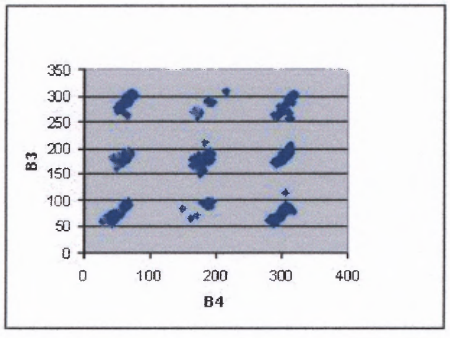


Figure 4.13 Raw data plot of TP250 plotted against B4 and B3.

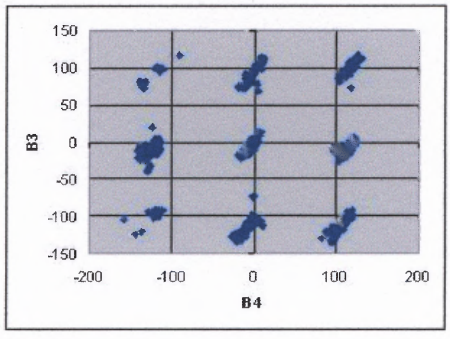


Figure 4.14 GEM-scaled data plot of TP250 plotted against B4 and B3.

A check was made of the descaling procedure in the following way. A conformer was selected from each of the nine clusters in the score plot in Figure 3.11. The data was descaled to see to which cluster that conformer belonged on the raw data plot of Figure 4.13. Using the notation of Chapter 3, the clusters in each figure were numbered consecutively across from the upper left (Group 1) to the upper right (Group 3) and so on down to the lower right (Group 9). Each conformer was found in the same group in both the figures. For example conformer number 326 is in Group 5 on both plots. It has B3 and B4 values similar to the GEM and is found in the same cluster as the GEM on both plots. Conformer 157 is found in Group 4 in both plots. It has a B3 value similar to that of the GEM, but a B4 value smaller than the GEM's B4 value. Conformer 558 is in Group 7 in both figures. This conformer has a value of 84.8° and 149.8° for B3 and B4, respectively, both of which are less than those of the GEM. Conformer 473 is in Group 8 on both plots. This conformer has a value of 63° and 281° for B3 and B4, respectively. This conformer has a value less than that of the B3 of the GEM but similar to the B4 of

the GEM. Conformer 322 has a B3 value of 300.6° and a B4 value of 317.5°. This conformer's B3 value is much larger than that of the GEM, but its B4 value is similar to that of the GEM. This conformer falls into Group 2 in both plots. Conformer 727 has a B3 value of 297.8° and a B4 value of 73.8°, both of which are larger than the GEM (in a clockwise direction). This conformer is in Group 3 in both plots. Conformer 151 has a B3 value of 181.7° which is similar to that of the GEM, and a B4 value of 61.5° which is larger than that of the GEM in a clockwise direction. This conformer falls into Group 2 in both plots. Conformer 684 has a B3 value of 73° which is much less than that of the GEM and a B4 value of 55.6° which is greater than that of the GEM. This conformer falls in Group 9 on both plots.

4.4 Evaluation of Combined Data Analysis

The purpose of analyzing the data together was to see if the analogs separate from each other. The data was combined into one large matrix and analyzed in two different ways. The first approach scaled all the data to the GEM angles of analog DM324. The second approach scaled all the data to the GEM angles of analog of TP250. The first approach produced three major groups with both DM324 and TP250 conformers occupying all three groups. Instead of the three groups separating along one axis, these groups were slightly slanted with respect to PC1 and PC2. The major contributors to PC1 were still A1 and A2; however the TP250 data now separated along PC1. This did not occur when TP250 was analyzed alone nor did it occur when the data were combined and the data scaled to the GEM of analog TP250. The first approach also showed the data separating

along PC8, where the major contributor was B3. This was not observed when DM324 was analyzed separately.

The second approach also produced three major groups along PC1 (due to B1, instead of A1 and A2) and PC8 (due to B3). Furthermore, B4 was still the major contributor of PC7 but did not contribute heavily to the separation of the data. In contrast to the first approach, angles A1 and A2 were not responsible separating the data along PC1. This is similar to the results of the TP250 data when analyzed separately. However, when the DM324 data were analyzed separately, it did not separate on the "B" side of the molecule. Therefore it seems that when analyzed together, the behavior of the analogs is influenced by the GEM analog to which they were scaled. Therefore this leads to the conclusion that DM324 truly separates on the "A" side (due to angles A1 and A2), while TP250 truly separates on the "B" side (due to angles B3 and B4).

CHAPTER 5

CONCLUSION

The purpose of applying SVD to two analogs of GBR 12909 was to attempt to uncover the relationships of torsional angles to the subtle differences in the conformations of the analogs. SVD was demonstrated to be able to separate conformers of analogs into clusters once a novel scaling technique based on the GEM conformer was introduced to treat the problem of circular data. The clusters were defined by torsional angles that have the largest correlation to the principal components which separate the data. Analog DM324, the piperazine analog, showed separation along PC1 when the data was examined in a score plot in (PC1, PC2) space. The major contributors to PC1 are angles A1 and A2, which represent the "A" side of the molecule. Separation of the DM324 data along A1 and A2 was noted when the data were plotted in (B5, A1) and (B5, A2) torsional angle space, where B5 is the chief contributor to PC2. For both the raw and GEM-scaled data, angle A1 separated the data into three large clusters (due to three-fold rotational symmetry around A1), whereas angle A2 separated the data into six smaller clusters (due to approximate six-fold rotational symmetry around A2). The conformers identified by SVD to be in the three large clusters were found to be quite similar to the cluster memberships determined by fuzzy and hierarchical clustering.

Analog TP250, the piperadine analog, showed separation along PC7 (due to B4) and PC8 (due to B3). These angles represent the "B" side of the molecule. Each score plot involving PC7 or PC8 showed three clusters, corresponding to three-fold rotational symmetry around either B3 or B4. Therefore, nine clusters were obtained when the data

were plotted in (PC7, PC8) space. Nine clusters were also obtained when the data were plotted in (B4, B3) torsional angle space using either the raw or GEM-scaled data.

Representative structures from regions of both the DM324 and TP250 score plots were chosen to validate the results of SVD against the raw data plots. The SVD results were shown to be consistent with the raw data plots in defining which regions of torsional angle space can be occupied by the low energy conformers and which regions are forbidden.

Singular Value Decomposition uncovered a difference in DM324 compared to TP250. In DM324 angles A1 and A2 separate the data into clusters, whereas in TP250 angles B3 and B4 separate the data into clusters. It is not obvious why this should be true, given the similarity of the two molecular structures (Figures 1.4 and 1.5). This result may be an artifact of scaling each data set to each analog's GEM torsional angles. Additional studies were carried out in which the GEM of each analog was used to scale the data. In both cases the data separated along PC8 (due to B3). For the combined data scaled to the DM324 GEM, the data also separated along PC1 (due to A1 and A2). For the combined data scaled to the TP250 GEM, the data also separated along PC1 (due to B1) Therefore, when the TP250 GEM was used to scale the data, separation only occurred on the "B" side. This proved to be useful to validate the results of analyzing the data separately.

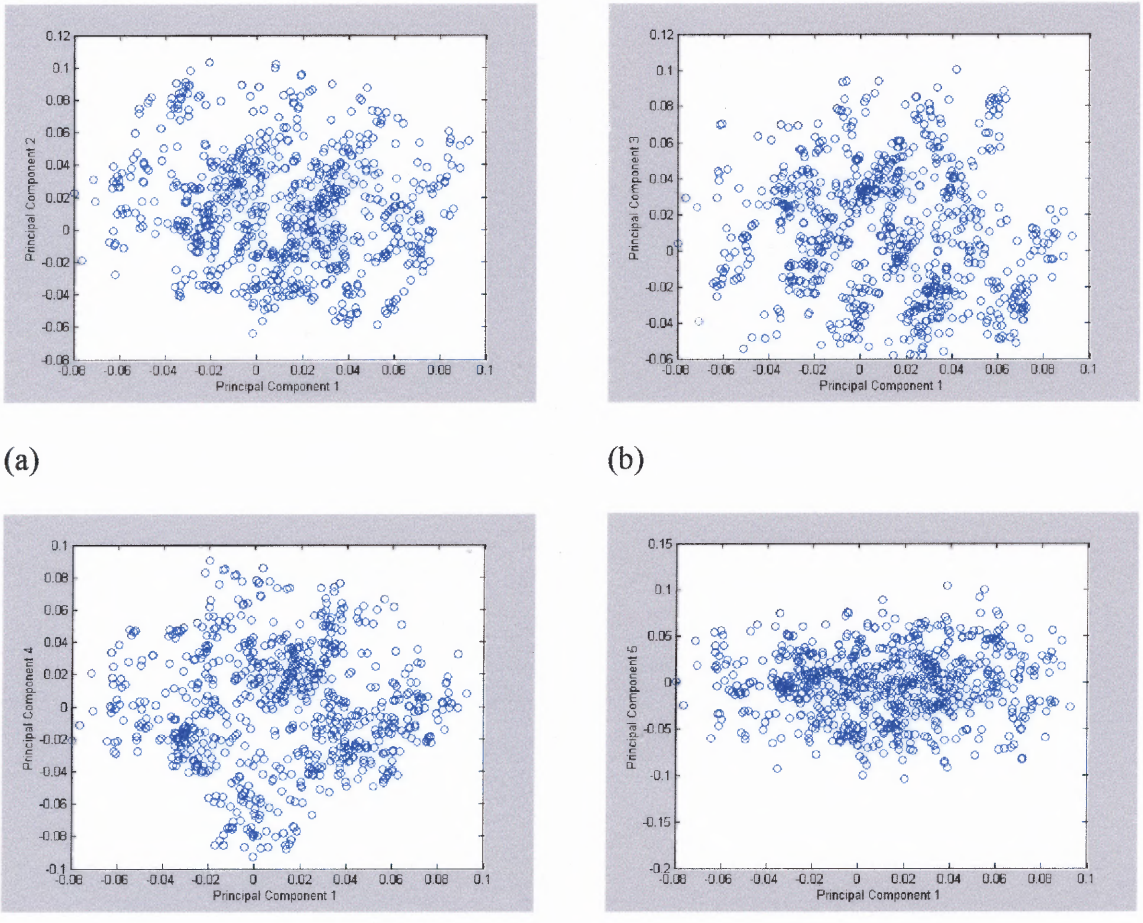
Taken together, the results seem to indicate that DM324 truly separates on the "A" side, while TP250 truly separates on the "B" side. Therefore when the DM324 and TP250 data are analyzed separately, one is able to see the subtle differences between the analogs; when the data are analyzed together, the similarity of the analogs is apparent.

The significance of this work lies in the development of a novel scaling technique for circular data and in the identification of clusters containing sets of molecular conformations. The present work is the first application of SVD to the analysis of very flexible molecules, such as the GBR 12909 analogs. In the future, representative conformations of these analogs will be used in pharmacophore modeling with the ultimate goal of designing a drug useful in the treatment of cocaine abuse.

APPENDIX A

MEDIAN-SCALED DATA FOR DM324

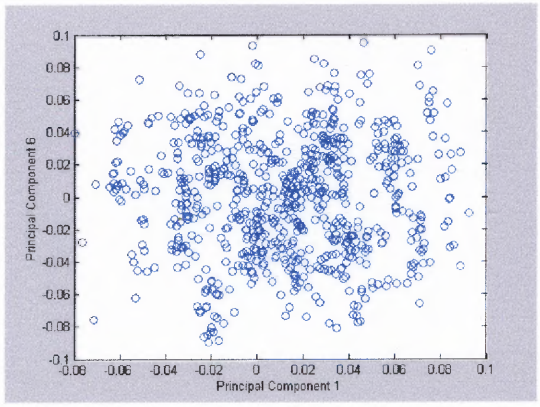
The following figures are the result of median scaling the data for analog DM324. The DM324 data set is contained in the following research directory:
/afs/cad/research/chem/venanzi/3/fuzzy/backup/planesNo/dm324_aligned.mdb
The MATLAB program which median scales the data and creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/MATLAB programs/med/DM324/runsvdscoresDM324Nov11.m



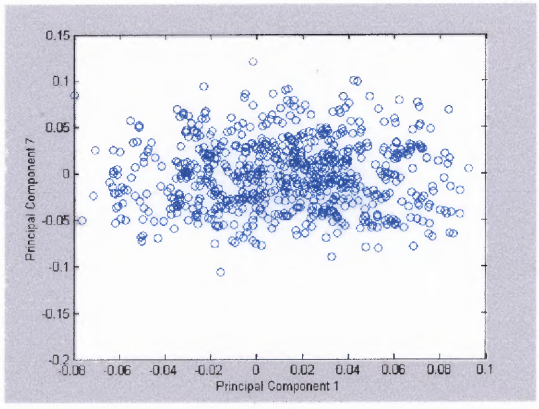
(a) (b)

(c) (d)

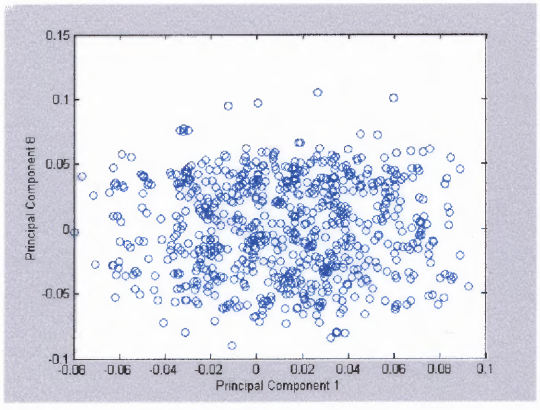
Figure A.1 Score plots for median-scaled DM324 data.



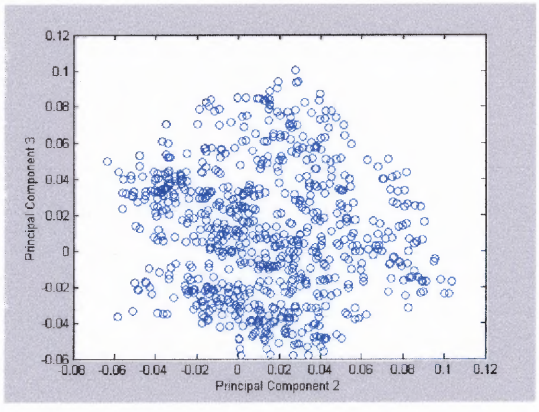
(a)



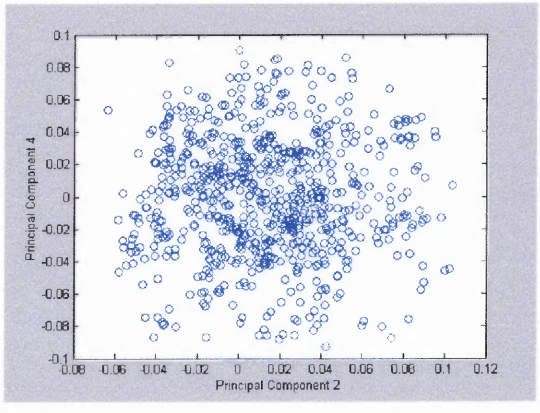
(b)



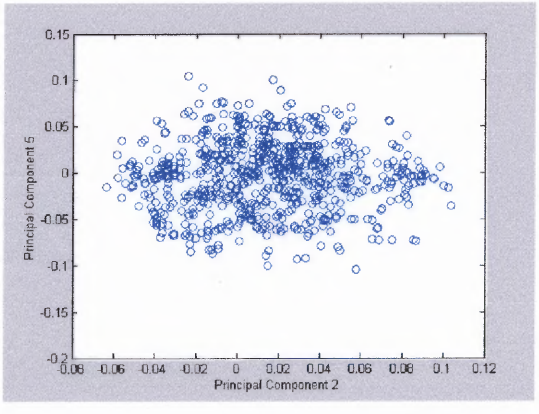
(c)



(d)

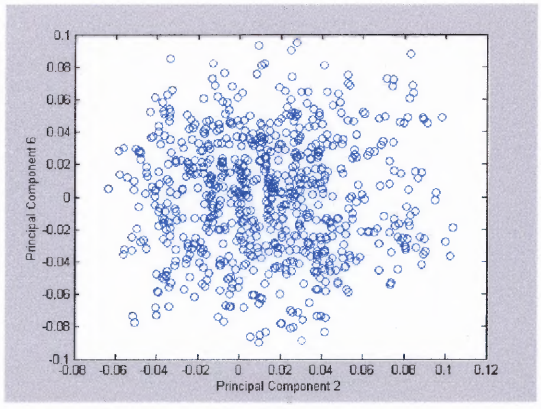


(e)

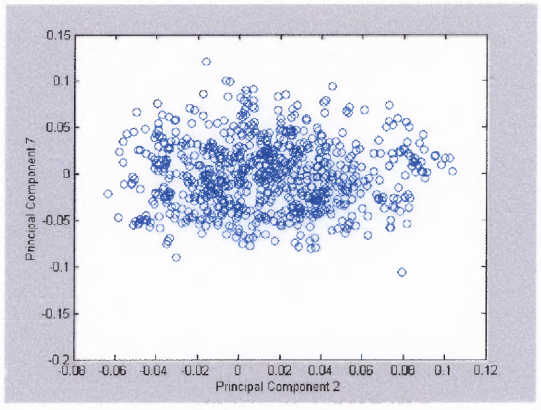


(f)

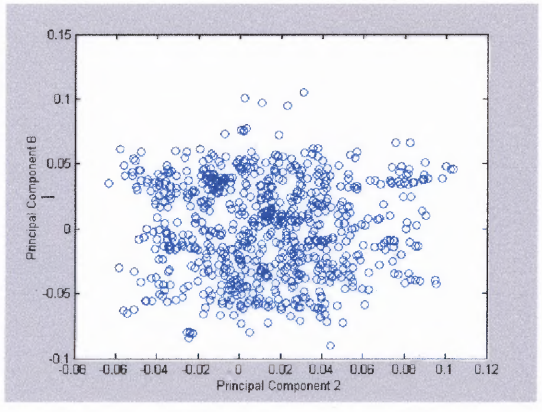
Figure A.2 Score plots for median-scaled DM324 data.



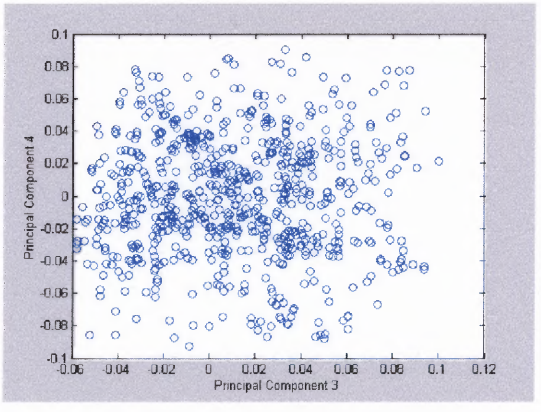
(a)



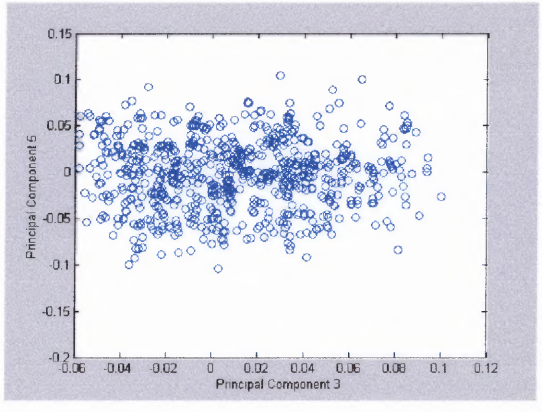
(b)



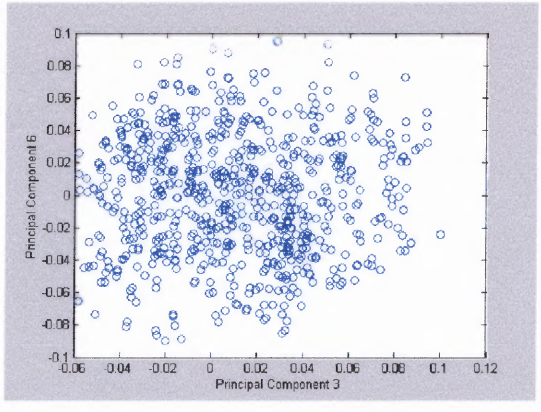
(c)



(d)

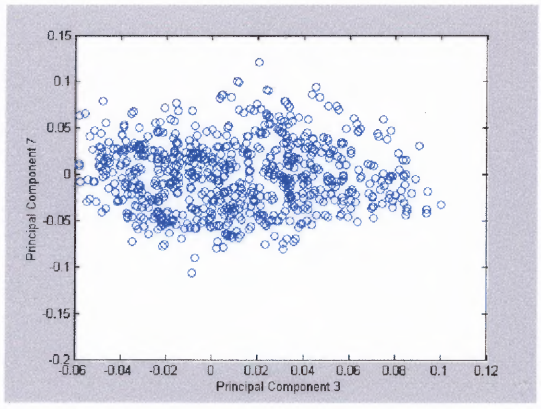


(e)

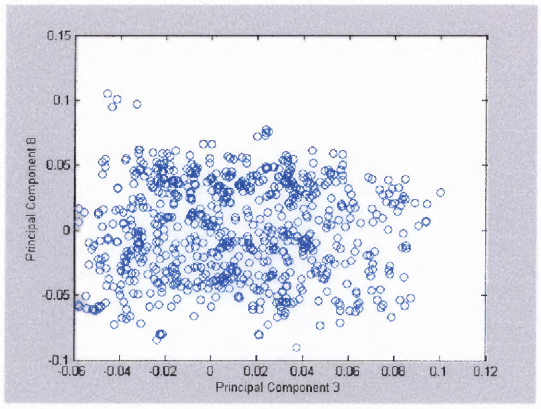


(f)

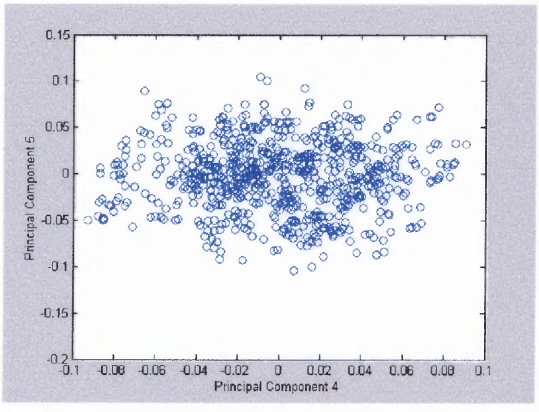
Figure A.3 Score plots for median-scaled DM324 data.



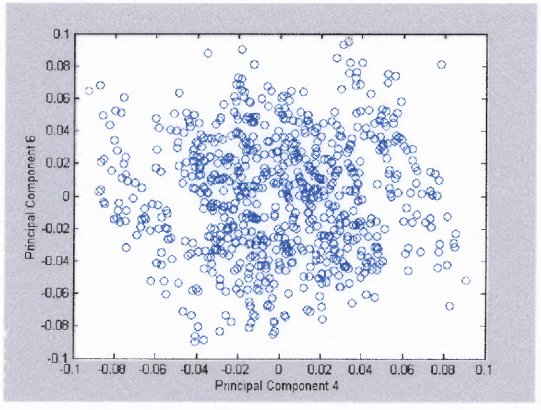
(a)



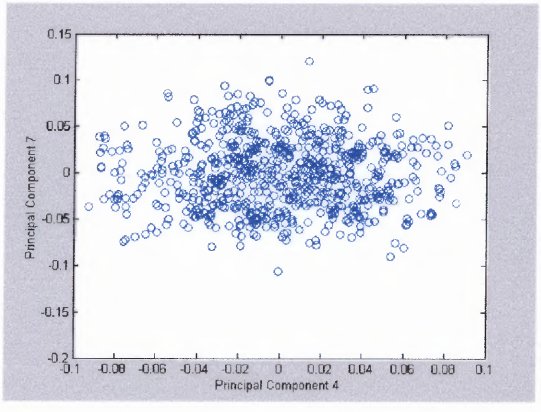
(b)



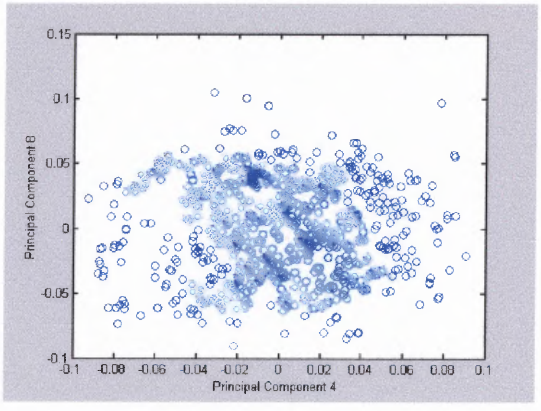
(c)



(d)

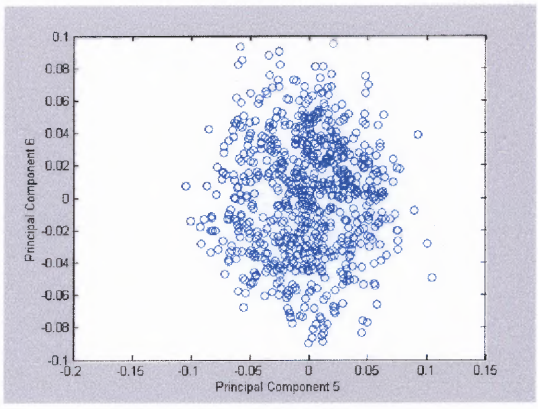


(e)

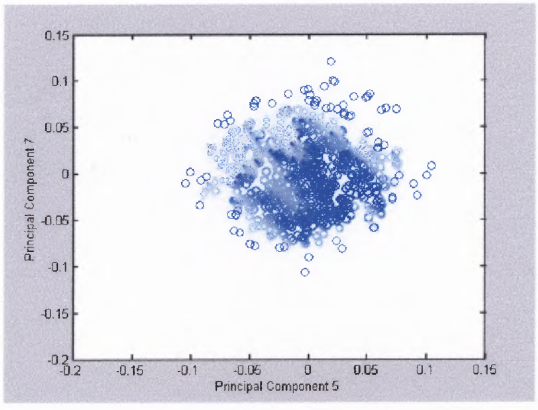


(f)

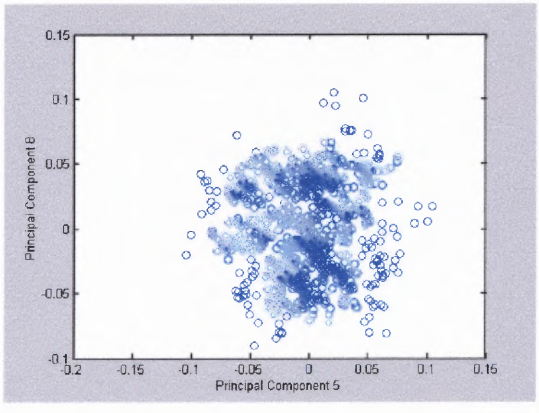
Figure A.4 Score plots for median-scaled DM324 data.



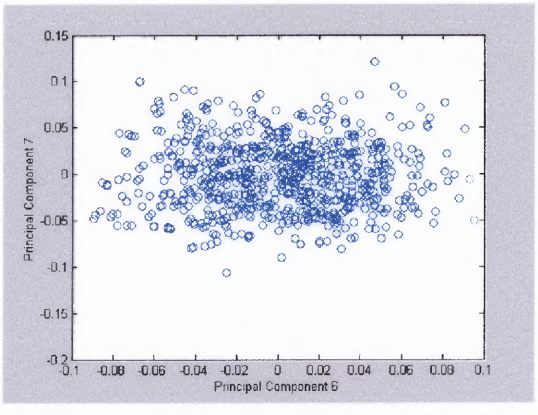
(a)



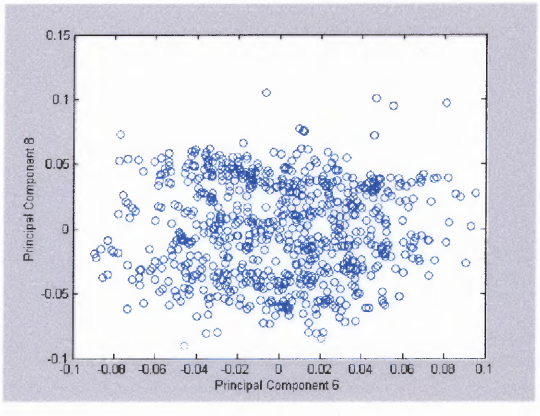
(b)



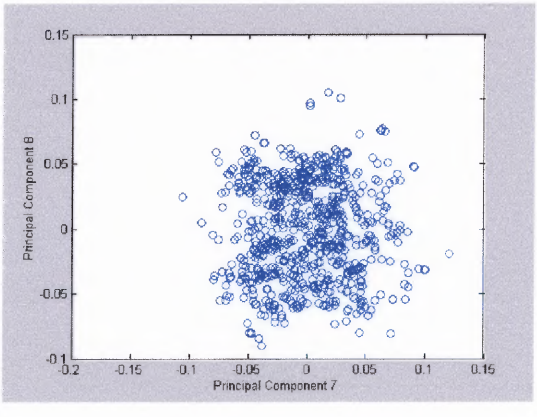
(c)



(d)



(e)



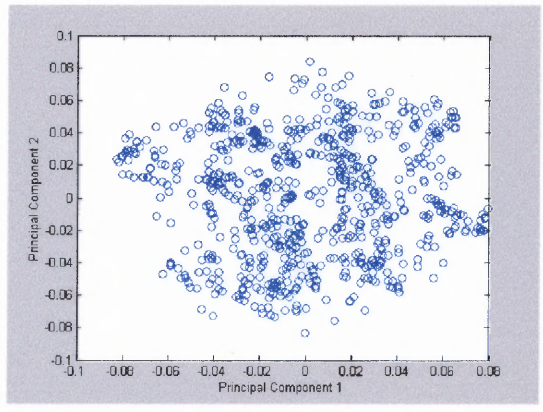
(f)

Figure A.5 Score plots for median-scaled DM324 data.

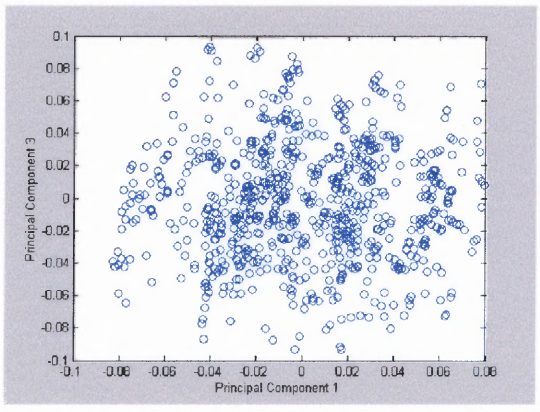
APPENDIX B

MEDIAN-SCALED DATA FOR TP250

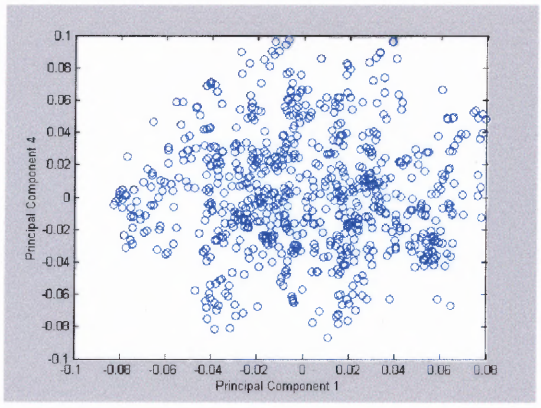
The following figures are the result of median scaling the data for analog TP250.
The TP250 data set is contained in the following research directory:
/afs/cad/research/chem/venanzi/2/eq_eq/prt_tp250_new_.mdb
The MATLAB program which median scales the data and creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/med/TP250/runsvdscoresTP250Nov11.m
The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/med/TP250/TP250Nov11.mat



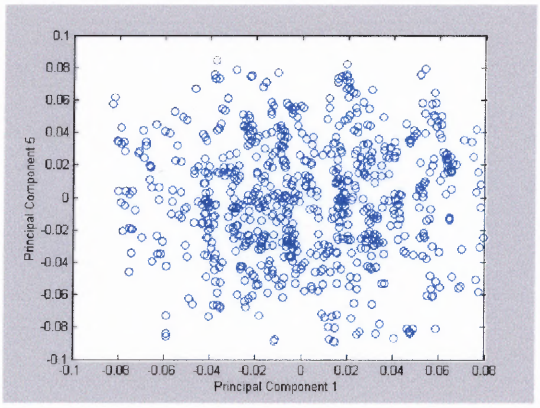
(a)



(b)

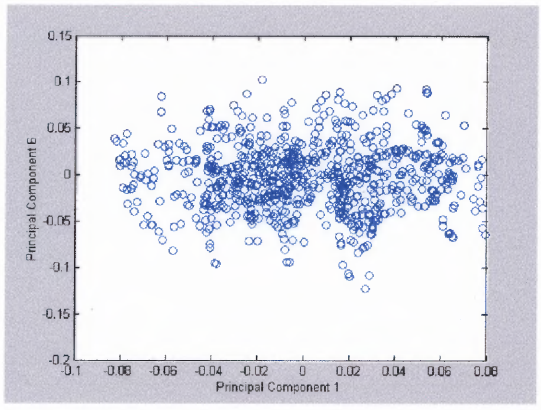


(c)

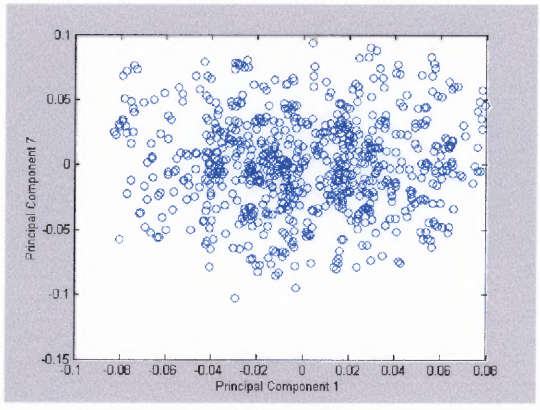


(d)

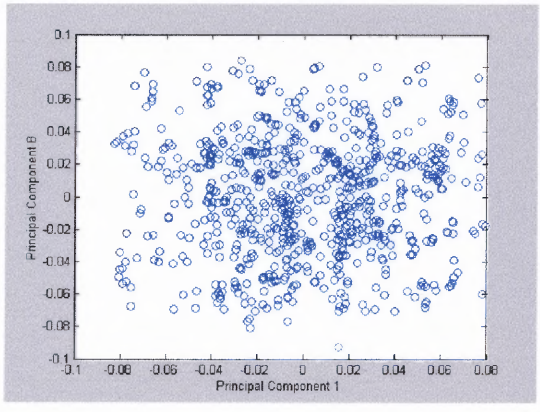
Figure B.1 Score plots of median-scaled TP250 data.



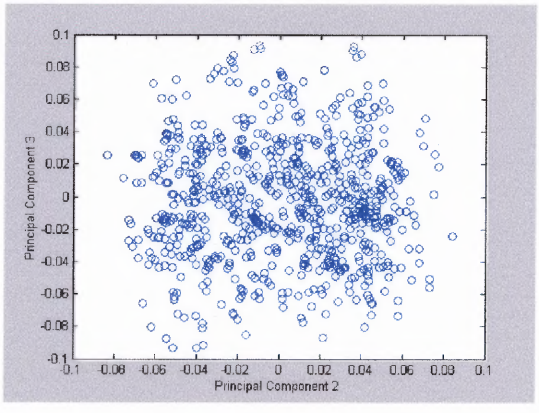
(a)



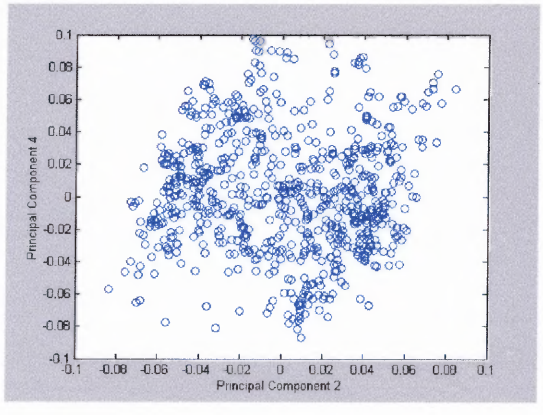
(b)



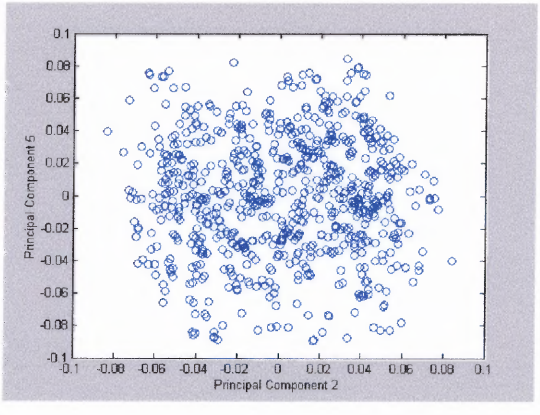
(c)



(d)

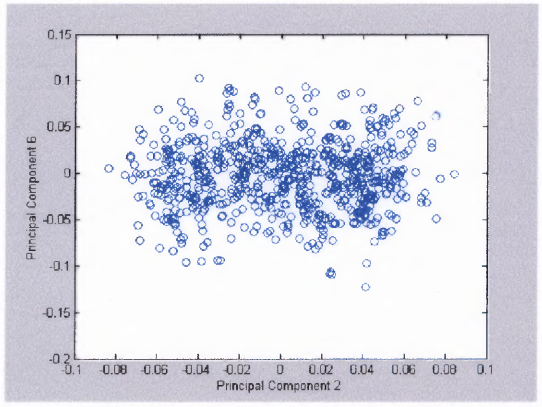


(e)

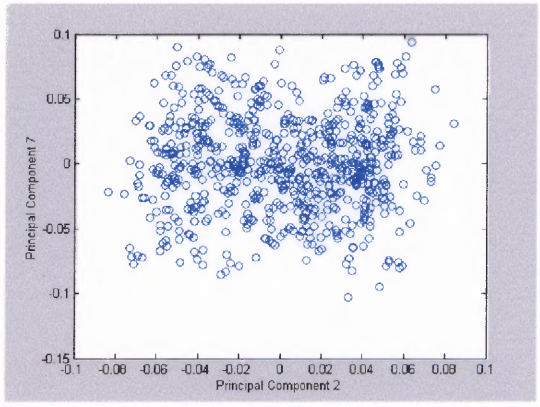


(f)

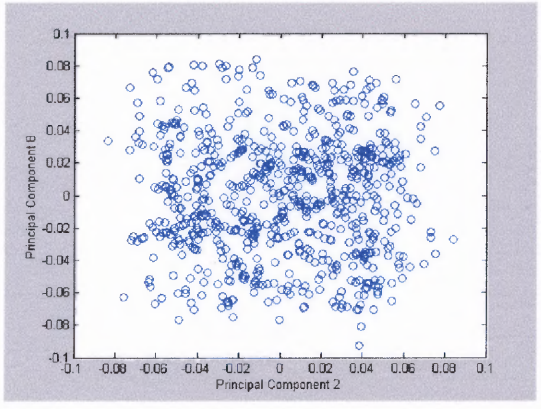
Figure B.2 Score plots of median-scaled TP250 data.



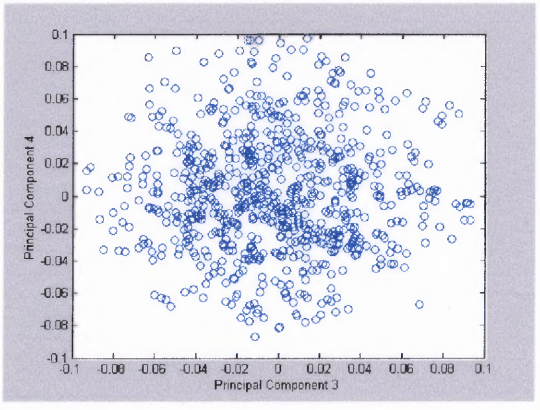
(a)



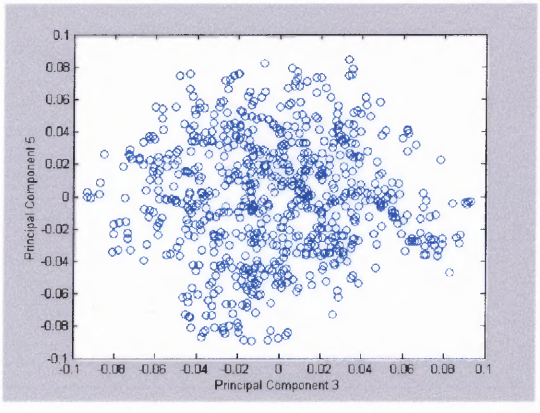
(b)



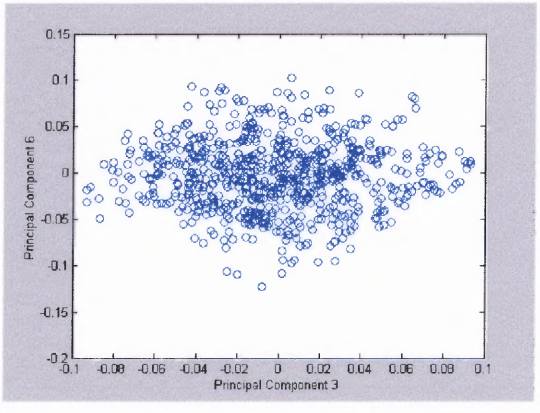
(c)



(d)

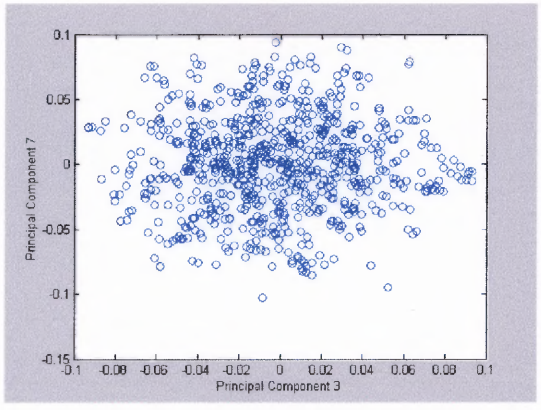


(e)

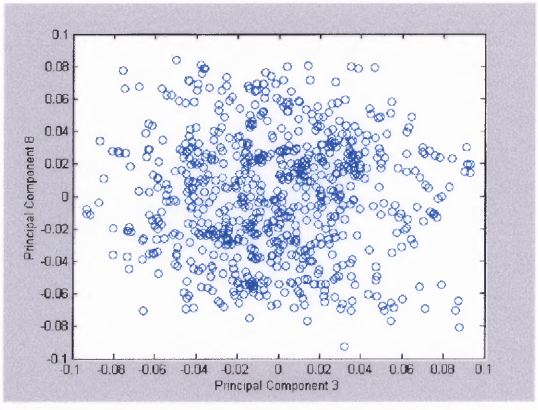


(f)

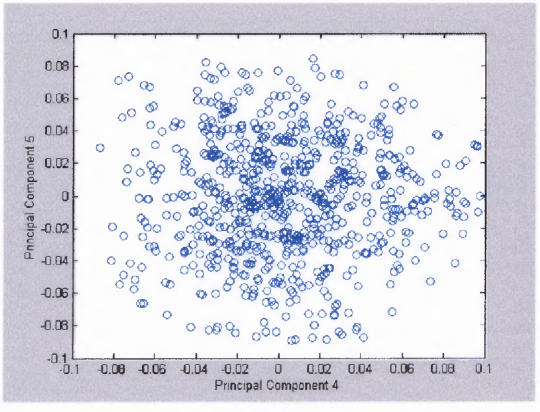
Figure B.3 Score plots of median-scaled TP250 data.



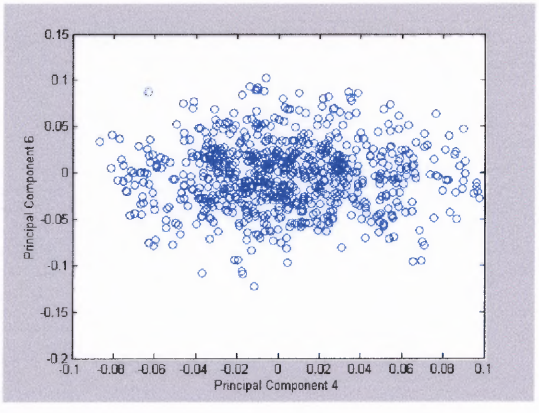
(a)



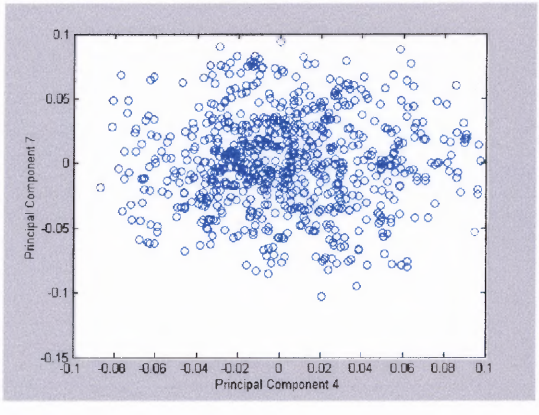
(b)



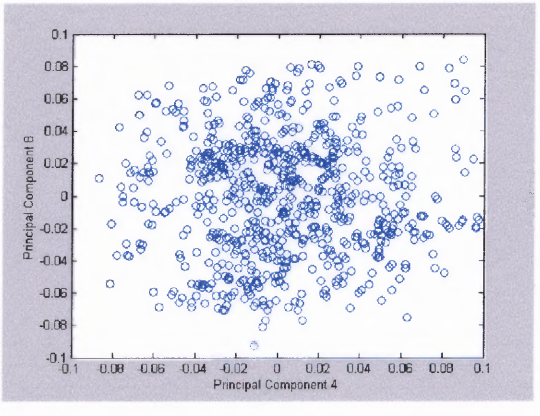
(c)



(d)

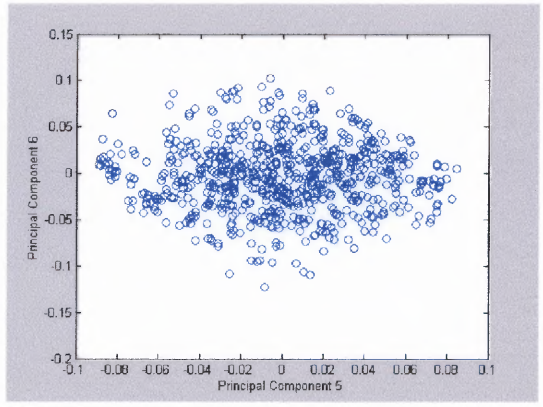


(e)

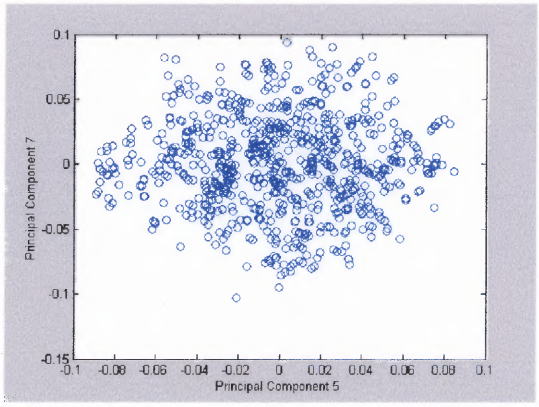


(f)

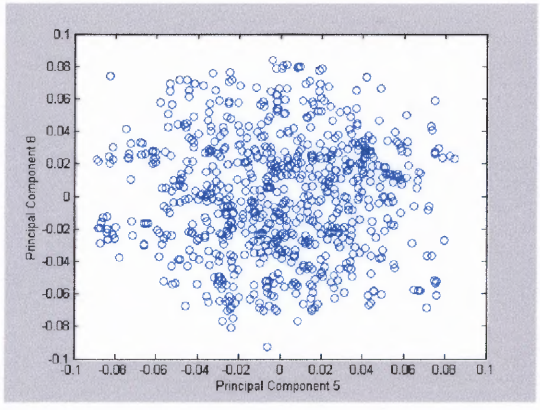
Figure B.4 Score plot of median-scaled TP250 data.



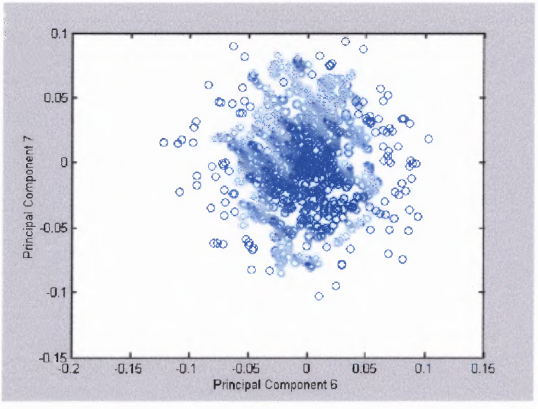
(a)



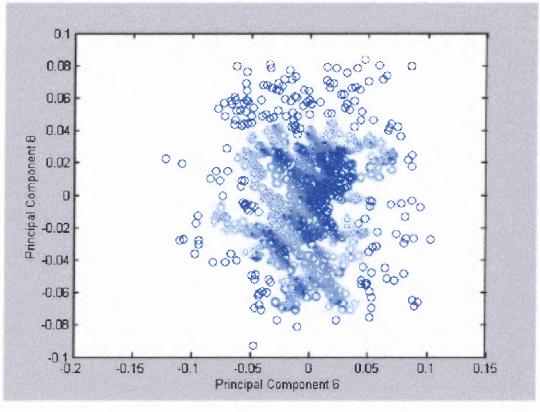
(b)



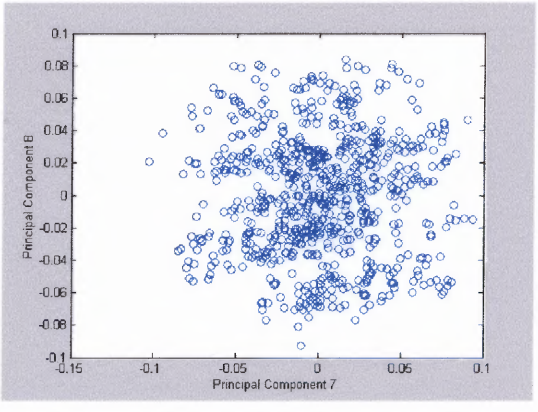
(c)



(d)



(e)



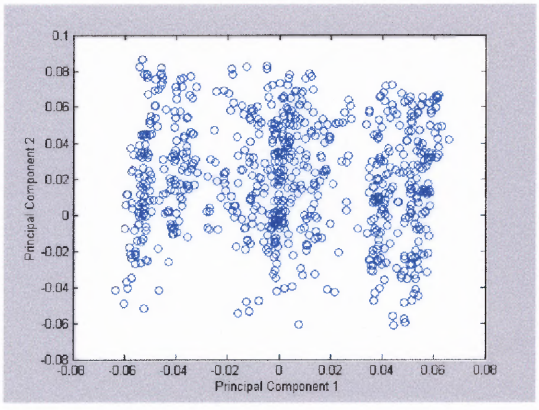
(f)

Figure B.5 Score plot of median-scaled TP250 data.

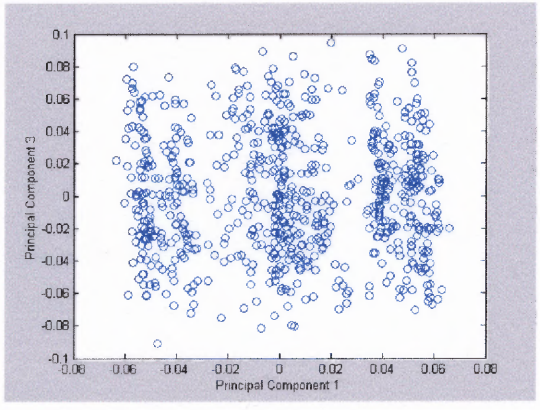
APPENDIX C

GEM-SCALED DATA FOR DM324

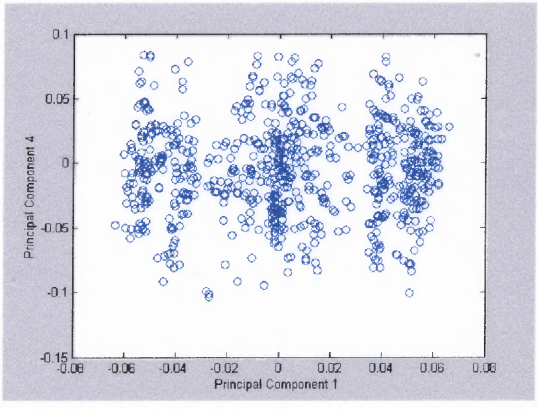
The following figures are the result of GEM scaling the data for analog DM324.
The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/gem/DM324/runsvdscoresDM324March.m
The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/gem/DM324/March11circularDM324.mat



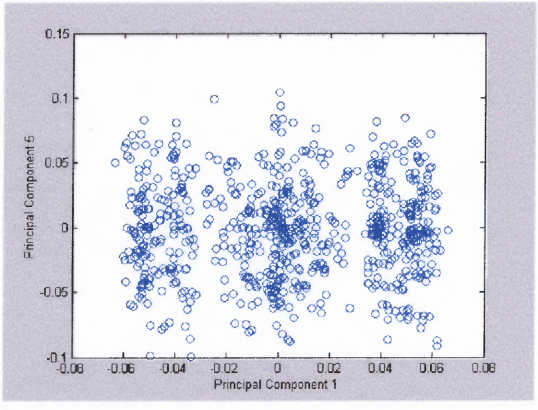
(a)



(b)

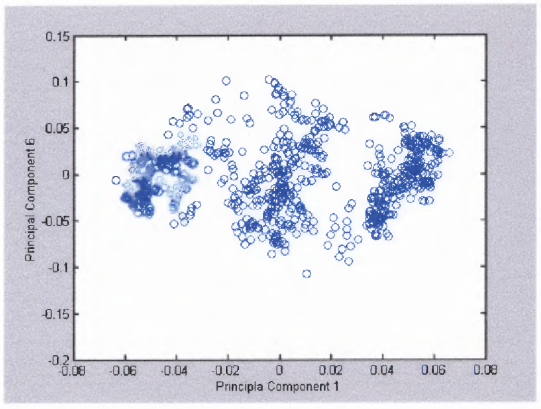


(c)

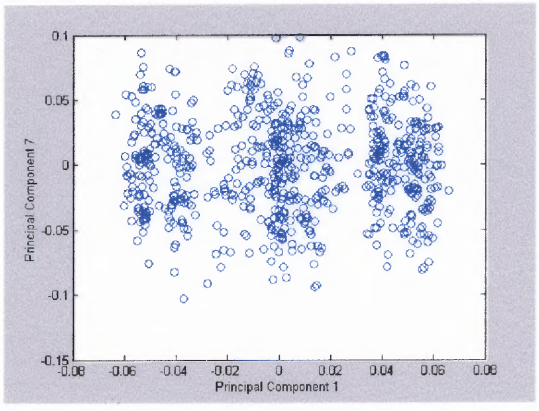


(d)

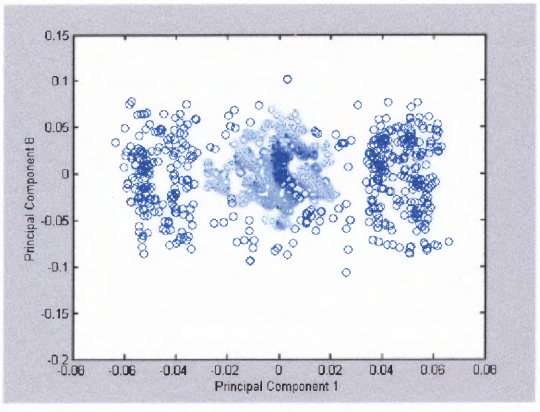
Figure C.1 Score plots of GEM-scaled DM324 data.



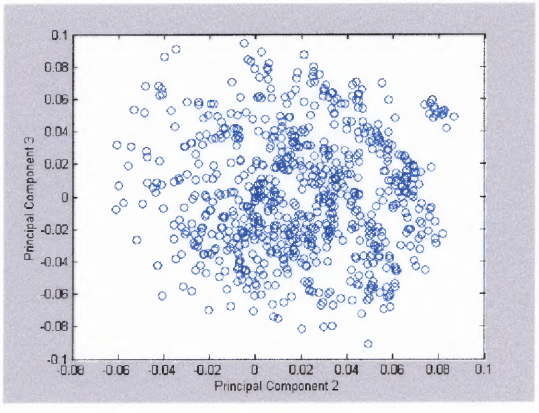
(a)



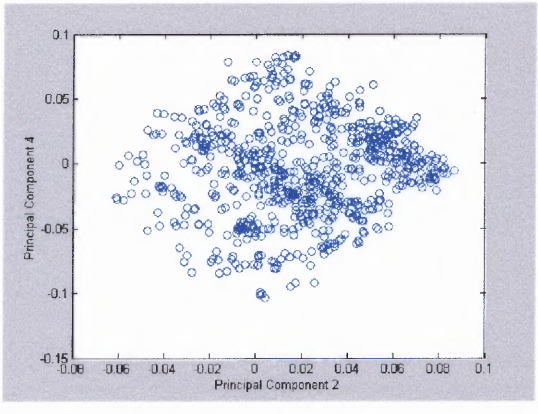
(b)



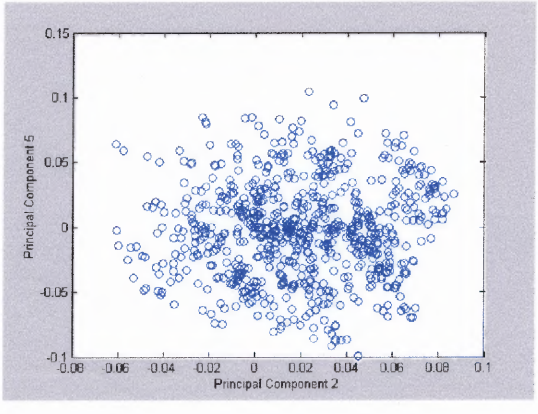
(c)



(d)

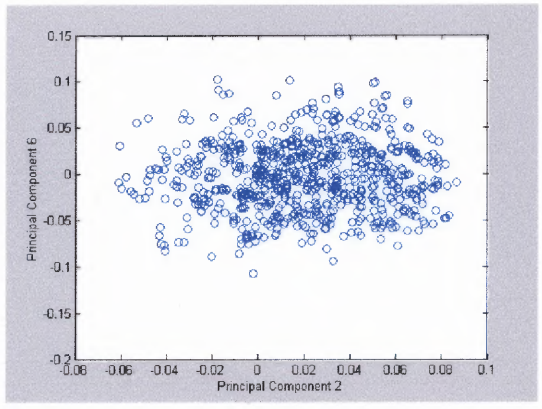


(e)

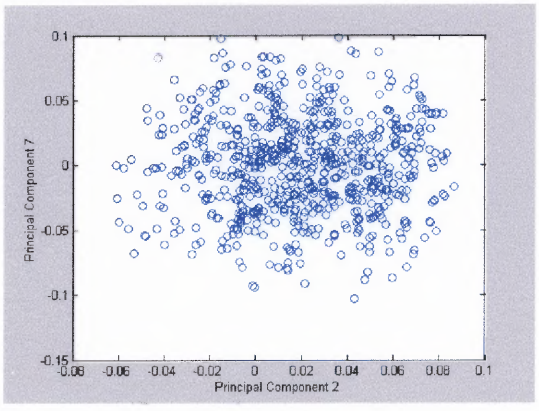


(f)

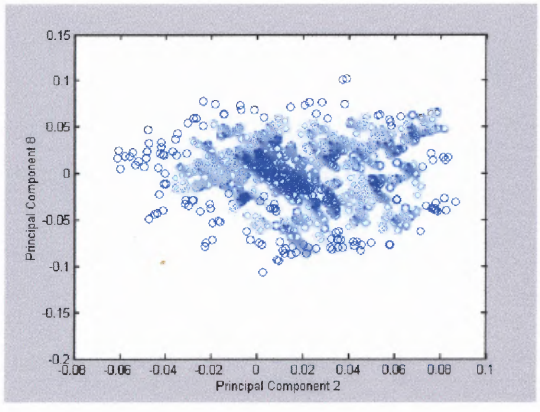
Figure C.2 Score plots of GEM-scaled DM324 data.



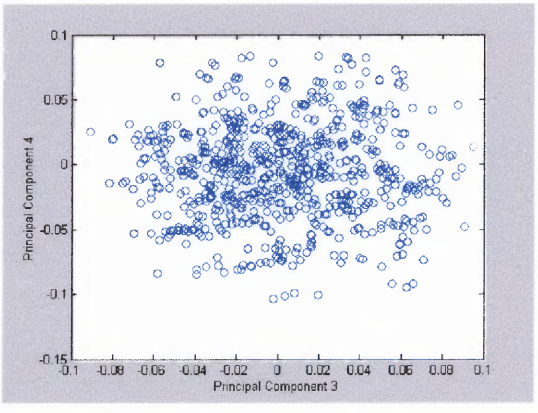
(a)



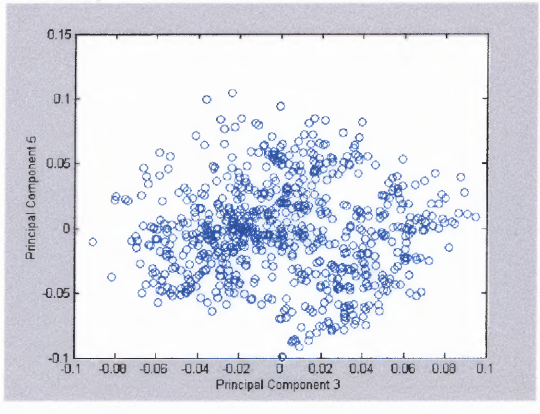
(b)



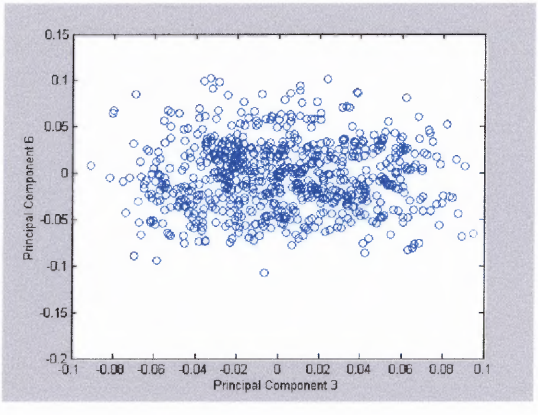
(c)



(d)

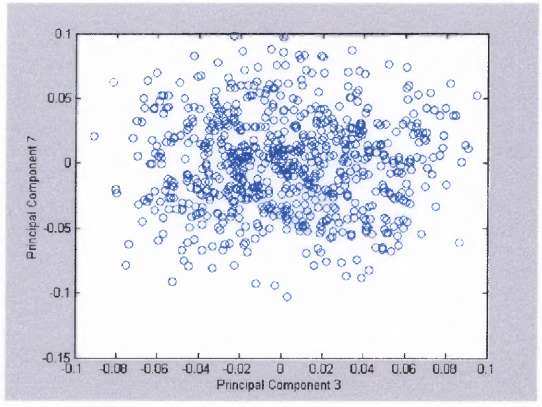


(e)

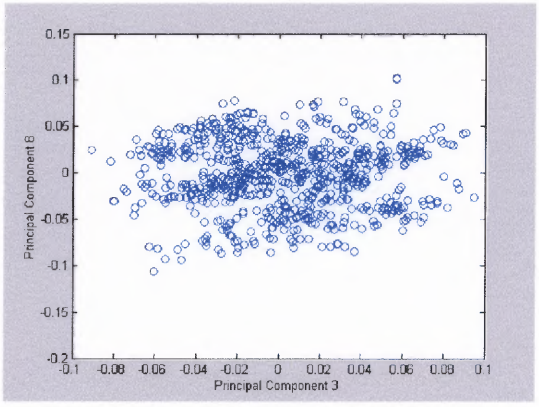


(f)

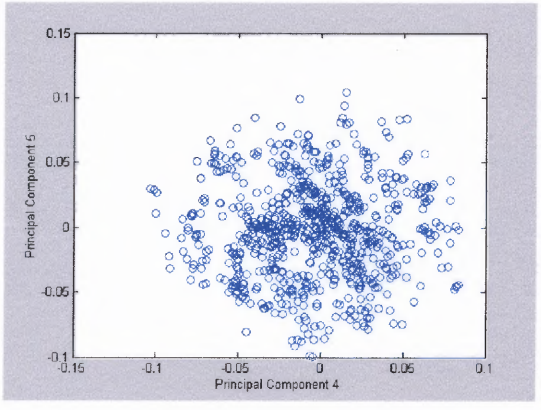
Figure C.3 Score plots of GEM-scaled DM324 data.



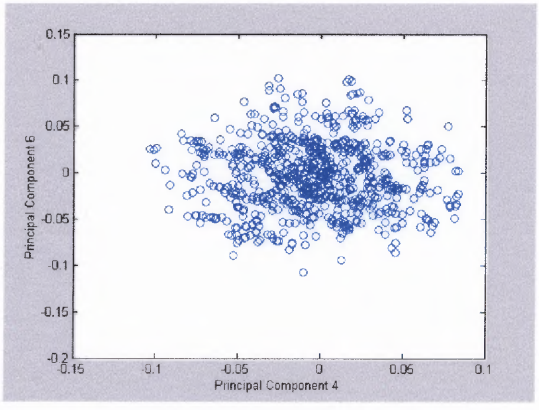
(a)



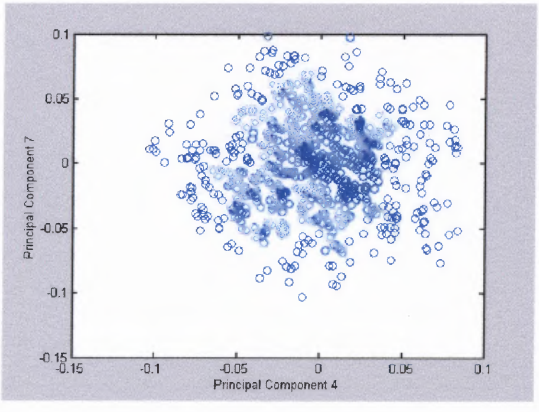
(b)



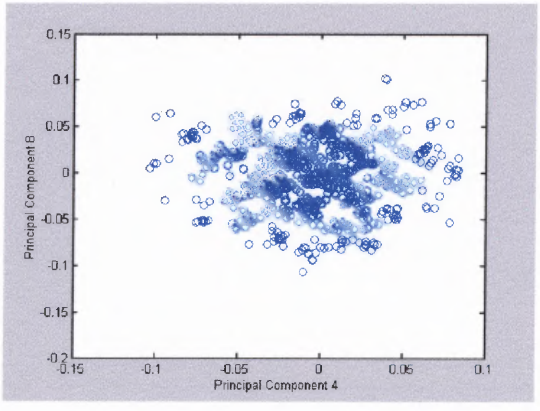
(c)



(d)

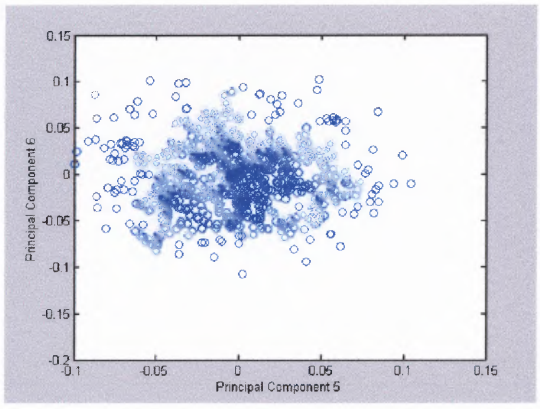


(e)

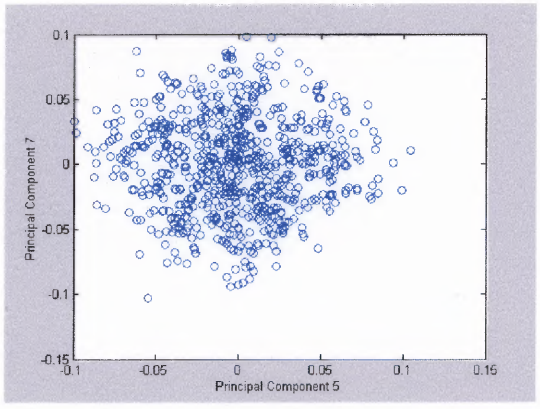


(f)

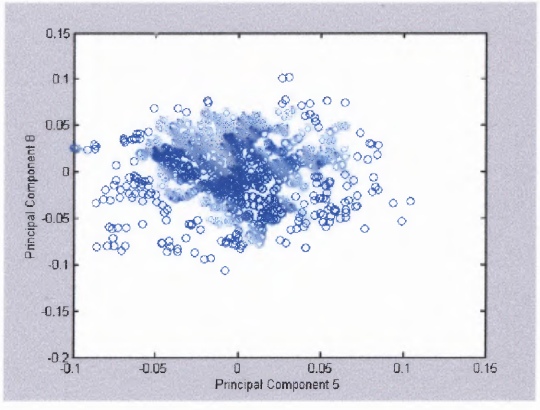
Figure C.4 Score plots of GEM-scaled DM324 data.



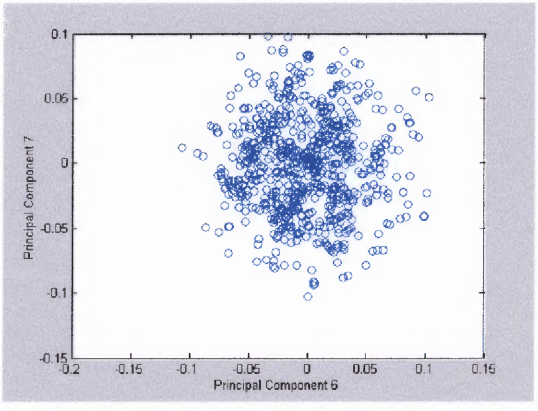
(a)



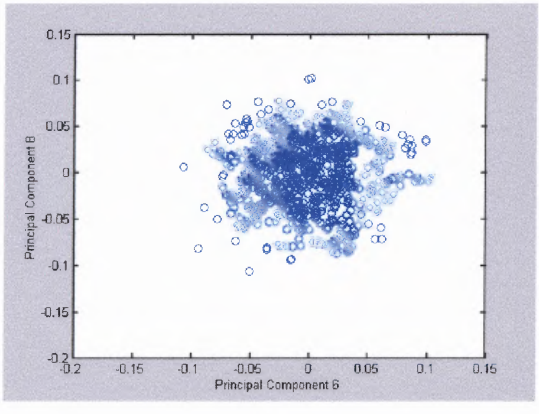
(b)



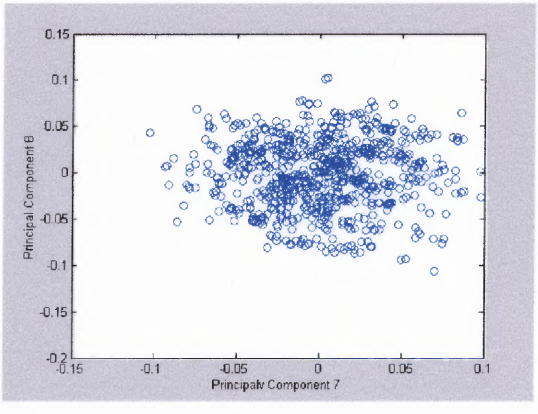
(c)



(d)



(e)



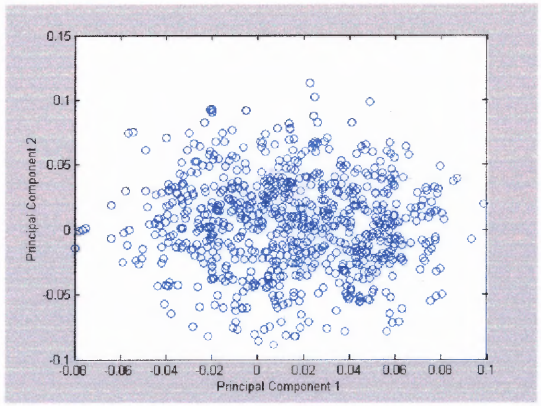
(f)

Figure C.5 Score plots of GEM-scaled DM324 data.

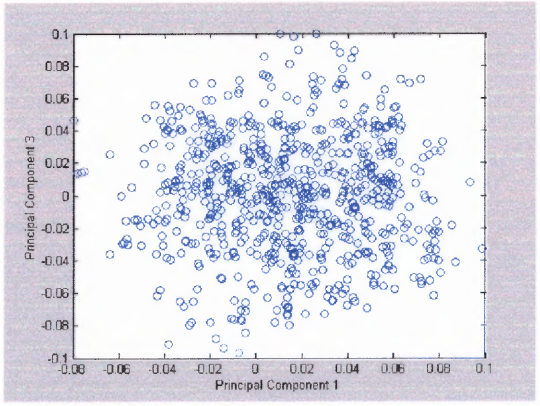
APPENDIX D

GEM-SCALED DATA FOR TP250

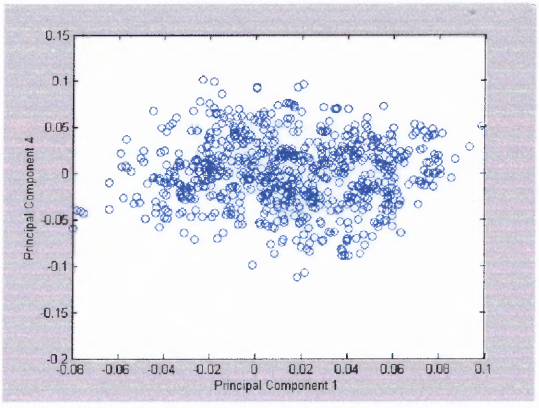
The following figures are the result of GEM scaling the data for analog TP250.
The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/gem/TP250/runsvdscoresTP250March.m
The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/gem/TP250/March11circularTp250.mat



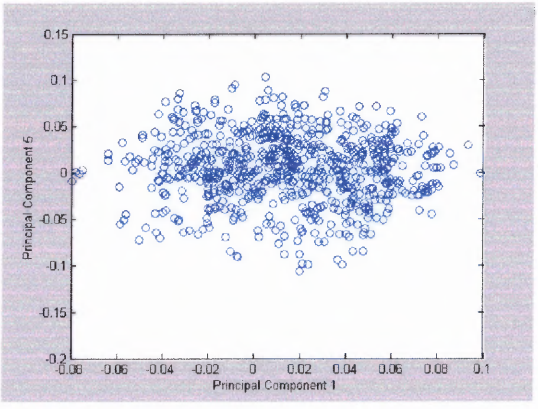
(a)



(b)

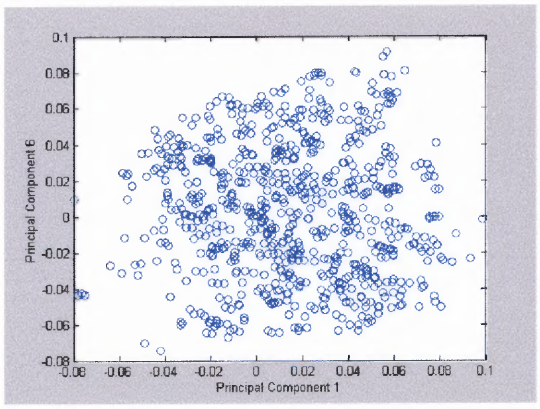


(c)

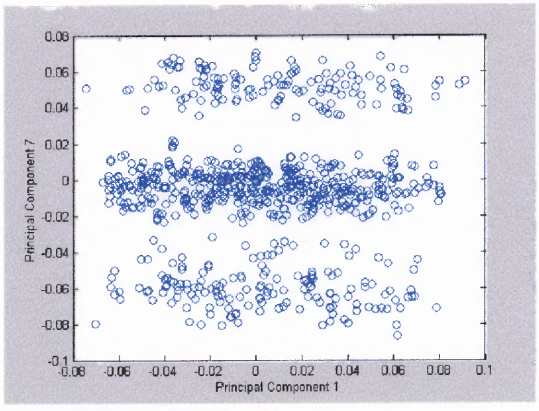


(d)

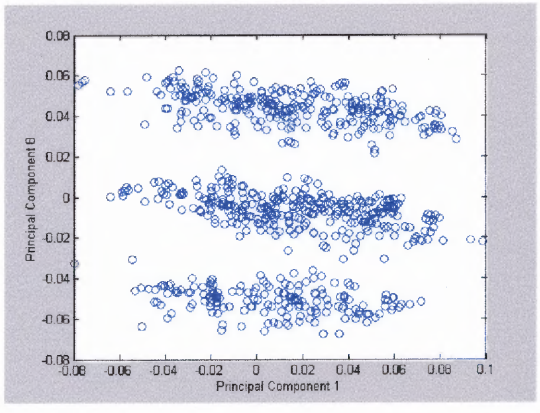
Figure D.1 Score plots of GEM-scaled TP250 data.



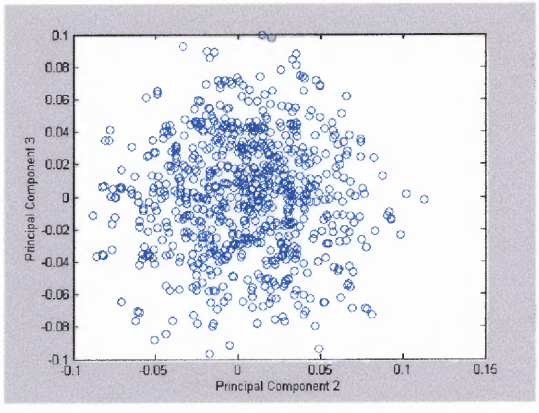
(a)



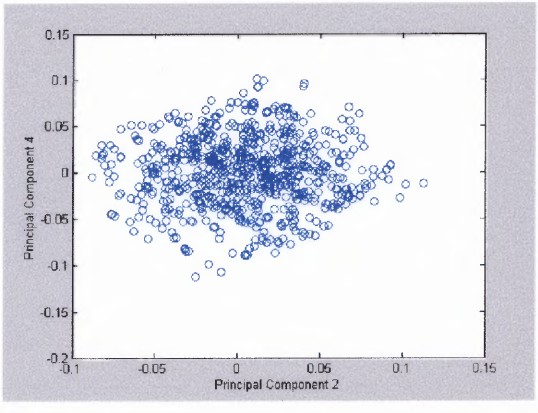
(b)



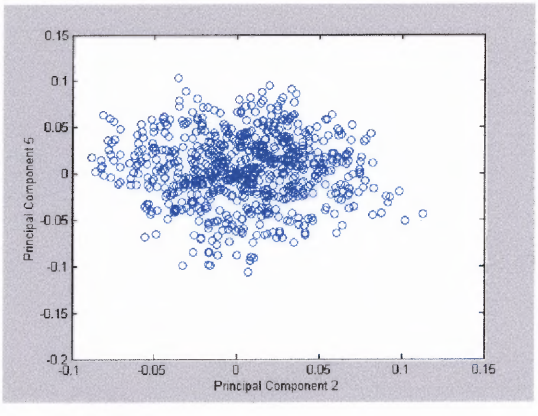
(c)



(d)

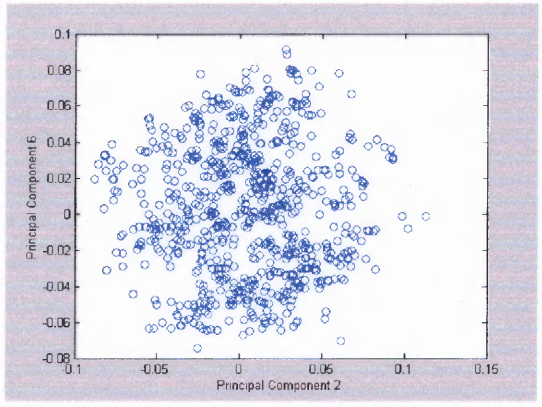


(e)

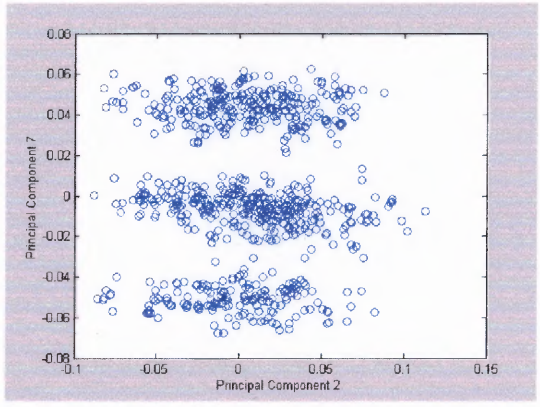


(f)

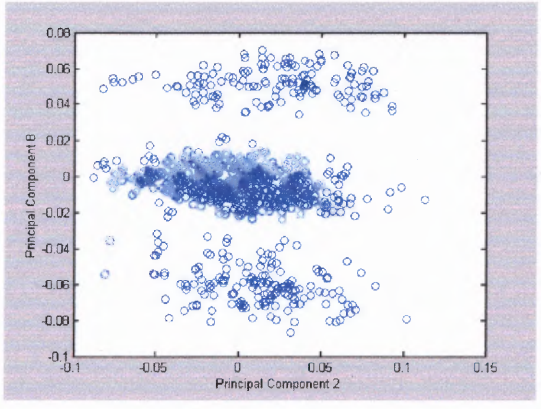
Figure D.2 Score plots of GEM-scaled TP250 data.



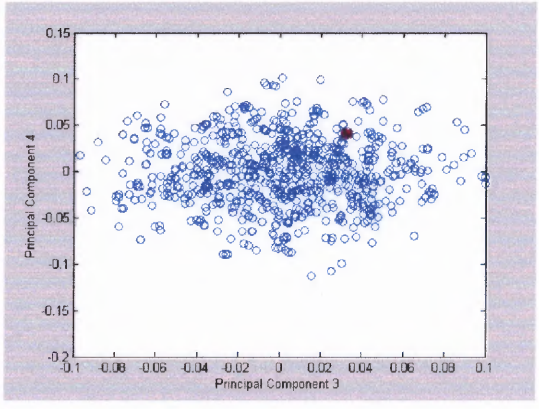
(a)



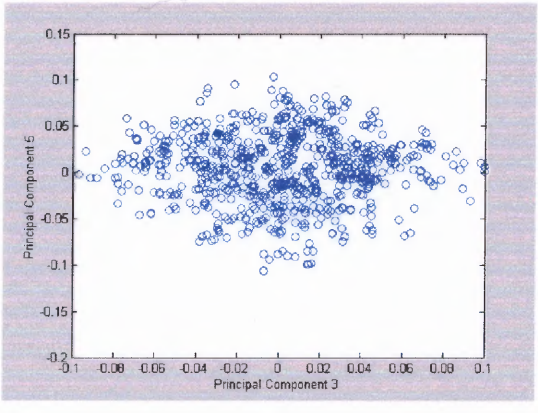
(b)



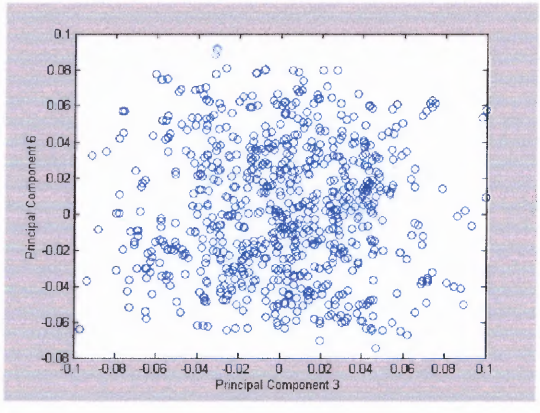
(c)



(d)

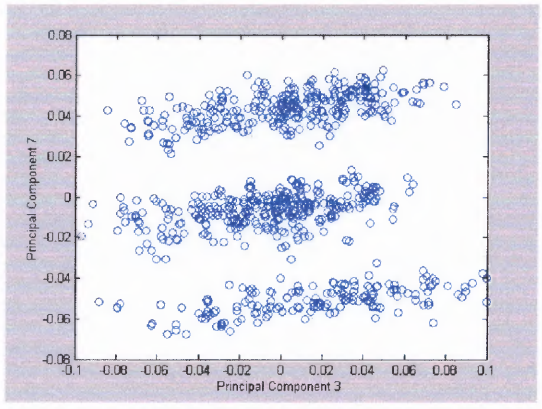


(e)

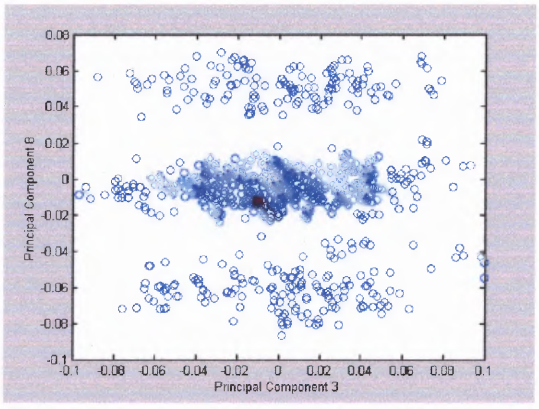


(f)

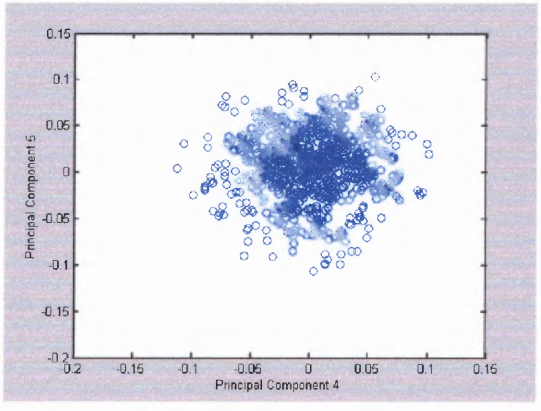
Figure D.3 Score plots of GEM-scaled TP250 data.



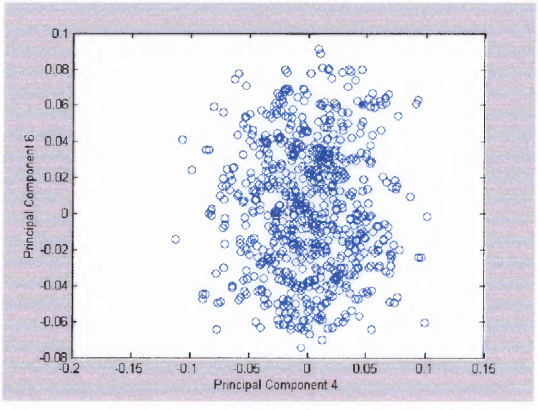
(a)



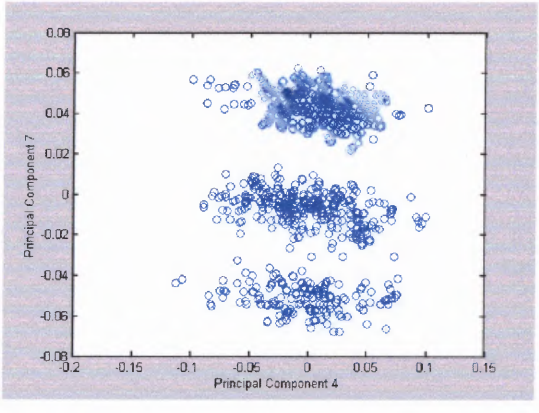
(b)



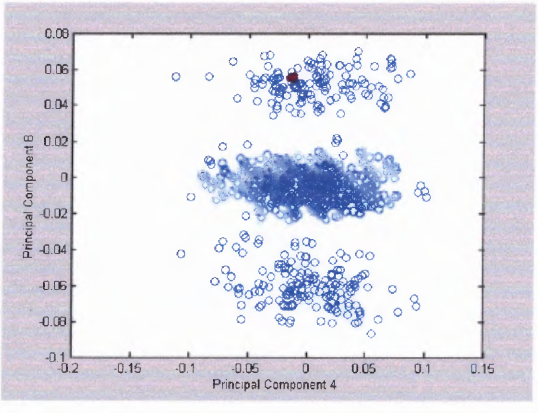
(c)



(d)

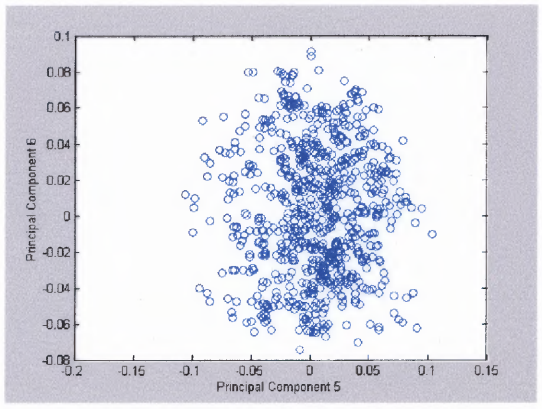


(e)

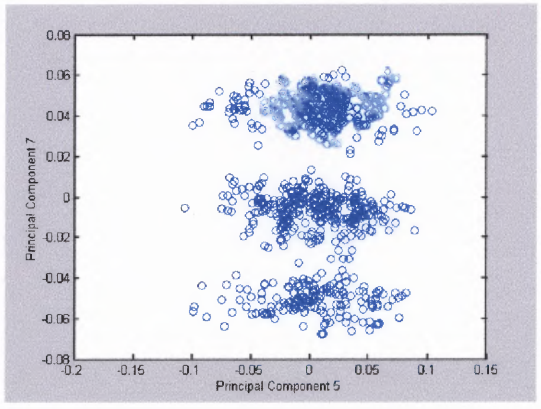


(f)

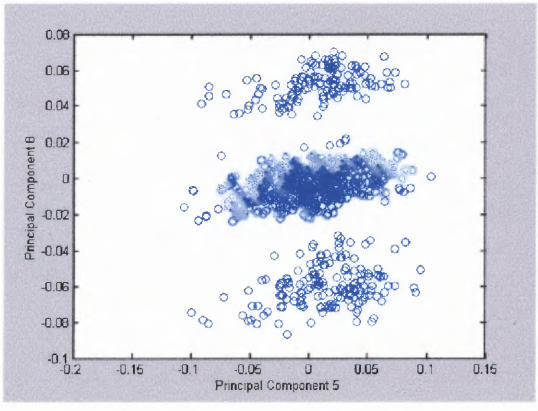
Figure D.4 Score plots of GEM-scaled TP250 data.



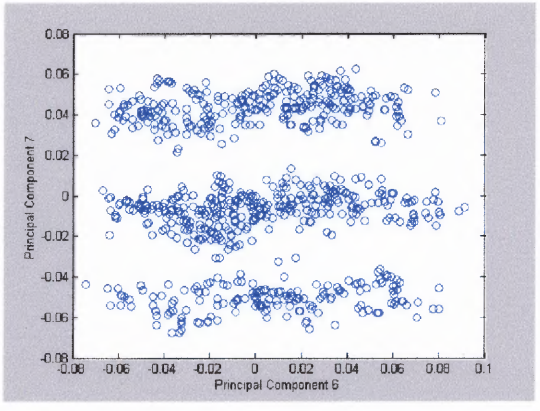
(a)



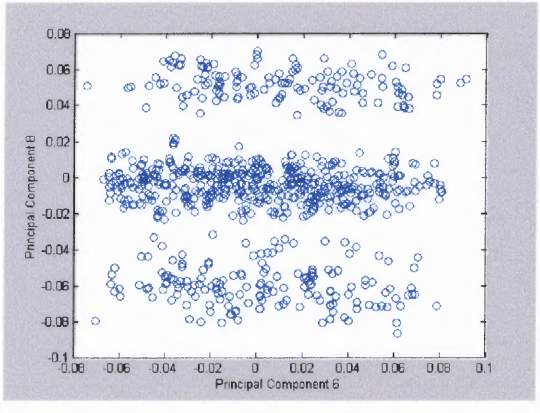
(b)



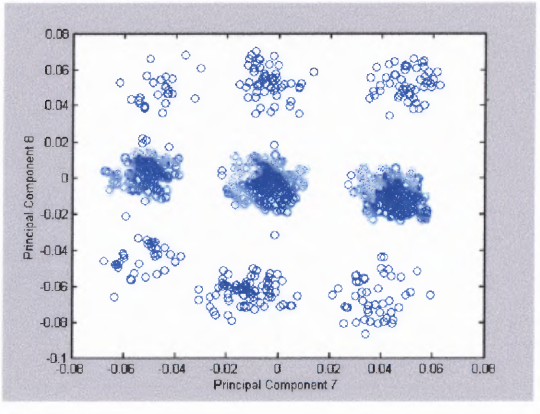
(c)



(d)



(e)



(f)

Figure D.5 Score plots of GEM-scaled TP250 data.

APPENDIX E

COMBINED DATA FOR DM324 AND TP250 SCALED TO DM324 GEM

The following figures are the result of GEM scaling the data for both analogs using the GEM of DM324 to scale the data.

The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/combined/usingDM324/runsvdscoresbothcircular.m

The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/combined/usingDM324/DMTPcircularscaled.mat

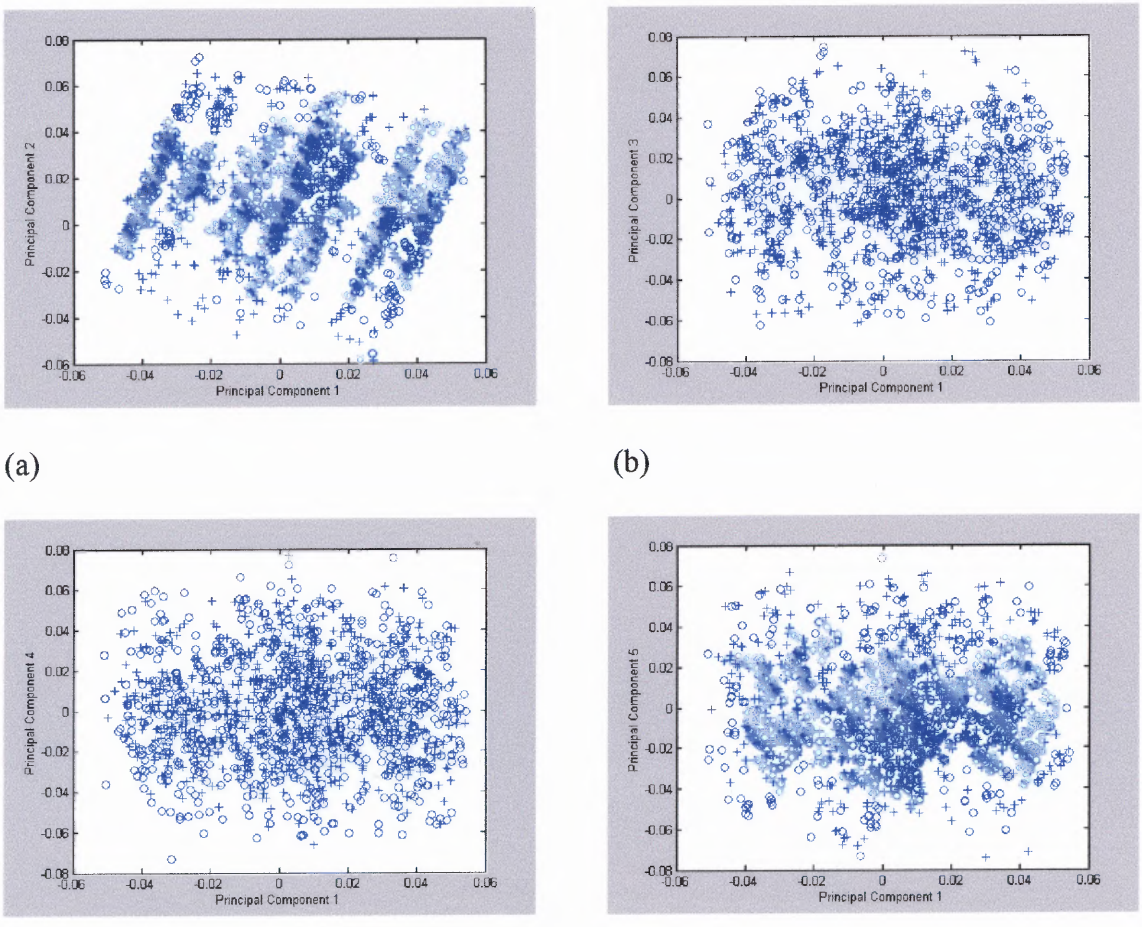
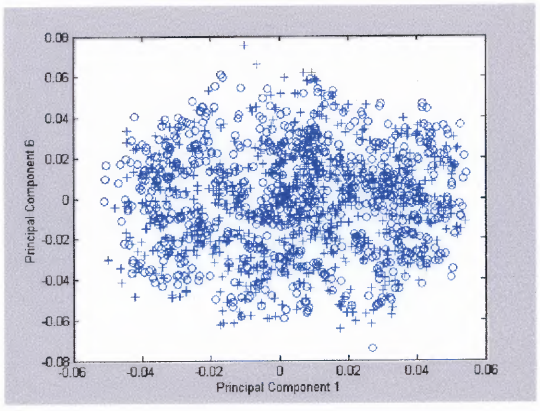
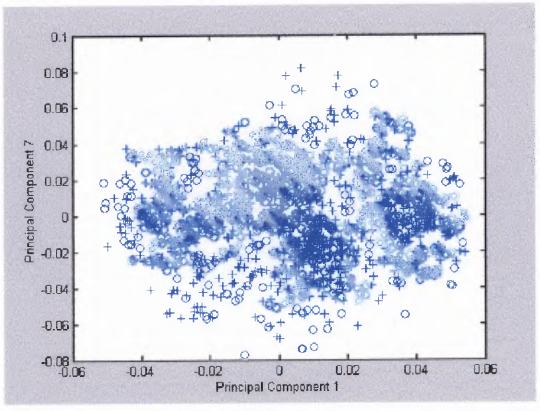


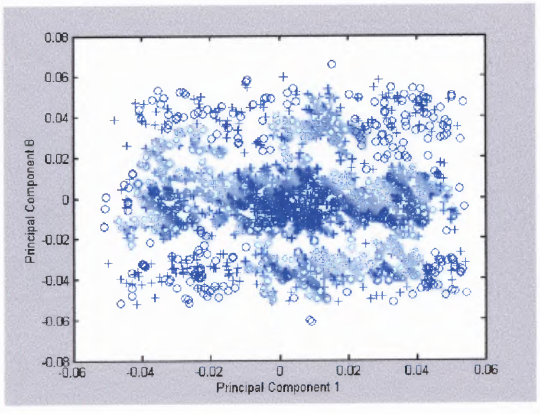
Figure E.1 Score plots of DM324 and TP250 GEM-scaled to DM324 GEM.



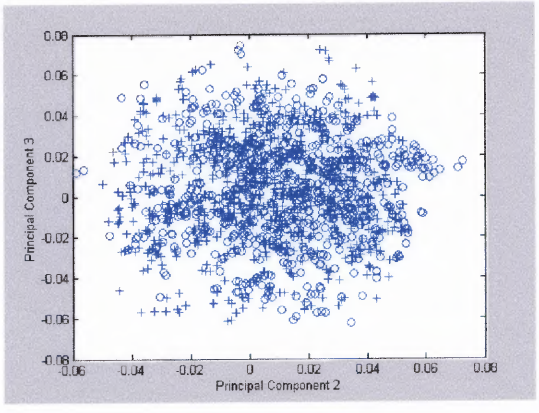
(a)



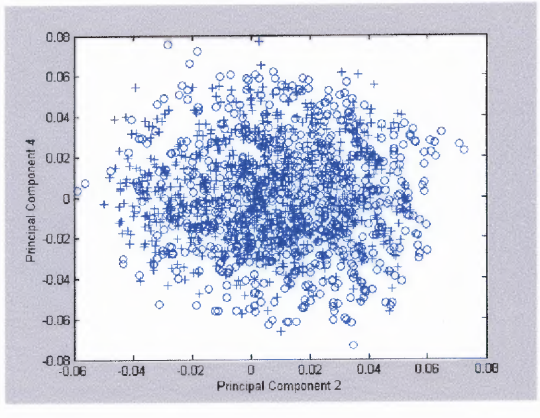
(b)



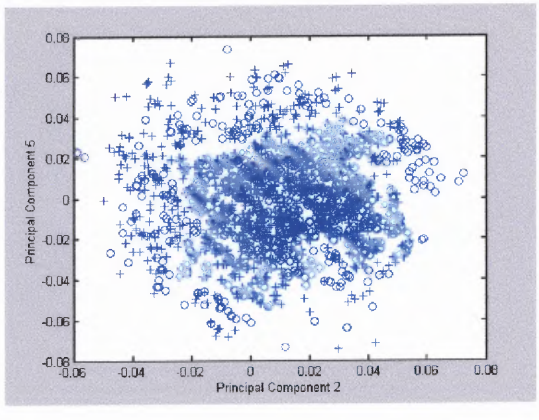
(c)



(d)

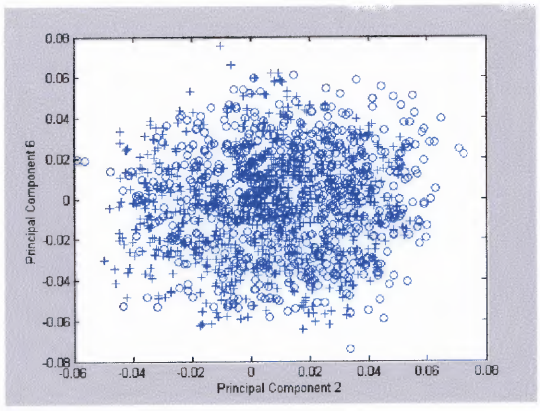


(e)

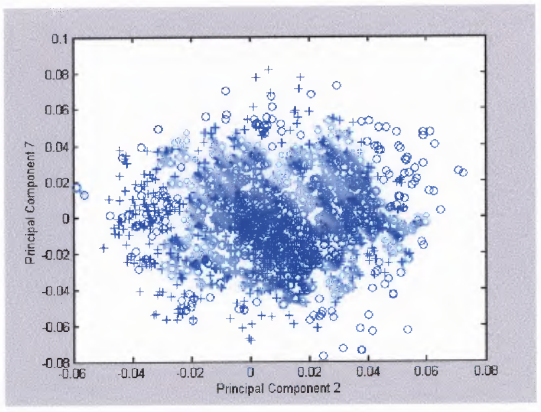


(f)

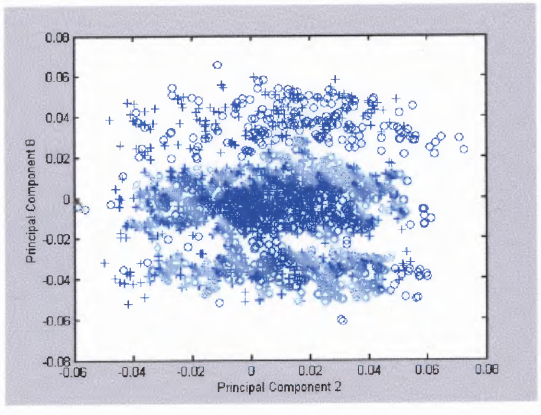
Figure E.2 Score plots of DM324 and TP250 GEM-scaled to DM324 GEM.



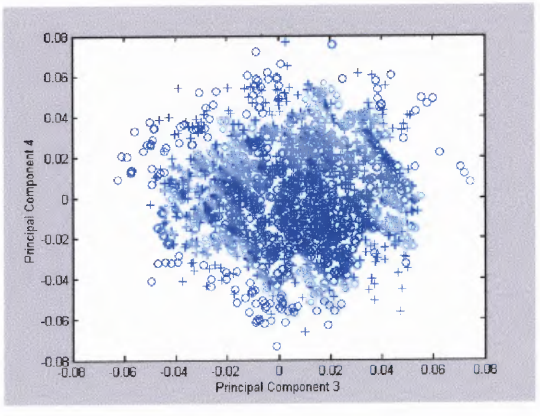
(a)



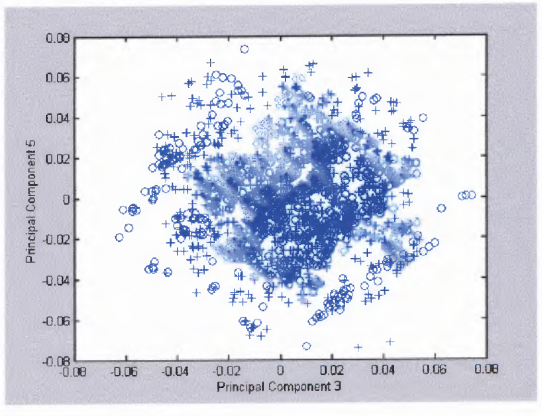
(b)



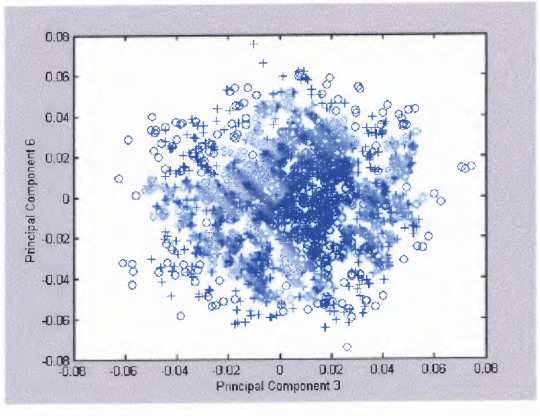
(c)



(d)

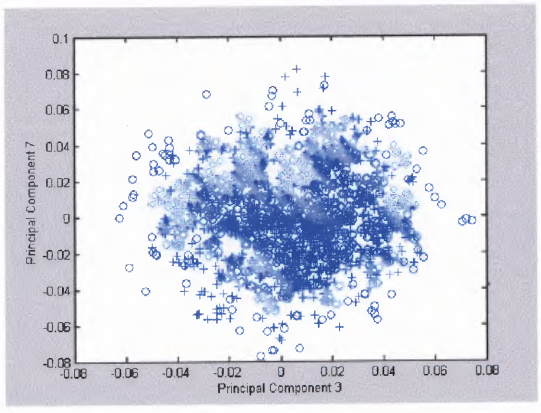


(e)

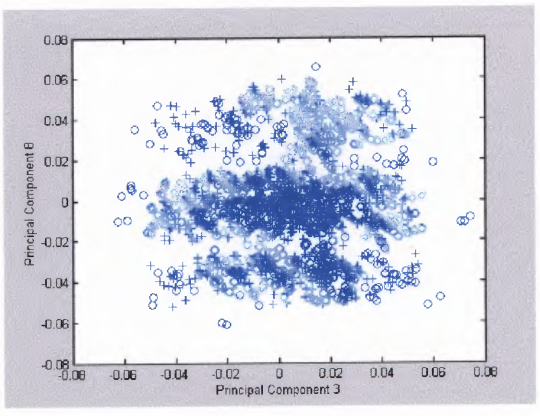


(f)

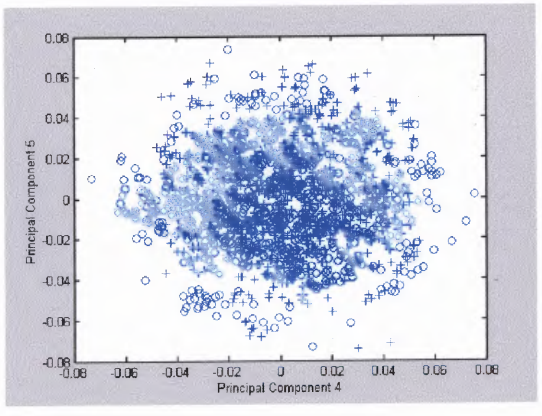
Figure E.3 Score plots of DM324 and TP250 GEM-scaled to DM324 GEM.



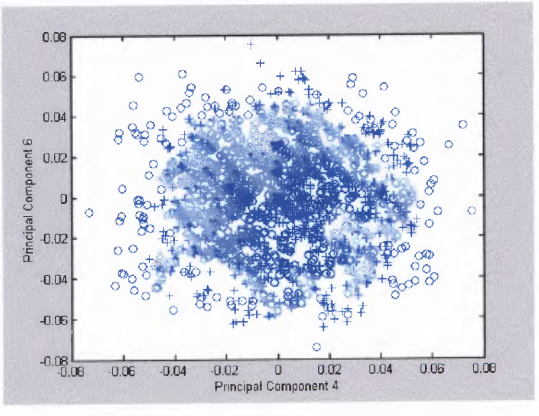
(a)



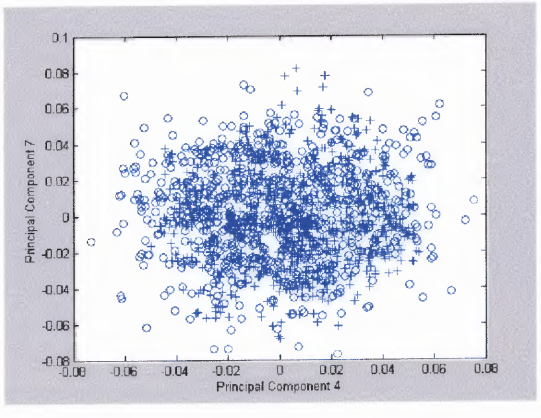
(b)



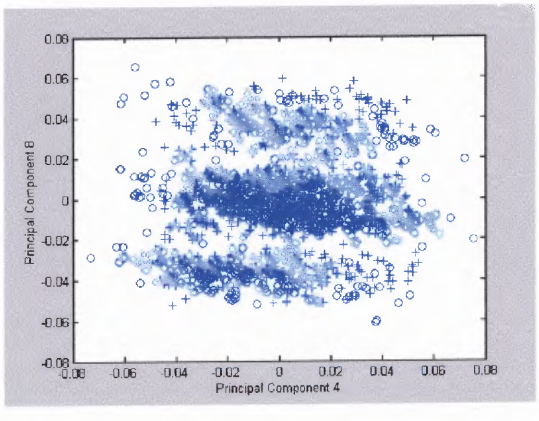
(c)



(d)

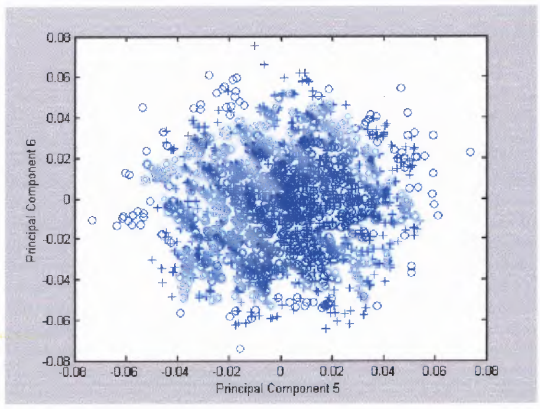


(e)

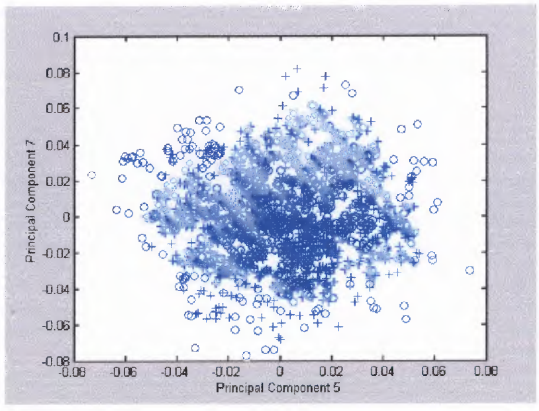


(f)

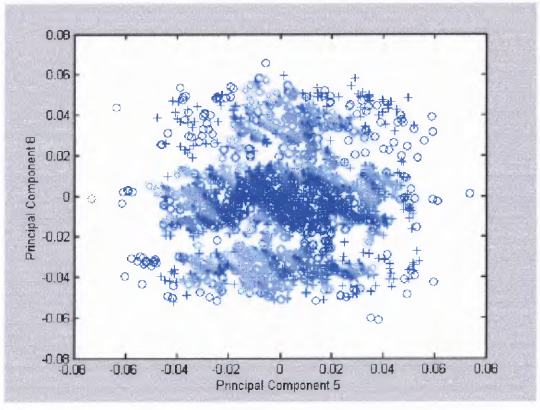
Figure E.4 Score plots of DM324 and TP250 GEM-scaled to DM324 GEM.



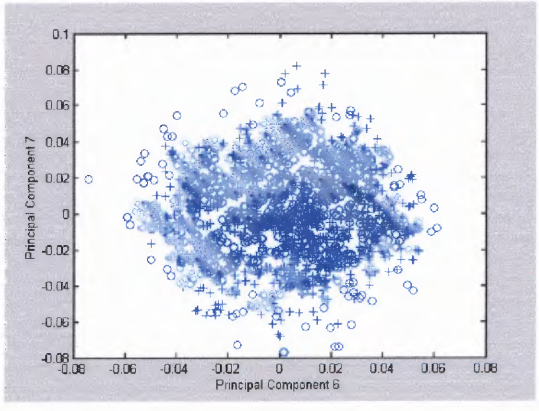
(a)



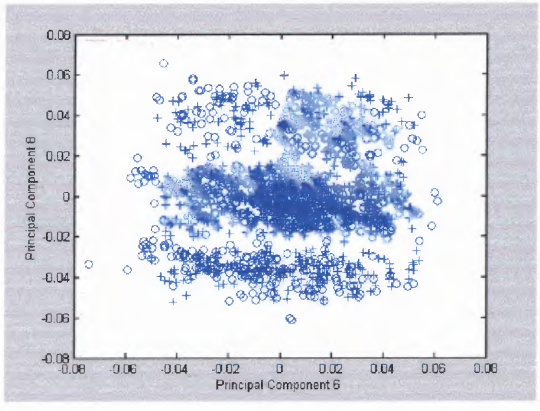
(b)



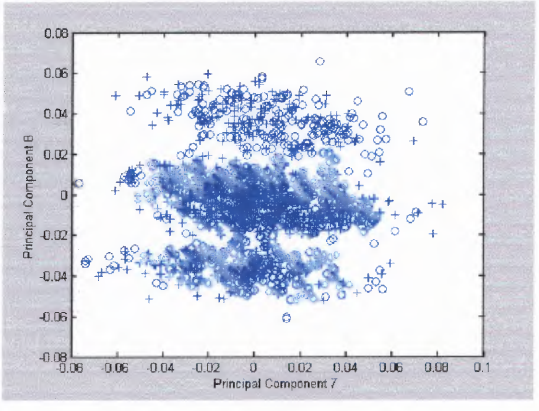
(c)



(d)



(e)



(f)

Figure E.5 Score plots of DM324 and TP250 GEM-scaled to DM324 GEM.

APPENDIX F

COMBINED DATA FOR DM324 AND TP250 SCALED TO TP250 GEM

The following figures are the result of GEM scaling the data for both analogs using the GEM of TP250 to scale the data.

The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/combined/usingTP250/runsvdscoresbothcircularTPgem.m

The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/combined/usingTP250/DMTPcircularscaledTPgem.mat

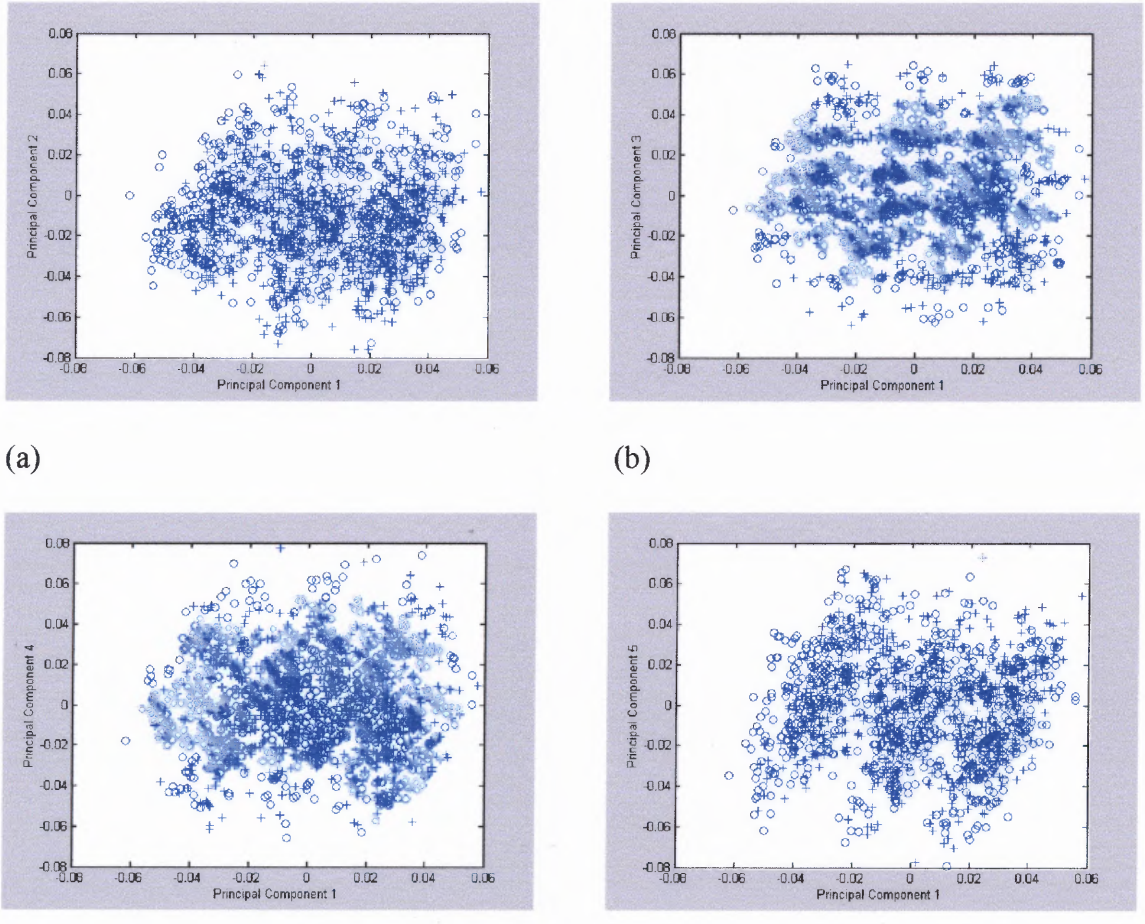
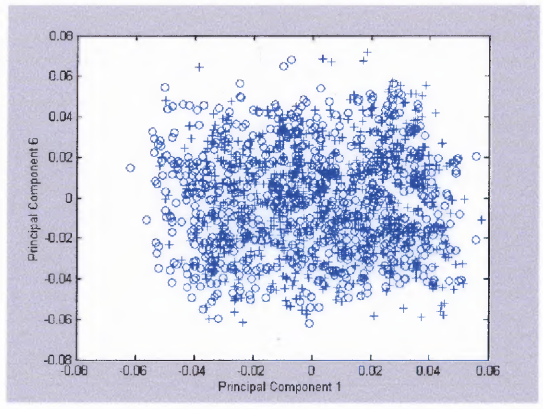
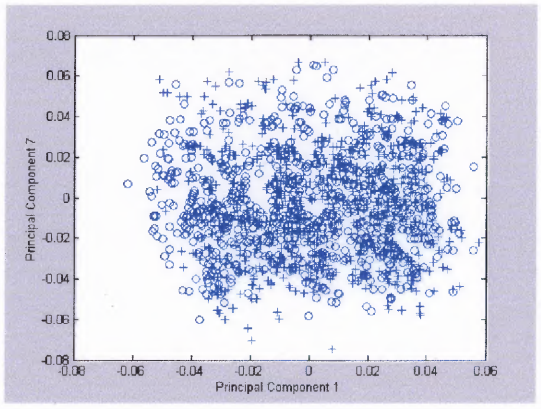


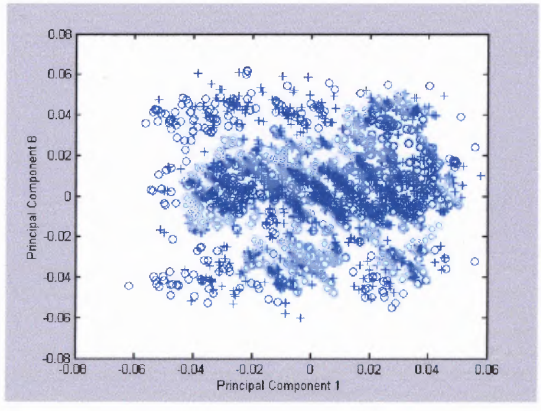
Figure F.1 Score plots for DM324 and TP250 GEM-scaled to TP250 GEM.



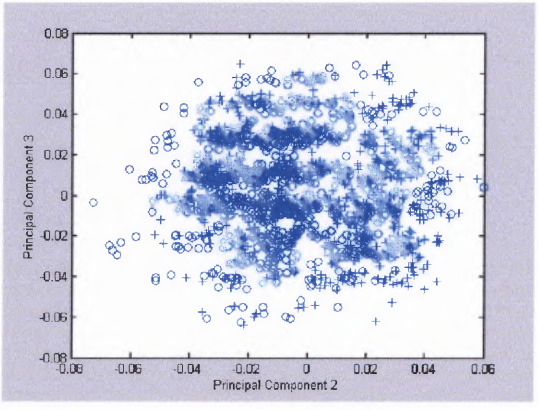
(a)



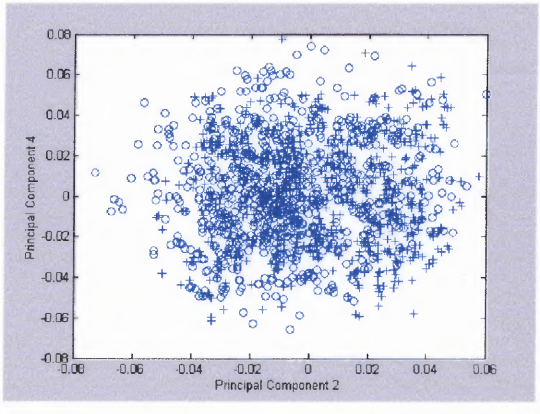
(b)



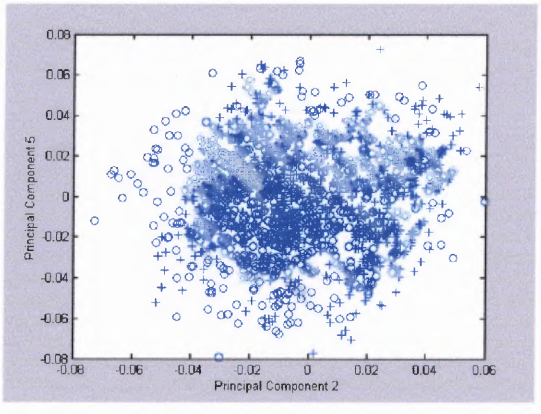
(c)



(d)

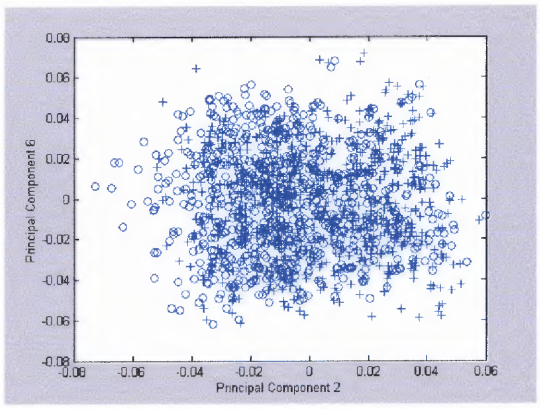


(e)

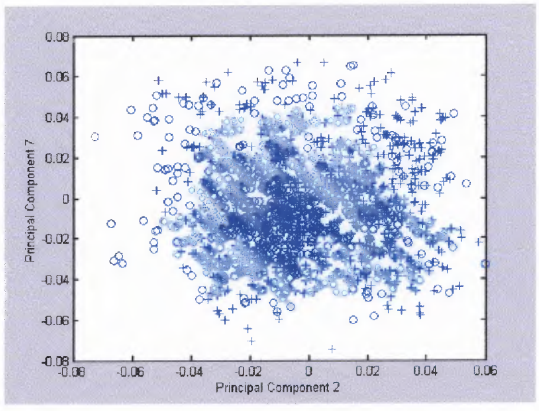


(f)

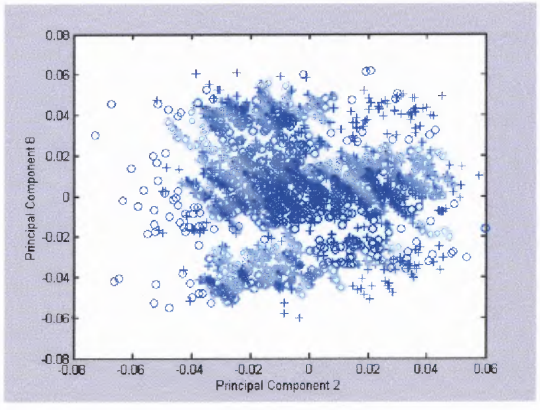
Figure F.2 Score plots for DM324 and TP250 GEM-scaled to TP250 GEM.



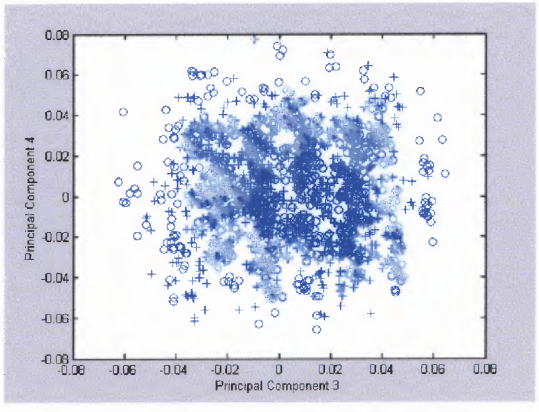
(a)



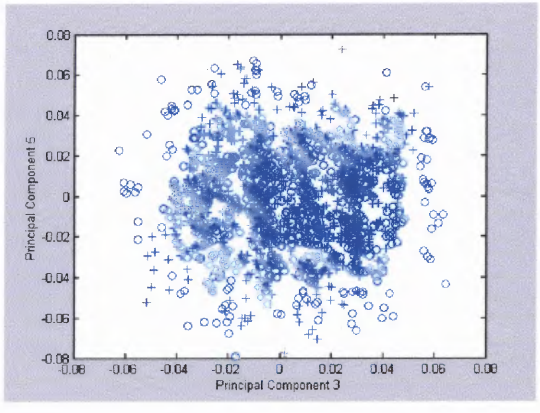
(b)



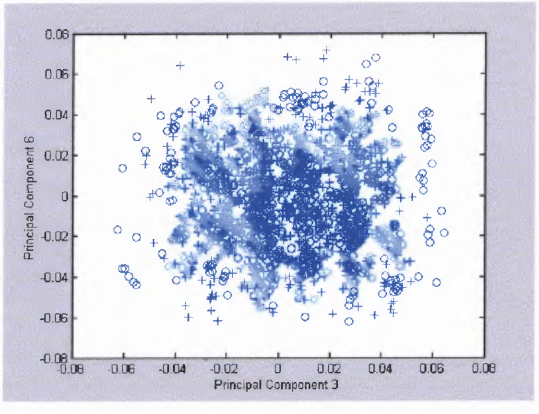
(c)



(d)

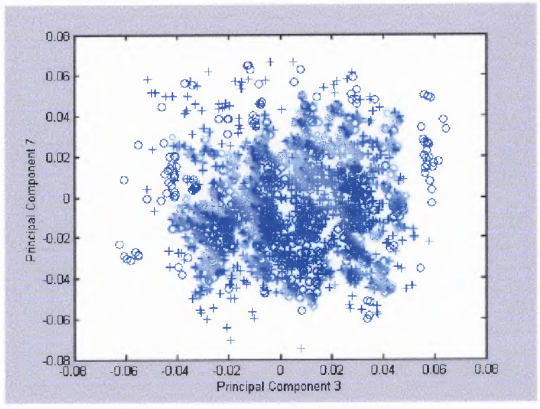


(e)

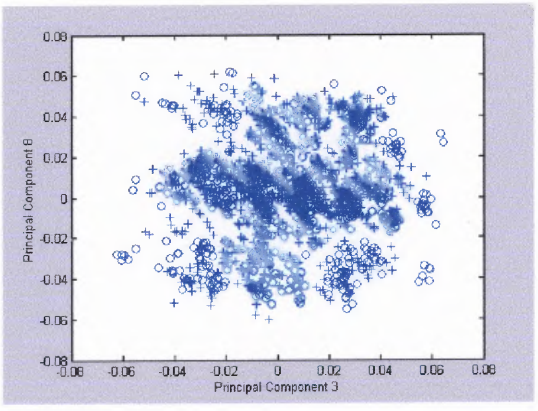


(f)

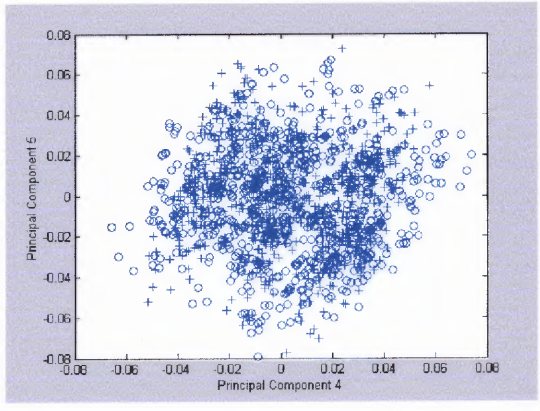
Figure F.3 Score plots for DM324 and TP250 GEM-scaled to TP250 GEM.



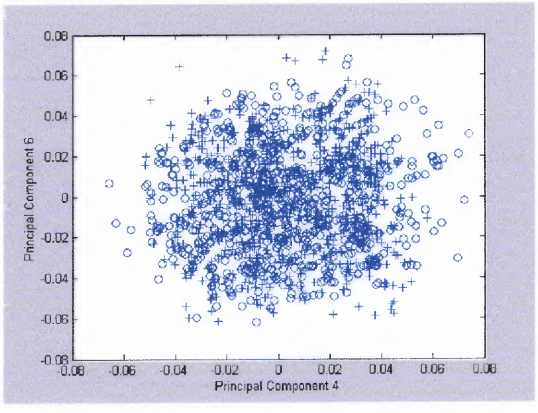
(a)



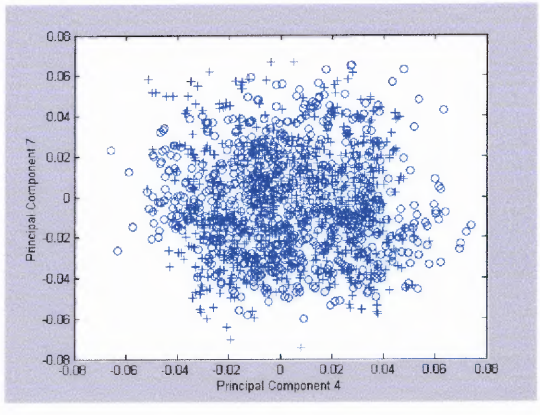
(b)



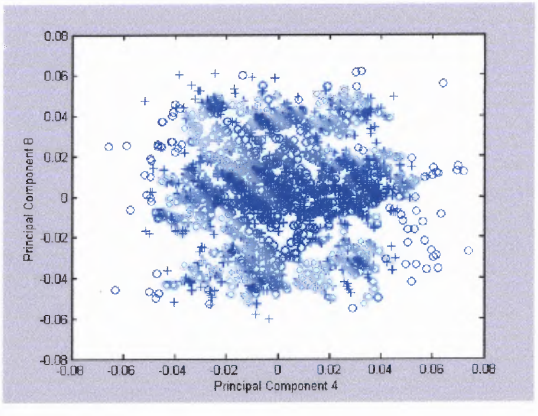
(c)



(d)

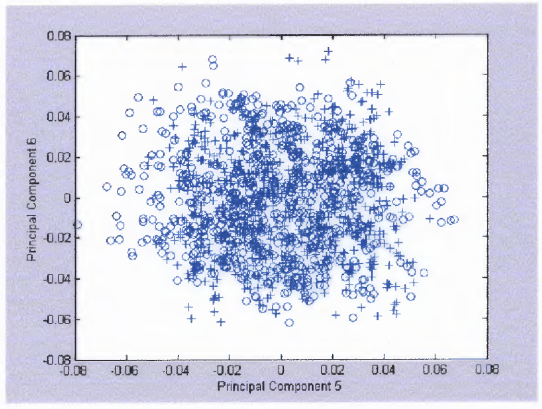


(e)

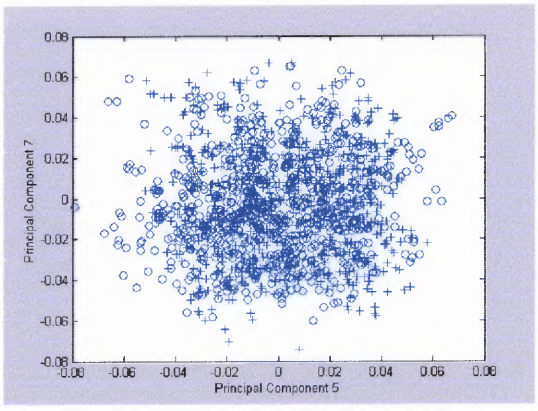


(f)

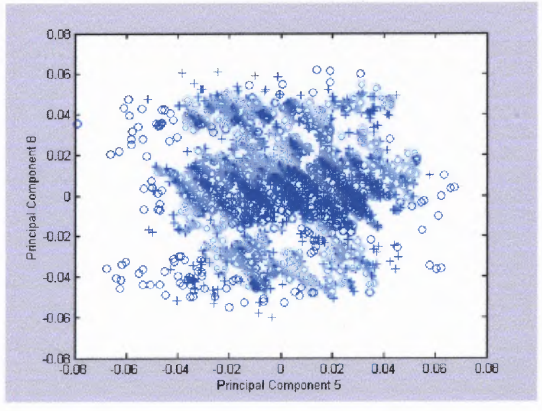
Figure F.4 Score plots for DM324 and TP250 GEM-scaled to TP250 GEM.



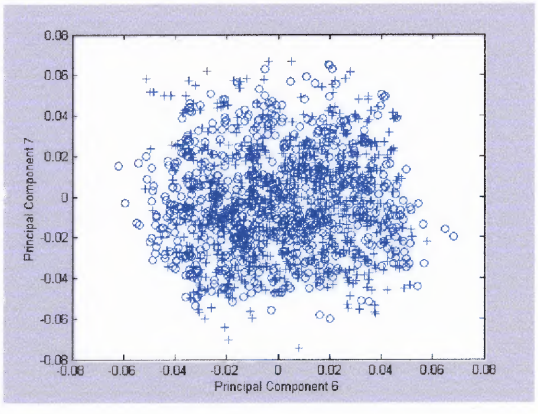
(a)



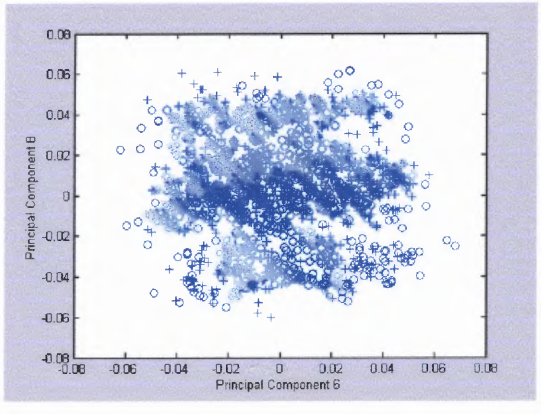
(b)



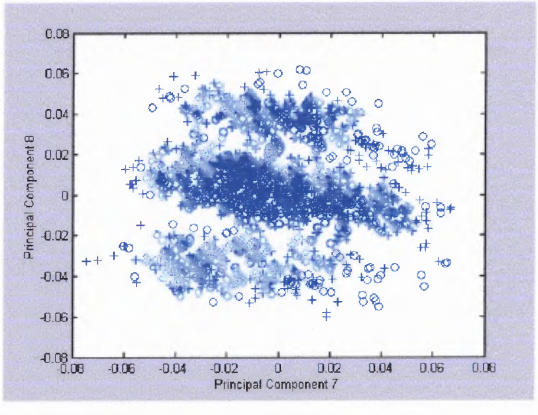
(c)



(d)



(e)



(f)

Figure F.5 Score plots for DM324 and TP250 GEM-scaled to TP250 GEM.

APPENDIX G

DATA FOR DNA

The following figures are the result of GEM scaling the data for DNA provided by Dr. Ron Wehrens of the University of Nijmegen. The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/dna/gem/theirdata.m
The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/dna/gem/theirdata.mat

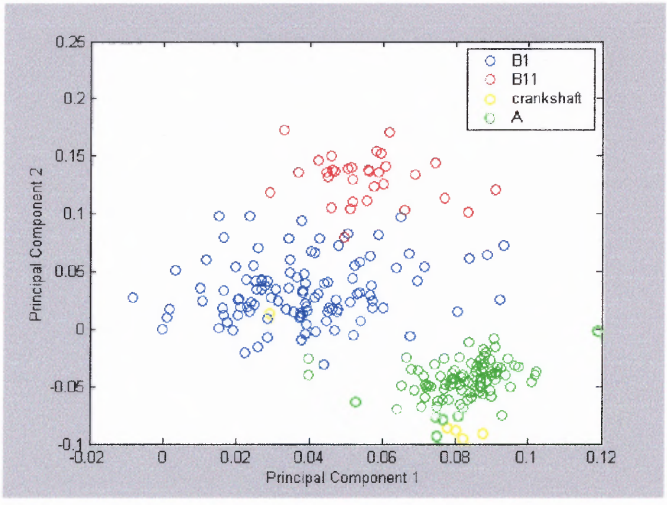


Figure G.1 GEM-scaled DNA data of PC1 vs PC2.

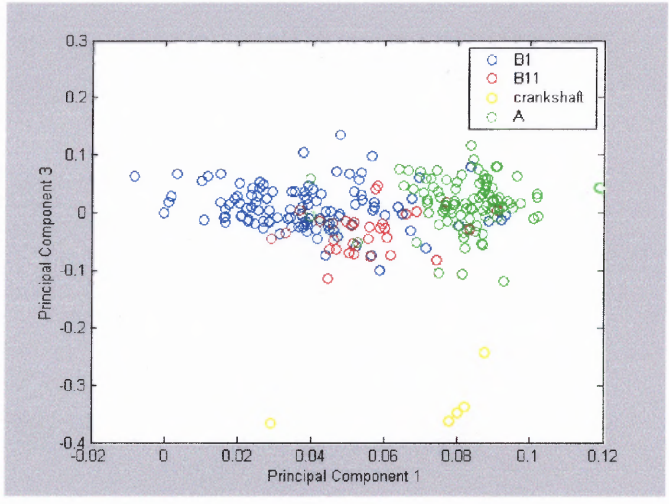


Figure G.2 GEM-scaled DNA data of PC1 vs PC3.

The following figures are the result of median scaling the data for DNA provided by Dr. Ron Wehrens of the University of Nijmegen. The MATLAB program which creates the graphs is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/dna/med/runsvd.m
The MATALB data which is needed to run the program is contained in the following directory:
/afs/cad/research/chem/venanzi/6/SVD/MATLAB programs/dna/med/test.mat

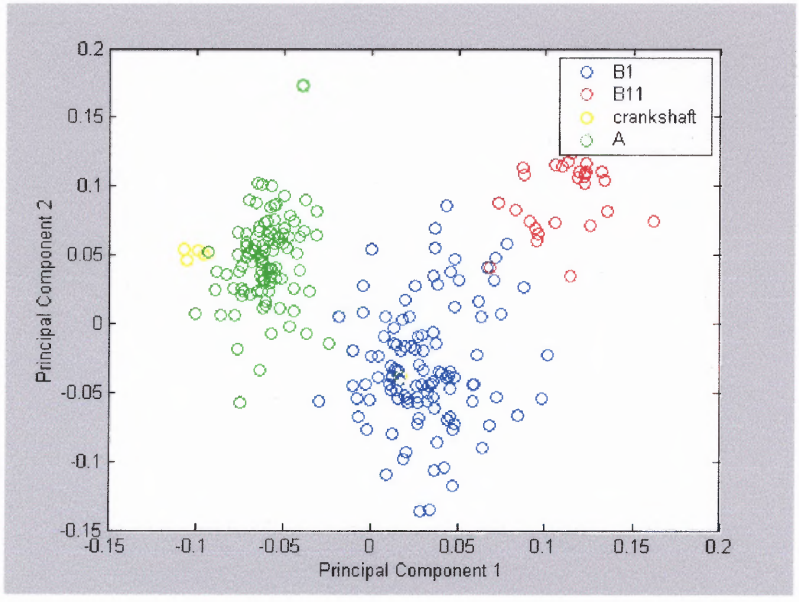


Figure G.3 Median-scaled score plot of DNA on PC1 vs PC2.

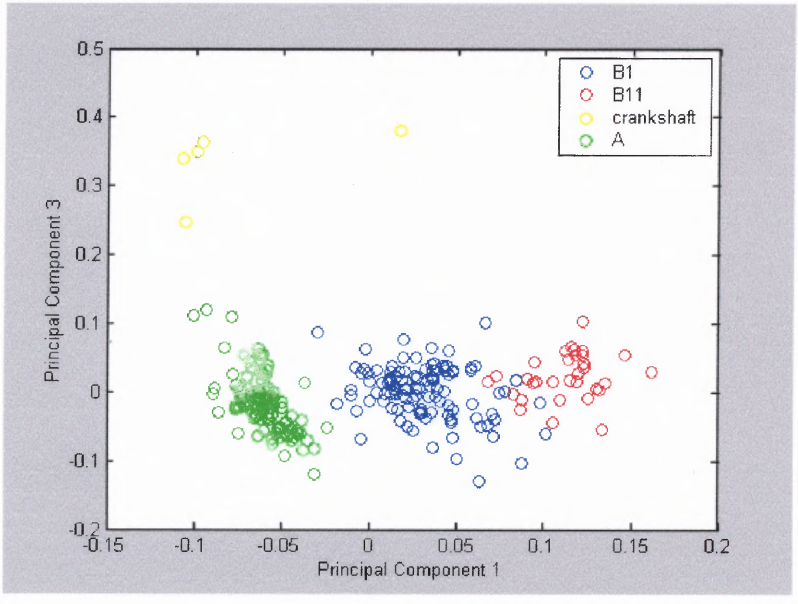


Figure G.4 Median-scaled data plot of DNA on PC1 vs PC3.

APPENDIX H

CLUSTER LISTS FOR DM324

The following lists the clusters that have been designated to the three major groups of DM324 conformers. The Fuzzy Clustering (FC) groups were provided by Milind Misra and Amit Banerjee. The XCluster (XC) groups were provided by Kathleen Gilbert.

Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC
001	2	2	1	046	3	1	2	091	2	2	1
002	2	2	1	047	3	1	2	092	2	2	1
003	2	2	1	048	1	2	1	093	2	2	1
004	2	2	1	049	1	2	1	094	2	3	3
005	2	2	1	050	1	2	1	095	3	3	3
006	2	3	3	051	2	1	2	096	3	3	3
007	1	2	1	052	2	3	3	097	3	3	3
008	2	3	3	053	2	3	3	098	3	1	2
009	1	2	1	054	2	3	3	099	2	1	2
010	1	2	1	055	2	3	3	100	2	2	1
011	1	2	1	056	2	3	3	101	2	3	3
012	1	2	1	057	3	3	3	102	2	1	2
013	1	2	1	058	3	3	3	103	2	1	2
014	2	2	1	059	1	3	3	104	3	1	2
015	3	1	2	060	1	2	1	105	3	1	2
016	3	1	2	061	3	1	2	106	3	2	1
017	3	1	2	062	3	1	2	107	1	2	1
018	3	3	3	063	1	2	1	108	1	2	1
019	2	2	1	064	1	2	1	109	1	2	1
020	2	2	1	065	1	2	1	110	1	3	3
021	2	1	2	066	1	2	1	111	1	3	3
022	3	1	2	067	2	1	2	112	1	3	3
023	2	2	1	068	3	1	2	113	1	3	3
024	2	2	1	069	3	1	2	114	3	3	3
025	2	2	1	070	3	1	2	115	2	3	3
026	2	2	1	071	3	1	2	116	3	3	3
027	2	2	1	072	3	1	2	117	1	2	1
028	3	1	2	073	3	1	2	118	1	2	1
029	3	1	2	074	2	2	1	119	1	2	1
030	3	1	2	075	3	1	2	120	1	2	1
031	1	2	1	076	2	2	1	121	1	2	1
032	1	2	1	077	3	1	2	122	1	2	1
033	3	1	2	078	1	2	1	123	1	2	1
034	3	1	2	079	1	2	1	124	2	2	1
035	2	2	1	080	3	1	2	125	2	2	1
036	1	3	3	081	3	1	2	126	2	2	1
037	1	3	3	082	3	1	2	127	2	2	1
038	1	3	3	083	3	1	2	128	1	3	3
039	3	3	3	084	3	3	3	129	1	3	3
040	3	3	3	085	3	3	3	130	2	1	2
041	3	3	3	086	1	2	1	131	2	1	2
042	3	3	3	087	1	2	1	132	3	1	2
043	3	1	2	088	1	2	1	133	1	2	1
044	3	1	2	089	1	2	1	134	2	2	1
045	3	1	2	090	1	2	1	135	2	3	3

Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC
136	2	3	3	186	2	2	1	236	2	1	2
137	2	3	3	187	2	1	2	237	2	1	2
138	1	3	3	188	2	1	2	238	2	1	2
139	3	3	3	189	3	3	3	239	2	1	2
140	1	3	3	190	3	3	3	240	2	1	2
141	1	3	3	191	1	3	3	241	2	1	2
142	1	3	3	192	3	1	2	242	2	3	3
143	2	3	3	193	3	1	2	243	2	3	3
144	2	3	3	194	2	1	2	244	2	2	1
145	2	3	3	195	2	1	2	245	2	2	1
146	2	3	3	196	3	1	2	246	2	2	1
147	2	2	Minor D	197	3	1	2	247	1	2	1
148	1	2	1	198	3	1	2	248	1	2	1
149	1	2	1	199	2	2	1	249	2	2	1
150	1	2	1	200	2	2	1	250	1	2	1
151	1	2	1	201	2	2	1	251	2	2	1
152	1	2	1	202	2	2	1	252	2	2	1
153	3	3	3	203	3	1	2	253	1	2	1
154	2	1	2	204	1	2	1	254	2	2	1
155	2	3	3	205	2	1	2	255	2	2	1
156	1	2	1	206	1	3	3	256	2	2	1
157	2	2	1	207	1	2	1	257	2	2	1
158	2	2	1	208	1	2	1	258	1	2	1
159	2	3	3	209	1	3	3	259	1	2	1
160	2	3	3	210	3	3	3	260	1	2	1
161	2	2	1	211	3	3	3	261	1	2	1
162	1	2	1	212	2	2	1	262	2	1	2
163	2	2	1	213	2	2	1	263	2	1	2
164	2	2	1	214	1	2	1	264	2	2	1
165	2	2	1	215	1	2	1	265	2	3	3
166	2	2	1	216	3	3	3	266	2	3	3
167	2	2	1	217	1	2	1	267	2	3	3
168	2	2	1	218	1	2	1	268	2	3	3
169	2	2	1	219	1	2	1	269	2	3	3
170	2	2	1	220	1	2	1	270	2	3	3
171	2	2	1	221	2	3	3	271	2	3	3
172	1	2	1	222	2	3	3	272	1	2	1
173	1	2	1	223	2	3	3	273	1	2	1
174	1	2	1	224	2	3	3	274	2	2	1
175	1	2	1	225	2	3	3	275	2	2	1
176	1	3	3	226	2	3	3	276	2	2	1
177	3	3	3	227	2	3	3	277	2	3	Minor C
178	3	3	3	228	2	3	3	278	3	1	2
179	3	3	3	229	2	3	3	279	2	1	2
180	2	3	3	230	2	3	3	280	2	2	1
181	3	3	3	231	2	3	3	281	1	2	1
182	3	3	3	232	2	3	3	282	2	2	1
183	2	2	1	233	2	2	1	283	2	2	1
184	2	2	1	234	2	1	2	284	2	2	1
185	2	2	1	235	2	1	2	285	2	2	1

Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC
286	2	2	1	336	3	3	3	386	3	3	3
287	2	2	2	337	1	2	1	387	2	3	3
288	3	1	2	338	1	2	1	388	3	3	3
289	1	2	1	339	1	2	1	389	3	3	3
290	1	2	1	340	1	2	1	390	3	3	3
291	2	2	1	341	1	2	1	391	3	3	3
292	1	2	1	342	1	2	1	392	3	3	3
293	3	3	3	343	1	2	1	393	3	3	3
294	2	3	Minor B	344	3	1	2	394	3	3	3
295	2	3	3	345	3	1	2	395	3	1	2
296	2	3	3	346	2	1	2	396	3	3	3
297	2	3	3	347	2	1	2	397	1	2	1
298	2	1	2	348	3	1	2	398	2	2	1
299	2	3	3	349	3	1	2	399	1	2	1
300	2	3	3	350	3	1	2	400	2	2	1
301	3	1	2	351	3	1	2	401	2	2	1
302	2	1	2	352	3	1	2	402	1	2	1
303	2	2	1	353	2	1	2	403	3	3	3
304	2	2	1	354	2	3	3	404	3	3	3
305	2	2	1	355	1	3	3	405	1	2	1
306	3	1	2	356	1	2	1	406	1	2	1
307	1	2	1	357	1	2	1	407	1	2	1
308	1	2	1	358	2	3	3	408	3	1	2
309	3	1	2	359	2	3	3	409	3	1	2
310	3	1	2	360	2	3	3	410	3	3	3
311	2	3	3	361	2	3	3	411	1	3	3
312	2	3	Minor B	362	2	3	3	412	3	1	2
313	2	2	1	363	2	3	3	413	3	1	2
314	3	1	2	364	2	2	1	414	3	1	2
315	3	1	2	365	2	2	1	415	3	3	3
316	3	3	3	366	2	1	2	416	3	3	3
317	3	3	3	367	2	1	2	417	1	2	1
318	1	2	1	368	2	1	2	418	3	1	2
319	1	2	1	369	3	1	2	419	3	1	2
320	1	2	1	370	3	1	2	420	3	1	2
321	2	1	2	371	3	1	2	421	3	1	2
322	2	1	2	372	2	1	2	422	3	1	2
323	2	1	2	373	2	1	2	423	3	3	3
324	2	1	2	374	2	1	2	424	1	2	1
325	3	1	2	375	3	1	2	425	2	2	1
326	3	1	2	376	1	3	3	426	2	2	1
327	3	1	2	377	3	3	3	427	2	2	1
328	1	2	2	378	3	3	3	428	2	2	1
329	1	2	2	379	3	3	3	429	2	2	1
330	1	2	2	380	1	3	3	430	2	2	1
331	3	1	2	381	2	3	3	431	2	1	2
332	3	1	2	382	2	3	3	432	2	1	2
333	1	3	3	383	2	3	3	433	2	1	2
334	2	3	3	384	2	2	1	434	1	2	1
335	1	3	3	385	3	3	3	435	2	3	3

Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC
436	1	3	3	486	2	2	1	536	3	1	2
437	2	3	3	487	2	2	1	537	3	1	2
438	2	1	2	488	1	2	1	538	3	1	2
439	2	3	3	489	1	2	1	539	2	1	2
440	2	2	1	490	1	2	1	540	2	2	1
441	2	3	3	491	1	2	1	541	2	1	2
442	2	3	3	492	1	3	3	542	2	2	1
443	2	1	2	493	2	3	3	543	3	1	2
444	2	1	2	494	2	3	3	544	2	1	2
445	2	2	1	495	2	3	3	545	3	1	2
446	2	2	1	496	2	3	3	546	3	1	2
447	2	1	2	497	2	3	3	547	3	3	Minor B
448	2	1	2	498	3	3	3	548	3	3	Minor B
449	3	1	2	499	1	3	3	549	3	3	3
450	3	1	2	500	2	3	Minor B	550	3	3	3
451	2	2	1	501	2	3	3	551	3	3	3
452	1	2	1	502	2	3	3	552	2	3	3
453	1	2	1	503	2	1	2	553	2	3	3
454	1	2	1	504	2	1	2	554	1	2	1
455	1	2	1	505	1	2	1	555	2	3	3
456	1	3	3	506	1	2	1	556	2	3	3
457	1	3	3	507	2	2	2	557	1	3	3
458	1	2	1	508	3	1	2	558	3	3	3
459	1	2	1	509	3	1	2	559	3	1	2
460	3	1	2	510	3	1	2	560	3	1	2
461	2	1	2	511	3	1	2	561	3	1	2
462	2	1	2	512	3	1	2	562	3	1	2
463	2	1	2	513	3	1	2	563	3	1	2
464	2	1	2	514	3	1	2	564	3	1	2
465	2	1	2	515	3	1	2	565	3	1	2
466	2	1	2	516	1	2	1	566	3	1	2
467	2	1	2	517	1	2	1	567	3	1	2
468	2	1	2	518	1	2	1	568	3	1	2
469	2	3	3	519	1	2	1	569	3	1	2
470	1	2	1	520	1	2	1	570	3	1	2
471	1	2	1	521	3	1	2	571	3	1	2
472	1	2	1	522	1	2	1	572	3	3	3
473	1	2	1	523	1	2	1	573	1	3	3
474	3	3	3	524	2	2	1	574	3	3	3
475	3	2	1	525	2	2	1	575	1	3	3
476	2	2	1	526	2	2	1	576	2	3	3
477	2	2	1	527	2	1	Minor A	577	2	3	3
478	2	2	1	528	2	1	2	578	2	3	3
479	3	3	3	529	3	1	2	579	2	3	3
480	1	3	3	530	3	1	2	580	2	3	3
481	3	3	3	531	3	1	2	581	2	3	3
482	3	3	3	532	3	1	2	582	1	3	3
483	3	3	3	533	3	1	2	583	3	1	2
484	1	2	1	534	3	1	2	584	2	2	2
485	2	2	1	535	3	1	2	585	2	2	2

Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC	Conf.#	SVD	FC	XC
586	1	2	1	636	2	1	2	686	2	2	1
587	1	2	1	637	1	3	3	687	2	3	Minor A
588	2	2	1	638	3	3	3	688	2	1	2
589	3	3	3	639	3	3	3	689	2	1	2
590	1	2	1	640	3	3	3	690	2	1	2
591	1	2	1	641	1	3	3	691	2	1	2
592	2	2	1	642	2	3	3	692	2	1	2
593	2	1	2	643	2	1	2	693	3	1	2
594	3	1	2	644	1	2	1	694	3	1	2
595	3	1	2	645	2	2	1	695	2	2	1
596	2	1	2	646	1	2	1	696	2	2	1
597	2	1	2	647	1	2	1	697	2	2	1
598	2	1	2	648	2	2	1	698	2	2	1
599	2	1	2	649	3	1	2	699	2	2	1
600	1	2	1	650	3	1	2	700	2	2	1
601	1	2	1	651	3	1	2	701	2	2	1
602	3	1	2	652	3	1	2	702	2	2	1
603	3	1	2	653	1	2	1	703	2	2	1
604	3	3	3	654	1	2	1	704	2	2	1
605	3	3	3	655	1	2	1	705	1	3	3
606	3	3	3	656	3	1	2	706	3	3	3
607	3	1	2	657	3	1	2	707	3	3	3
608	3	1	2	658	3	1	2	708	3	3	3
609	3	1	2	659	3	1	2	709	3	3	3
610	2	1	2	660	3	1	2	710	3	3	3
611	2	2	1	661	3	1	2	711	3	3	3
612	3	1	2	662	3	1	2	712	2	2	1
613	3	1	2	663	3	3	3	713	2	2	1
614	3	3	3	664	3	3	Minor B	714	1	2	1
615	3	3	3	665	3	3	3	715	1	2	1
616	3	3	3	666	2	3	3	716	1	2	1
617	2	3	3	667	2	1	2	717	2	3	3
618	1	3	3	668	2	1	2	718	3	3	3
619	2	3	3	669	3	1	2	719	2	2	1
620	2	3	3	670	3	1	2	720	1	2	1
621	2	3	3	671	3	1	2	721	1	2	1
622	1	3	3	672	3	3	3	722	3	1	2
623	1	2	1	673	1	3	3	723	2	1	2
624	3	1	2	674	2	3	3	724	2	3	3
625	3	1	2	675	3	3	3	725	1	2	1
626	1	2	1	676	1	2	1	726	1	2	1
627	3	1	2	677	3	3	3	727	1	2	1
628	3	1	2	678	3	1	2	728	2	2	1
629	3	1	2	679	2	3	3				
630	3	1	2	680	2	3	3				
631	3	1	2	681	2	3	3				
632	2	1	2	682	2	3	3				
633	2	1	2	683	2	3	3				
634	2	1	2	684	2	3	3				
635	2	1	2	685	2	2	1				

Minor A, B, C, and D indicates groups that do not belong to the three major clusters.

REFERENCES

1. Beckers, M.L.M., Buydens, L.M.C. (1998). Multivariate analysis of a data matrix containing A-DNA and B-DNA dinucleoside monophosphate steps: Multidimensional Ramachandran plots for nucleic acids. *Journal of Computational Chemistry, 19*(7), 695-715.

2. Loland, C.J., Norregaard, L., Gether, U. (1999). Defining proximity relationships in the tertiary structure of the dopamine transporter. *Journal of Biological Chemistry, 274*(52), 36928-36934.

3. Lin, Z., Wang, W. Kopajtic, T., Revay, R.S., Uhl, G.R.. (1999). Dopamine Transporter: Transmembrane phenylalanine mutations can selectively influence dopamine uptake and cocaine analog recognition. *Molecular Pharmacology, 56*, 434-447.

4. Carroll, F.I., Lewin, A.H., Boja, J.W., Kuhar, M.J. (1992). Cocaine receptor: Biochemical characterization and structure-activity relationship of cocaine analogues at the dopamine transporter. *Journal of Medicinal Chemistry, 22*, 3099-3108.

5. Lin, Z., Wang, W., Uhl, G.R. (2000). Dopamine transporter tryptophan mutants highlight candidate dopamine and cocaine selective domains. *Molecular Pharmacology, 58*, 1581-1592.

6. Kosten, T.R., George, T.P., Kosten, T.A. (2002). The potential of dopamine agonists in drug addiction. *Expert Opinion on Investigational Drugs, 11*(4), 491-499.

7. Reith, M.E.A., Berfield, J.L., Wang, L.C., Ferrer, J.V., Javitch, J.A. (2001). The uptake inhibitors cocaine and benztropine differentially alter the conformation of the human dopamine transporter. *Journal of Biological Chemistry, 276*(31), 29012-29018.

8. Schenck, J.O. (2002). The functioning neuronal transporter for dopamine: kinetic mechanisms and effects of amphetamines, cocaine, and methylphenidate. *Progress in Drug Research, 59*, 113-131.

9. Kimmel, H.L., Joyce, A.R. (2003). Dopamine Transporters: Structure-activity relationships and regulation in neurons. *Neurological Disease and Therapy, 56*, 467-500.

10. Singh, S. (2000). Design and structure-activity relationships of cocaine antagonists. *Chemical Reviews, 100*, 925-1024.

11. Dutta, A.K., Meltzer, P.C., Madras, B.K. (1993). Positional importance of the nitrogen atom in novel piperadine analogs of GBR 12909: Affinity and selectivity for the dopamine transporter. *Medicinal Chemistry Research, 3*, 209-222.

12. Prisinzano, T., Greiner, E., Johnsons, I., Dersch, C.M., Marcus, J., Partilla, J.S., Rothman, R.B., Jacobson, A.E., Rice, K.C. (2002). Piperadine analogues of GBR 12909: High affinity ligands for the dopamine transporter. *Journal of Medicinal Chemistry, 45*, 4371-4374.

13. Matecka, D., Lewis, D., Rothman, R.B., Dersch, C.M., Wojnicki, F.H.E., Glowa, J.R., De Vries, A.C., Pert, A., Rice. K.C. (1997). Heteroatomic analogs of 1-[2-(Diphenylmethoxy)ethyl]- and 1-[2-[Bis(4-flurophenyl)methoxy]ethyl]-4-3-phenylpropyl)piperazines (GBR 12935 and GBR 12909) as high-affinity dopamine reuptake inhibitors. *Journal of Medicinal Chemistry, 40*, 705-716.

14. Ben-Hur, A.G.I. (2003). Detecting stable clusters using principal component analysis. *Methods in Molecular Biology, 224*, 159-182.

15. Jackson, J.E. (1990). *A user's guide to principal component analysis*. New York: John Wiley and Sons, Inc.

16. Reijmers, T.H., Wehrens, R., Buydens, L.M.C. (2001). Circular effects in representations of an RNA nucleotides data set in relation with principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 56*, 61-71.

17. Wall, M.E., Rechtsteiner, A., Rocha, L.M. (2003). Singular value decomposition and principal component analysis. In D.P. Berrar (Ed.), *A practical approach to microarray data analysis* (pp.91-109). Norwell, MA: Kluwer.