

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen



The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

LIGAND-BASED DESIGN OF DOPAMINE REUPTAKE INHIBITORS: FUZZY RELATIONAL CLUSTERING AND 2-D AND 3-D QSAR MODELING

by

Milind Misra

As the three-dimensional structure of the dopamine transporter (DAT) remains undiscovered, any attempt to model the binding of drug-like ligands to this protein must necessarily include strategies that use ligand information. For flexible ligands that bind to the DAT, the identification of the binding conformation becomes an important but challenging task. In the first part of this work, the selection of a few representative structures as putative binding conformations from a large collection of conformations of a flexible GBR 12909 analogue was demonstrated by cluster analysis. Novel structure-based features that can be easily generalized to other molecules were developed and used for clustering. Since the feature space may or may not be Euclidean, a recently-developed fuzzy relational clustering algorithm capable of handling such data was used. Both superposition-dependent and superposition-independent features were used along with region-specific clustering that focused on separate pharmacophore elements in the molecule. Separate sets of representative structures were identified for the superposition-dependent and superposition-independent analyses.

In the second part of this work, several QSAR models were developed for a series of analogues of methylphenidate (MP), another potent dopamine reuptake inhibitor. In a novel method, the Electrotopological-state (E-state) indices for atoms of the scaffold common to all 80 compounds were used to develop an effective test set spanning both the structure space as well as the activity space. The utility of E-state indices in modeling a

series of analogues with a common scaffold was demonstrated. Several models were developed using various combinations of 2-D and 3-D descriptors in the Molconn-Z and MOE descriptor sets. The models derived from CoMFA descriptors were found to be the most predictive and explanatory. Progressive scrambling of all models indicated several stable models. The best models were used to predict the activity of the test set analogues and were found to produce reasonable residuals. Substitutions in the phenyl ring of MP, especially at the 3- and 4-positions, were found to be the most important for DAT-binding. It was predicted that for better DAT-binding the substituents at these positions should be relatively bulky, electron-rich atoms or groups.

**LIGAND-BASED DESIGN OF DOPAMINE REUPTAKE INHIBITORS: FUZZY
RELATIONAL CLUSTERING AND 2-D AND 3-D QSAR MODELING**

**by
Milind Misra**

**A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Chemistry**

Department of Chemistry and Environmental Science

January 2006

Copyright © 2006 by Milind Misra

ALL RIGHTS RESERVED

APPROVAL PAGE

LIGAND-BASED DESIGN OF DOPAMINE REUPTAKE INHIBITORS: FUZZY RELATIONAL CLUSTERING AND 2-D AND 3-D QSAR MODELING

Milind Misra

Dr. Carol A. Venanzi, Dissertation Advisor
Distinguished Professor of Chemistry, NJIT

Date

Dr. Joseph W. Bozzelli, Committee Member
Distinguished Professor of Chemistry, NJIT

Date

Dr. Tamara Gund, Committee Member
Professor of Chemistry, NJIT

Date

Dr. Michael L. Recce, Committee Member
Associate Professor of Information Systems, NJIT

Date

Dr. Edgardo T. Farinas, Committee Member
Assistant Professor of Chemistry, NJIT

Date

BIOGRAPHICAL SKETCH

Author: Milind Misra
Degree: Doctor of Philosophy
Date: January 2006

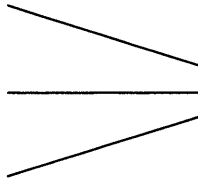
Undergraduate and Graduate Education:

- Doctor of Philosophy in Chemistry,
New Jersey Institute of Technology, Newark, NJ, 2006
- Master of Science in Computational Biology,
New Jersey Institute of Technology, Newark, NJ and
Rutgers the State University of New Jersey, Newark, NJ, 2003
- Master of Science in Applied Chemistry,
New Jersey Institute of Technology, Newark, NJ, 1999
- Bachelor of Science in Engineering Science,
New Jersey Institute of Technology, Newark, NJ, 1998

Major: Chemistry

Publications:

- A. Fiorentino, D. Pandit, K.M. Gilbert, M. Misra, R. Dios, and C. A. Venanzi, "Singular value decomposition of torsional angles of analogs of the dopamine reuptake inhibitor GBR 12909", *Journal of Computational Chemistry*, in press.
- M. Misra, A. Banerjee, R. N. Davé, and C. A. Venanzi, "Novel feature extraction technique for fuzzy relational clustering of a flexible dopamine reuptake inhibitor", *Journal of Chemical Information and Modeling*, 2005 **45**, 610-623.
- K. M. Gilbert, W. J. Skawinski, M. Misra, K. A. Paris, N. H. Naik, R. A. Buono, H. M. Deutsch and C. A. Venanzi, "Conformational analysis of methylphenidate: comparison of molecular orbital and molecular mechanics methods", *Journal of Computer-Aided Molecular Design*, 2004 **18**, 719-738.



“From darkness to light” is a common refrain in scriptures of all pasts; “Create a one-way from ignorance to enlightenment,” say the ancients, with supreme sanction. My simple *Three Line Diagram* is no less of an authority: For it states the Old in the new guise of Geometry, and helps examine the Point that is the Self.

To my family

ACKNOWLEDGMENT

I am grateful to my advisor, Professor Carol A. Venanzi, for giving me the space to exercise some of my ideas and for her support. I am also indebted to her for my introduction to professional meetings and organizations, the development of my scientific writing, and the funding for my last several months at NJIT.

I thank the Department of Chemistry and Environmental Science for providing tuition assistance and a stipend for many semesters of my Ph.D. I am grateful to Ms. Gayle Katz and the Office of Graduate Studies for making the funding process and other administrative procedures easy.

I am grateful to Professor Joseph W. Bozzelli, Professor Tamara Gund, Professor Michael L. Recce, and Professor Edgardo T. Farinas for comments, suggestions, and corrections as members of my dissertation committee.

Dr. Amit Banerjee, my colleague and friend, provided critical insights and stimulating conversations. I thank Dr. Banerjee and Professor Rajesh N. Davé for their invaluable collaboration. Professor William J. Skawinski merits mention for continuous inspiration and for showing interest in special conversations that were almost always unrelated to chemistry. He helped me with the title of this dissertation. Dr. Kathleen M. Gilbert deserves mention for providing useful contrasts of opinion and style. I thank her for suggesting the website on which I found the advertisement that led to a post-doctorate position. I am indebted to Ms. Deepangi N. Pandit for her ready help in preparing me for my interview for the above position and for her understanding of personal issues.

Finally, I acknowledge the tremendous role that Ms. Tulika Singh has played in my transformation from one who thought he knew to one who knows he does not.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION.....	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Significance	2
1.4 Background	3
1.4.1 Dopamine Reuptake Inhibitors	3
1.4.2 Pharmacophore Modeling	5
1.4.3 Ligand-Based Drug Design	6
1.4.4 Cluster Analysis	7
1.4.5 3-D QSAR Approach	8
1.4.6 2-D QSAR Approach	9
1.4.7 Molecular Descriptors	9
1.5 Overview	10
2 CLUSTER ANALYSIS USING NOVEL FEATURES.....	11
2.1 Data Clustering	12
2.2 Fuzzy Relational Clustering	17
2.3 Fuzzy Cluster Validity Measures	19
2.4 Overview of Novel Feature Extraction Techniques	21
2.4.1 Superposition-Dependent Features	22
2.4.2 Superposition-Independent Features	23

TABLE OF CONTENTS (Continued)

Chapter	Page
2.4.2.1 Nucleic Acid Structure Analysis Programs	24
2.4.2.2 The 3DNA Program	24
2.4.2.3 The Planes Program	26
3 FUZZY CLUSTERING: METHODS	28
3.1 Conformational Analysis	28
3.2 Superposition-Dependent Features	29
3.2.1 Superposition 1	30
3.2.1.1 Feature Extraction for A-Side Clustering, Superposition 1	30
3.2.1.2 Feature Extraction for B-Side Clustering, Superposition 1	32
3.2.1.3 Feature Extraction for Full-Molecule Clustering, Superposition 1	33
3.2.2 Superposition 2	33
3.2.2.1 Feature Extraction for B'-Side Clustering, Superposition 2	33
3.2.3 Determining the Angle between Two Planes	34
3.2.4 Fuzzy Clustering	35
3.3 Superposition-Independent Features	36
3.3.1 Calculation of Planes Parameters	36
3.3.1.1 Definition of Standards	36
3.3.1.2 Calculation of Translational and Rotational Parameters	37
3.3.2 Feature Vectors	39
3.3.3 Proximity Matrices	39

TABLE OF CONTENTS

(Continued)

Chapter	Page
3.3.4 Fuzzy Clustering	41
4 SUPERPOSITION-DEPENDENT CLUSTERING RESULTS	42
4.1 Conformational Analysis	42
4.2 Clustering	42
4.2.1 Full-Molecule Clustering, Superposition 1	42
4.2.2 A-Side Clustering, Superposition 1	43
4.2.3 B-Side Clustering, Superposition 1	50
4.2.4 B'-Side Clustering, Superposition 2	51
4.3 Identification of Full-Molecule Representative Structures	54
4.4 Generalization of the Feature Extraction Method	55
4.5 Comparison to Hierarchical Clustering Using XCluster	58
4.6 Use of Torsional Angles as Feature Vectors?	61
4.7 Different Superpositions	62
5 SUPERPOSITION-INDEPENDENT CLUSTERING RESULTS	63
5.1 Clustering	63
5.1.1 Full-Molecule Clustering	64
5.1.2 B-Side Clustering	68
5.1.3 A-Side Clustering	72
5.2 Identification of Representative Conformations	74
5.3 Discussion	75

TABLE OF CONTENTS (Continued)

Chapter	Page
6 QSAR OF METHYLPHENIDATE ANALOGUES: METHODS	78
6.1 Introduction	78
6.2 Data Preparation	83
6.2.1 MPN Data Set	84
6.2.2 MPP Data Set	85
6.3 Data Analysis I: Exploratory Models	87
6.3.1 Forward Stepwise Regression Analyses and Scaling	87
6.3.2 Identification of Test Set Analogues	89
6.3.2.1 Tanimoto Coefficient	89
6.3.2.2 Sphere Exclusion Algorithms	90
6.3.2.3 D-SIM Version 1.0	92
6.3.2.4 Unsupervised Forward Selection of Redundant Descriptors	93
6.3.3 All Possible Subsets Regression Analyses	93
6.4 Data Analysis II: Robust Models	94
6.4.1 Calculation of Descriptors	94
6.4.2 PLS Analyses	94
6.5 Model Validation	95
6.5.1 Progressive Scrambling	95
6.5.2 Test Set Validation	96
7 QSAR OF METHYLPHENIDATE ANALOGUES: RESULTS	97

TABLE OF CONTENTS (Continued)

Chapter	Page
7.1 Forward Stepwise Regression	97
7.2 Test Set Identification	100
7.3 All Possible Subsets Regression	103
7.4 Partial Least Squares Analyses	105
7.5 Model Validation	110
7.5.1 Progressive Scrambling Results	110
7.5.2 External Validation	111
7.6 Data Interpretation and Predictions	112
7.7 Discussion	116
8 CONCLUSIONS	122
APPENDIX A ADDITIONAL RESULTS FOR SUPERPOSITION- INDEPENDENT CLUSTERING	124
APPENDIX B MP ANALOGUE DATA SET	131
APPENDIX C MINIMIZATION AND SEARCH PARAMETERS	135
APPENDIX D TORSIONAL ANGLES FOR GEM CONFORMATIONS	136
APPENDIX E COMFA PARAMETERS	137
APPENDIX F MATLAB COMMANDS	138
APPENDIX G D-SIM VERSION 1.0 PROGRAM CODE	146
APPENDIX H ALL POSSIBLE SUBSETS REGRESSION RESULTS	163
APPENDIX I PATH INFORMATION FOR RESEARCH FILES	171
REFERENCES	172

LIST OF TABLES

Table	Page
3.1 Summary of Feature Vectors Used for the Two Superpositions	32
4.1 Torsional Angles and Relative Energies of Full-Molecule Representatives	55
4.2 Summary of XCluster Studies	60
5.1 Superposition-Independent Clustering Results	63
5.2 Torsional Angles and Relative Energies for Representative Conformations ...	74
7.1 Forward Stepwise Regression Results	98
7.2 Test Set Identification Results	101
7.3 All Possible Subsets Regression: Test Set Residuals	105
7.4 Description of PLS Models	106
7.5 Results of PLS Analyses for the MPN Data Set	107
7.6 Results of PLS Analyses for the MPP Data Set	108
7.7 Test Set (TS3) Activity Prediction	112
B.1 MP Analogue Data Set	127
B.2 Data Set and Test Set Groups	130
D.1 Torsional Angles for MPN and MPP GEM Conformations	130

LIST OF FIGURES

Figure	Page
1.1 Some of the classes of DAT reuptake inhibitors	4
2.1 Structures of GBR 12909 and 1	11
2.2 Elements of the modified feature vector for the B'-side only	17
2.3 Base-step parameters used in nucleic acid structure analysis	26
3.1 Reconstruction sequence for the A-side	31
3.2 Determination of the angle between two planes	35
4.1 Side view of the 728 conformations of 1 , Superposition 1	43
4.2 Cluster validity plots for partitions on the A-side	44
4.3 Torsional angles definitions and conformations of 1 in (A1, A2) space	48
4.4 Results for the A-side clustering at $c = 3$ and $c = 6$, Superposition 1	50
4.5 Cluster validity plots for partitions on the B-side	51
4.6 Cluster validity plots for partitions on the B'-side	51
4.7 Conformations of 1 in (B3, B4) space	53
4.8 Clustering results for the B'-side at $c = 9$	54
4.9 Full-molecule representative structures	55
4.10 Identification of reduced feature set for cocaine	58
4.11 Features selected in XCluster studies	59
5.1 Cluster validity plots for the $[N \times P2]_{T+R}$ proximity matrix	64
5.2 (Slide, Shift, Rise) space for $[N \times P2]_{T+R}$, $c = 5$	65
5.3 (Slide, Rise) space for $[N \times P2]_{T+R}$, $c = 5$	65

LIST OF FIGURES (Continued)

Figure	Page
5.4 Cluster validity plots for the $[N \times P2]_T$ proximity matrix	66
5.5 (Slide, Shift, Rise) space for $[N \times P2]_T$, $c = 5$	66
5.6 (Roll, Tilt, Twist) space for $[N \times P2]_{T+R}$, $c = 5$	67
5.7 Cluster validity plots for the $[N \times P2]_R$ proximity matrix	67
5.8 Cluster validity plots for the $[C \times P2]_{T+R}$ proximity matrix	68
5.9 (Slide, Shift, Rise) space for $[C \times P2]_{T+R}$, $c = 3$	69
5.10 (Shift, Rise) space for $[C \times P2]_{T+R}$, $c = 3$	69
5.11 (Slide, Rise) space for $[C \times P2]_{T+R}$, $c = 3$	70
5.12 Cluster validity plots for the $[C \times P2]_T$ proximity matrix	71
5.13 Cluster validity plots for the $[C \times P2]_R$ proximity matrix	71
5.14 (Slide, Shift, Rise) space for $[C \times P2]_T$, $c = 3$	71
5.15 Cluster validity plots for the $[N \times C]_{T+R}$ proximity matrix	73
5.16 (Slide, Shift, Rise) space for $[N \times C]_{T+R}$, $c = 9$	73
5.17 (Roll, Tilt, Twist) space for $[N \times C]_{T+R}$, $c = 9$	73
6.1 The scaffold for the MP analogue data set	83
6.2 The MPN and MPP GEM conformations	85
6.3 Schematic for sphere-exclusion algorithms	92
7.1 CoMFA maps for mpn_c1_trn and mpp_c1_trn models	114
A.1 (Shift, Rise) space for $[N \times P2]_{T+R}$, $c = 5$	124
A.2 (Shift, Slide) space for $[N \times P2]_{T+R}$, $c = 5$	124

LIST OF FIGURES (Continued)

Figure	Page
A.3 (Roll, Twist) space for $[N \times P2]_{T+R}$, $c = 5$	125
A.4 (Tilt, Roll) space for $[N \times P2]_{T+R}$, $c = 5$	125
A.5 (Tilt, Twist) space for $[N \times P2]_{T+R}$, $c = 5$	125
A.6 Cluster validity plots for the $[N \times P1]_{T+R}$ proximity matrix	126
A.7 (Slide, Shift, Rise) space for $[N \times P1]_{T+R}$, $c = 5$	126
A.8 Cluster validity plots for the $[N \times P1]_T$ proximity matrix	126
A.9 (Roll, Tilt, Twist) space for $[C \times P2]_{T+R}$, $c = 3$	127
A.10 (Shift, Slide) space for $[C \times P2]_{T+R}$, $c = 3$	127
A.11 (Roll, Twist) space for $[C \times P2]_{T+R}$, $c = 3$	127
A.12 (Tilt, Roll) space for $[C \times P2]_{T+R}$, $c = 3$	128
A.13 (Tilt, Twist) space for $[C \times P2]_{T+R}$, $c = 3$	128
A.14 Cluster validity plots for the $[C \times P1]_{T+R}$ proximity matrix	128
A.15 (Slide, Shift, Rise) space for $[C \times P1]_{T+R}$, $c = 3$	129
A.16 Cluster validity plots for the $[C \times P1]_T$ proximity matrix	129
A.17 Cluster validity plots for the $[N \times C]_T$ proximity matrix	129
A.18 Cluster validity plots for the $[N \times C]_R$ proximity matrix	130

CHAPTER 1

INTRODUCTION

1.1 Motivation

Since the introduction of crack in the mid-1980's, cocaine abuse has been an epidemic in the U.S. The abuse of cocaine and other central nervous system (CNS) stimulants has greatly affected public health because of the associated spread of HIV-1, hepatitis B and C, and drug resistant tuberculosis. According to government estimates,¹ the annual demand for cocaine in the U.S. is about 300 metric tons, which is about half the total world production of cocaine. Another estimate places the number of hardcore users of cocaine in the U.S. at about 3.5 million every year.¹ The high associated social and economic costs of treatment and prevention are a continuing motivation for the development of effective cocaine and other CNS stimulant abuse therapeutics.

Cocaine binds to the dopamine transporter (DAT) in the brain and is believed to produce its euphorogenic and addictive effects by inhibiting the reuptake of synaptic dopamine into presynaptic neurons. Consequently, several classes of dopamine reuptake inhibitor compounds are currently being pursued as possible cocaine abuse therapeutics. These include the 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine (GBR 12909) and methylphenidate (MP) classes. For the current work, a large number of analogues from both these classes was available through separate collaborations with chemists who synthesized these analogues and measured their binding affinity with the DAT. Several computational strategies were explored to aid in the development of lead compounds for treatment of cocaine abuse.

1.2 Objectives

There were two main objectives of this work:

- To identify putative bioactive conformations for use in 3-D quantitative structure activity relationship (QSAR) analyses by applying fuzzy relational clustering to over 700 conformations of a flexible GBR 12909 analogue. To develop novel feature selection schemes by using feature spaces based upon molecule-specific structural properties. To demonstrate the usefulness of this approach in the classification of conformers of flexible molecules.
- To perform QSAR analyses using structural descriptors for 80 MP analogues. To derive models that could be used to predict the bioactivity of a test set of analogues.

1.3 Significance

The significance of this work lies in the methodology and the data set to which it is applied. The GBR 12909 and MP classes of dopamine reuptake inhibitor studied here are two of the most promising for the treatment of cocaine abuse. To date no extensive molecular modeling studies have been carried out on these drugs.

The present work includes the first application of fuzzy relational clustering to the classification of molecular conformations and demonstrates the utility of this approach for all types of flexible molecules. It also includes the first calculation of topological and other structural descriptors for a large data set of MP analogues and the attempt to find the relationship between these descriptors and the biological activity of these analogues.

1.4 Background

1.4.1 Dopamine Reuptake Inhibitors

The DAT has been implicated in cocaine abuse and addiction by the “dopamine hypothesis”.² The DAT is located on the cell membrane of dopaminergic nerve terminals and is responsible for the termination of dopamine neurotransmission and maintenance of homeostasis by transporting synaptic dopamine into the presynaptic neuron.³ According to the dopamine hypothesis, cocaine blocks this reuptake of dopamine by binding to the DAT and leads to an elevated level of extracellular dopamine that is believed to be the main reason for addiction. One approach to testing the dopamine hypothesis and finding an effective treatment for cocaine abuse is to develop a noncompetitive inhibitor of cocaine that should have a low intrinsic activity and should strongly bind to but slowly dissociate from the DAT.^{4,5} Such an inhibitor would be able to withstand an increase in cocaine self-administration resulting from loss of reward due to inhibitor action. Since cocaine also binds to the serotonin transporter (SERT), the ideal compound would have high selectivity for the DAT relative to the SERT.

Figure 1.1 shows some classes of dopamine reuptake inhibitors that act like cocaine by binding to the DAT but may not share the same abuse potential. Structure-activity relationship (SAR) studies for several classes of dopamine reuptake inhibitors have been reviewed⁶ and provide a large amount of data for pharmacophore modeling.

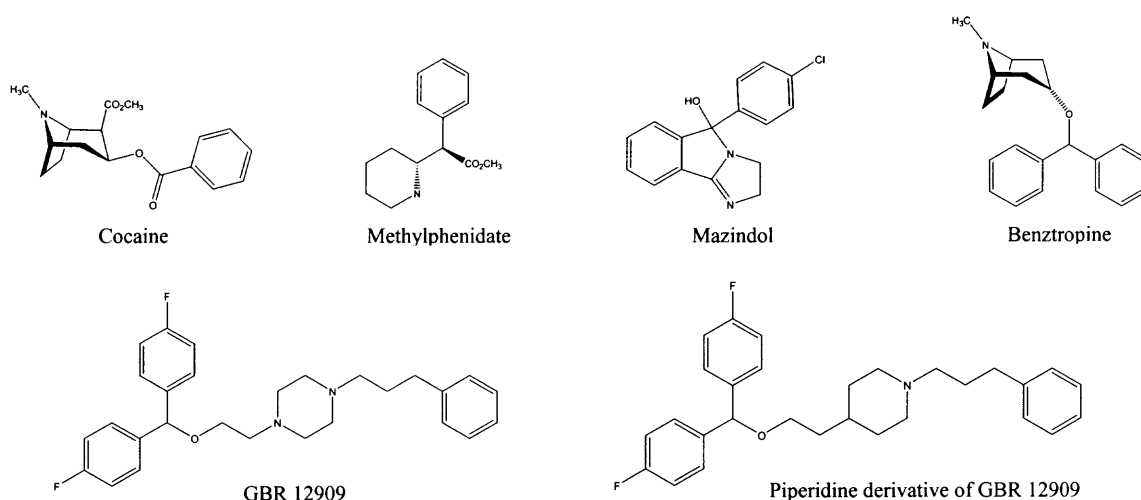


Figure 1.1 Some of the classes of DAT reuptake inhibitors.

Quantitative Structure-Activity Relationship (QSAR) studies and molecular modeling have been carried out on various classes of dopamine reuptake inhibitors including tropanes,⁷⁻¹⁶ benztropine,^{17,18} BTCP,¹⁹ mazindol,²⁰ MP,²¹⁻²⁵ GBR 12909 analogues,²⁶⁻³⁰ and novel piperidinols.³¹ Two interesting classes of dopamine reuptake inhibitors are the MP and the GBR 12909 classes. The mechanism of action of MP (Ritalin®) is similar to that of cocaine. However, because of its limited abuse potential (it has been prescribed by pediatricians to children with attention deficit hyperactivity disorder), there has been considerable interest in MP and its analogues.^{32,33} A wide range of MP SAR studies has been produced by the Deutsch and Schweri laboratories.³⁴⁻⁴⁴ This MP data was provided to Professor Venanzi's group by Dr. Howard Deutsch of the Georgia Institute of Technology, who synthesized the analogues, and Dr. Margaret Schweri of the Mercer University School of Medicine, who measured the DAT binding affinity (IC₅₀) of the analogues.

The GBR 12909 class⁴⁵ is one of the most promising candidates for the treatment of cocaine dependence. GBR 12909 (1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine) has been found to be effective in reducing cocaine self-administration in rhesus monkeys without significantly affecting food-maintained responding^{46,47} and has completed phase I clinical trials.⁴⁵ Dutta et al.⁴⁸ showed that only one of the two nitrogens in the central piperazine ring was required for activity at the dopamine transporter and, subsequently, both the Rice^{49,50} and Dutta⁵¹ groups have pursued SAR studies of piperidine analogues of GBR 12909. However, the SAR of the piperidine analogues may be different than that of the piperazines^{49,52,53} since the same structural modifications result in different relative levels of activity for the members of each series. The GBR 12909 analogue data was provided to Professor Venanzi's group by Dr. Kenner Rice of the National Institute of Diabetes and Digestive and Kidney Diseases and Dr. Richard Rothman of the National Institutes of Health. By choosing to work on GBR 12909 analogues, the Venanzi group focuses on the difficult problems associated with 3-D QSAR analysis of flexible molecules, e.g., identification of putative bioactive conformations that can be used in such analyses.

1.4.2 Pharmacophore Modeling

The pharmacophore model for the ideal dopamine reuptake inhibitor defines the three-dimensional geometric arrangement between the chemical moieties (e.g., a carbonyl group or aromatic ring) or chemical features (e.g., a hydrogen bond donor or acceptor) that have been identified from the experimental structure-activity data as being required for biological activity (or "bioactivity"). A pharmacophore model (or pharmacophore) is

defined by the orientation of these features in the bioactive conformation of the ligand, i.e., the conformation in which it binds to the protein. Flexible ligands, in particular, may have more than one conformation that can adjust to fit the pharmacophore model with little expenditure of energy. The pharmacophore model can be used as a template from which synthetic chemists can design a new structure-activity series or a rigid analogue that can be used to validate and refine the pharmacophore model. Such validation can lend support to the hypothesized bioactive conformer from which the pharmacophore model was derived.

1.4.3 Ligand-Based Drug Design

In drug design, the identification of the bioactive conformation of a promising drug candidate is of great interest. In the absence of structural information about the receptor, the prediction of the bioactive conformation of the ligand can be very challenging. While the amino acid sequence of the DAT has been identified,⁵⁴ its three-dimensional structure is not known.⁵⁵ Further, there is considerable evidence that drug molecules do not bind to proteins in their vacuum-phase global energy minimum (GEM) conformation.⁵⁶⁻⁵⁹ Conformational searching techniques can be used to explore the conformational space of a ligand and locate minima on the ligand's potential energy surface. For a flexible molecule (for example, a GBR 12909 analogue that has 8-12 torsional angles), the number of minima generated can be very large. This prohibits consideration of every minimum conformation as a putative bioactive conformation. However, the importance of considering conformations other than the GEM has been shown to be significant in pharmacophore modeling.⁶⁰⁻⁶⁵ Therefore, selection of a suitable set of representative

conformers for analysis is an important first step in 3-D QSAR techniques, such as Comparative Molecular Field Analysis (CoMFA).⁶⁶ In such cases, it becomes essential to use data reduction techniques such as clustering in order to first identify well-defined groups of conformations and then select representative structures from each group. Each of the representative structures can then be used as a putative bioactive conformation for modeling studies.

1.4.4 Cluster Analysis

Data clustering as a means of classification has been used extensively in computational chemistry and has been recently reviewed.⁶⁷ Clustering of molecules in chemical databases has received considerably more attention than clustering of molecular conformations. Attempts to cluster conformations have been based on some sort of proximity measure between pairs of conformations. Fuzzy clustering is a partition-based clustering scheme and is particularly useful when there are no apparent clear groupings in the data set. For fuzzy clustering, since every individual conformation belongs to not one but all clusters with varying degrees of membership, the clustering results can provide a natural interpretation of the goodness of the partition. This is useful for clustering of a large number of conformations of a flexible molecule where overlap of cluster boundaries is expected. Davé and coworkers⁶⁸ have developed a fuzzy relational clustering algorithm. The present work was part of a collaboration between the Davé and Venanzi groups to apply this technique to the classification of conformations of GBR 12909 analogues. Specifically, the novel feature extraction techniques described in Chapter 3 were developed through close collaboration with Amit Banerjee of the Davé group. In

addition, Amit Banerjee used his C++ implementation of the fuzzy relational clustering algorithm to perform all clustering calculations. The conformations that most closely represent each identified cluster can be used as putative bioactive conformations in 3-D QSAR analyses.

1.4.5 3-D QSAR Approach

The putative bioactive conformations identified after conformational and cluster analyses serve as the starting points for 3-D QSAR analyses such as CoMFA. CoMFA is based on the assumption that the interactions between the ligand and its receptor site are primarily noncovalent in nature. CoMFA is performed by calculating and comparing the molecular steric (Lennard Jones) and electrostatic (Coulombic) “fields” of suitably superimposed analogues. Each putative bioactive conformation is used as the template for the superposition of all the analogues and a separate CoMFA study is carried out for each superposition. These fields are calculated at each lattice point around each analogue using a probe atom, such as a sp^3 carbon atom with +1 charge, at regularly-spaced points on the three-dimensional CoMFA grid. The energy values thus calculated are entered into columns in a CoMFA QSAR table. A multivariate statistical analysis routine such as partial least squares then attempts to find a relationship between the predictor variables (i.e., steric and electrostatic interaction energies) and the response variable (i.e., the experimental IC_{50} values). 3-D QSAR techniques like CoMFA are thus heavily dependent on 1) the selection of the putative bioactive conformation used as the template for superposition and 2) the selection of a suitable scheme for alignment of the analogues for placement in the three-dimensional CoMFA grid. Often these two procedures are

subjective and time-consuming, and could compound the problem by being interdependent.

1.4.6 2-D QSAR Approach

When a QSAR analysis is required for a dataset containing a large number of analogues, the 3-D approach can be time-consuming and impractical for reasons noted above. Since 2-D QSAR analyses depend upon molecular connectivity and not upon the conformation of the analogues, they can be used instead of 3-D QSAR analyses when, for example, high throughput screening of large datasets is required. A suitable selection of molecular descriptors is used to develop model(s) for prediction of bioactivity of a novel analogue. However, the number of molecular descriptors available can be very large and to obtain a valid QSAR model it becomes essential to restrict the selection to the ones that are the most important.

1.4.7 Molecular Descriptors

Structural descriptors such as topological indices represent nonempirical information about molecular structure that can be useful in relating structure to properties in 2-D QSAR analyses. Topological indices⁶⁹ (and the variable of molecular structure that they encode) include: Chi indices (molecular connectivity); Kappa indices (molecular shape and flexibility); and electrotopological state indices (E-State, which encodes both topological and electronic information and is correlated with electronegativity). Other structural descriptors include counts of graph paths, atoms, atoms types, bond types, rings, etc. and information indices such as the Shannon and the Bonchev-Trinajstić

information indices. These indices have been widely used in QSAR analyses and applications⁶⁹ such as anesthetic potency, hallucinogenic activity, enzyme inhibition, bioconcentration factors, toxicity, carcinogenicity, soil sorption, solubilities, boiling points, densities, molar refraction retention, and gas chromatographic retention. Three-dimensional descriptors are conformation-dependent and may or may not depend upon a coordinate reference frame. The CoMFA steric and electrostatic field interaction values are also 3-D descriptors.

1.5 Overview

This dissertation is divided into two main parts in terms of the research objectives and the methods used. The first part is comprised of Chapters 2-5 and focuses on identification of representative conformations of a flexible GBR 12909 analogue by using cluster analysis techniques. The material in these chapters has been the basis for published²⁷ and submitted²⁸ work done in the course of this research. Chapter 2 provides the background of data clustering in general and of the recently-developed fuzzy relational clustering in particular. It also presents the rationale for the novel feature extraction techniques developed in this work. Chapter 3 describes the feature extraction and clustering methods in detail. Chapters 4 and 5 provide the results of superposition-dependent and superposition-independent cluster analyses, respectively. The second part consists of Chapters 6 and 7 and deals with QSAR studies on methylphenidate and its analogues. Chapter 6 describes the methods used in these studies while Chapter 7 provides the results. The last chapter presents the overall conclusions garnered from this work.

CHAPTER 2

CLUSTER ANALYSIS USING NOVEL FEATURES

This chapter and the next present clustering studies of conformations of **1**⁴⁶ (Figure 2.1), an analogue of 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine (GBR 12909). Analogues of GBR 12909 belong to a class of dopamine reuptake inhibitors that may be potentially useful in the treatment of cocaine abuse.⁴⁵ GBR 12909 has been found to be effective in reducing cocaine self-administration in rhesus monkeys without significantly affecting food-maintained responding^{46,47} and has completed phase I clinical trials.⁴⁵

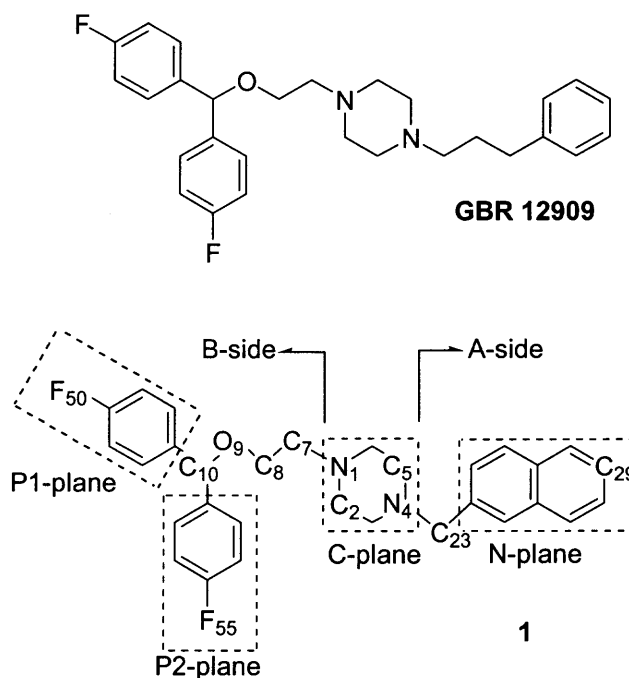


Figure 2.1 Structures of GBR 12909 and **1**.

The purpose of cluster analysis of the conformations of flexible GBR 12909 analogues is to identify a small number of representative conformations that could aid in

understanding the interaction between the analogues and the DAT. These representative conformations will be used in a future study as templates for molecular alignment in CoMFA studies of a large set of GBR 12909 analogues.

2.1 Data Clustering

Data clustering as a means of classification has been used extensively in computational chemistry and a thorough review of techniques is available.⁶⁷ The most popular clustering techniques for identification of representative conformers are hierarchical techniques such as the single-link clustering and the average-link clustering schemes. The single-link clustering package XCluster⁷⁰ clusters conformations based on a root mean square (RMS) distance matrix derived either from a set of atom coordinates (with or without rigid body superposition of conformations) or from a set of torsional angles. An average-link clustering technique has been used to demonstrate clustering of 63 conformations of a tripeptide fragment based on a Euclidean distance measure of proximity in a 36-dimensional space.⁷¹ However, as data sets become larger, techniques based on dendrograms are impractical for more than a few hundred patterns.⁷² Another problem is that such techniques may tend to find singleton clusters unless carefully-selected termination criteria are utilized. Other general disadvantages of hierarchical schemes include sensitivity to outliers, since sources of error and variance are not considered, and since there is no relocation of the objects along the hierarchy, objects once assigned *incorrectly* cannot be reassigned.⁷³ From an application point of view, hierarchical schemes are particularly useful for detecting sequential levels of clustering, such as in taxonomy and in biological classification.

In contrast, partitional clustering techniques have been developed to uncover a single best partition of the data.⁷⁴ Unlike hierarchical clustering, partitional clustering techniques produce a single c-partition of the data based on a clustering criterion, which can either be a global criterion (such as minimization of square-error or maximization of expectation) or a local criterion (such as estimation of regions of high density, a nearest neighbor (NN) criterion). The global clustering criterion is also called an objective-function-based optimization and such schemes represent each cluster by a *prototype* and assign objects to clusters according to the most similar prototype. A variant of the NN-based local clustering criterion, called the nearest single neighbor method,⁷⁵ was used to cluster different sets of peptide conformations and was based on a proximity measure derived from RMS distances between pairs of conformations defined by only the peptide backbone structure. While NN techniques have been shown to be useful, they tend to be computationally expensive for large data sets.

Recently, attempts have been made to cluster families of conformations using statistical scaling techniques as cluster analysis tools. For example, families of relatively small or rigid molecules such as dopamine, roseotoxin-B, and cycloheptadecane were clustered by first scaling the higher dimensional data in real space to a reduced 3-D conformational space using both multidimensional and metric scaling techniques. Then either visual inspection or a hierarchical technique applied to a proximity matrix derived from the reduced 3-D data set was used to complete the clustering.⁷⁶ Subsequently, the same 3-D data set was clustered using fuzzy clustering.⁷⁷ These appear to be the only instances where partitional clustering schemes have been successfully applied to cluster families of conformations.

Fuzzy c-means⁷⁸ and a large number of its derivative algorithms⁷⁹ are partitional clustering schemes, which are based on different types of square-error minimization and specifically target detection of cluster prototypes, ranging from a simple prototype that is an n-dimensional point to complex shapes that include non-linear surfaces and manifolds in n-dimensional space.⁸⁰ The motivation to use an objective-function-based partitional scheme in the present work, instead of hierarchical or local criterion-based partitional techniques, comes from the fact that such schemes are best suited for efficient representation and compression of large data sets. They not only inherently search for natural groupings, but also find the cluster prototypes, i.e., ideal representatives, which in the present case are the most representative molecular conformations. While there are many advantages of such clustering methods, they suffer from two problems that face all clustering methods, including hierarchical clustering, namely, susceptibility to noise or outliers and difficulty in determining the exact number of clusters within the data. Fortunately, the recent development of robust clustering algorithms⁸¹ and highly meaningful cluster validity measures^{82,83} address these two problems. Availability of these algorithms and recent examples of the application of fuzzy clustering by the Feher group^{76,77} provide a motivation for utilizing such methods for the present work.

The use of fuzzy memberships has several advantages. First, they are particularly useful when there are no easily-identifiable, clear groupings in the data set. Second, having fuzzy memberships is generally helpful in terms of smoother convergence of the numerical algorithm as compared to the use of non-fuzzy (also termed crisp or hard) memberships.⁸⁴ Moreover, partitioning schemes provide automatic detection of cluster boundaries. In the case of fuzzy clustering, these cluster boundaries can overlap. Use of

fuzzy memberships is therefore advantageous in the present application because overlap of cluster boundaries is expected since the flexibility of the molecule results in a continuous "spectrum" of closely-related conformers. For fuzzy clustering, since every individual data entity (a conformer, in the present case) belongs to not one but all clusters with varying degrees of membership, the clustering results can provide a natural interpretation of the goodness of the partition. The fuzzy membership (a numerical quantity between 0 and 1) is directly related to the structure of the partition. Hence almost all the fuzzy cluster validity measures are based on fuzzy memberships and even though they do not take the cluster geometry into account, they often provide meaningful, interpretable results.^{85,86} Based on these well-known advantages of fuzzy memberships, it was decided to utilize fuzzy clustering algorithms. However, the most important aspect of the present application is that it incorporates the use of an objective-function-based partitional scheme instead of methods such as hierarchical clustering or multivariate data analysis.

The present fuzzy clustering approach differs from that of Feher and Schmidt⁷⁷ in several ways. Since there are problems associated with labeling the symmetric heavy atoms in the phenyl ring (discussed further in the next section), the present approach is based on reducing the raw data to a manageable form, taking care of the symmetry issue in the process. A proximity-based distance measure using all available heavy atom information appears to be more relevant than Feher and Schmidt's approach of scaling down the data by considering only the symmetry-unique atoms. The use of a proximity matrix also means that the clustering should be performed on the relational data domain instead of the object-space-based clustering used by Feher and Schmidt. Further, the use

of a proximity measure facilitates incorporation of heuristic or many popular non-Euclidean similarity measures (for example, L_1 norm, p norm, Manhattan norm, etc.), implying that the relational data is not necessarily based on Euclidean measure. For clustering non-Euclidean relational data, the NERFCM (non-Euclidean relational fuzzy c-means),⁸⁷ which is the relational dual of fuzzy c-means (FCM),⁷⁸ is a popular choice. This method was initially utilized in the present work and was also compared with the newly-developed Fuzzy Relational Clustering (FRC) procedure.⁶⁸ Both the methods were found to provide nearly identical results. However, FRC was used for the results reported here²⁷ because it does not require the beta-spread transformation that is needed in NERFCM. Hence it is computationally attractive and its performance can be further improved by using a Seidel iterative scheme. In terms of the overall methodology developed here, the use of FRC is indicated for all future applications (as for example, in Banerjee et al.²⁸) because it can be easily utilized in its robust version to handle noisy data, although that is not highly relevant to the case considered here.

Besides the potential advantages listed above, the present approach is novel for several reasons. First, it is the only fuzzy clustering study of a very flexible molecule. Second, region-specific clustering that focused on individual pharmacophore elements of the molecule was made possible by defining feature vectors in terms of the A- and B-side (see Figure 2.1) or A'- and B'-side (see Figure 2.2) moieties which contain important chemical features of the pharmacophore. Third, fuzzy relational clustering was performed using a) a novel superposition-dependent feature extraction technique (see Section 2.4.1) and b) superposition-independent features derived from relative orientation of molecular planes (see Section 2.4.2). As described in the next section, the FRC

procedure used proximity matrices derived from feature vectors that contained real space elements (atom coordinates and angles between planes for superposition-dependent clustering and planes parameters for superposition-independent clustering) that were related to the pharmacophore elements of the molecule. As a result, the feature vectors described in this work are unique to **1** since they were derived from the geometric constraints of this particular molecule. However, as will be illustrated in Chapter 3, the feature extraction techniques can be generalized to other molecules.

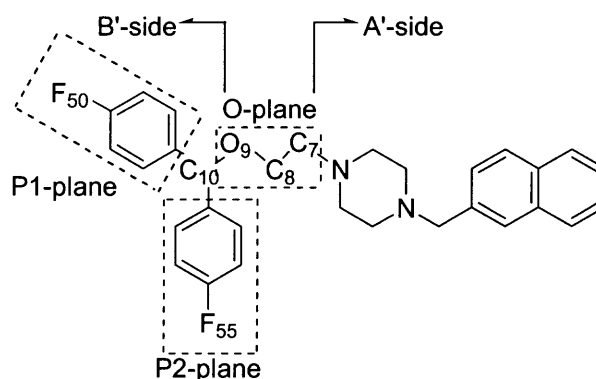


Figure 2.2 Elements of the modified feature vector for the B'-side only.

2.2 Fuzzy Relational Clustering

As discussed above, an object-space-based fuzzy clustering scheme such as FCM⁷⁸ could be used to cluster the data in the reduced feature space. However, to be consistent with the proposed general methodology, it was clear that converting the data in the reduced feature space into a proximity distance matrix would provide a better understanding of the inter-conformational similarities, and allow for introducing measures that are not strictly based on Euclidean distance. Moreover, such a proximity matrix could also handle any subjective similarity information which would be impossible to achieve in an

object space, while it would be easily handled through use of NERFCM or FRC. As will be demonstrated in Chapter 3, the feature vectors for **1** consist of "mixed" features (i.e., features with different physical units). Each feature vector is either a set of atom coordinates and angles between planes or a set of translational and rotational parameters. The relational data matrix obtained from such a feature vector can easily involve a non-Euclidean measure of dissimilarity. Accordingly, a relational clustering technique capable of handling non-Euclidean data to generate partitions was used.⁶⁸

FRC is a recently-developed relational clustering technique and is conceptually attractive because it works directly on the non-Euclidean data without first converting it to a Euclidean measure. The scheme is therefore less constrained than most of the other relational clustering techniques. Given a dissimilarity data matrix, $\mathbf{D} = [D_{jk}]$, $1 \leq j, k \leq n$, FRC only assumes that its elements are subject to the minimal constraints given below

$$D_{jj} = 0, \quad D_{jk} \geq 0, \quad D_{jk} = D_{kj}, \quad 1 \leq j, k \leq n. \quad (2.1)$$

The algorithm then alternates between optimizing the memberships, $\mathbf{U} = [u_{ik}]$, and a related distance matrix, $\mathbf{A} = [a_{ik}]$, $1 \leq i \leq c$, $1 \leq k \leq n$, using a successive-substitution method as described by Davé and Sen.⁶⁸ Here n is the number of data objects and c is the number of clusters fixed *a priori*. The update equations used for \mathbf{U} and \mathbf{A} are shown in equations 2.2 and 2.3,

$$u_{ik} = \frac{\left[\frac{1}{a_{ik}} \right]^{1/m-1}}{\sum_{w=1}^c \left[\frac{1}{a_{wk}} \right]^{1/m-1}}, \quad (2.2)$$

$$a_{ik} = \frac{m \sum_{j=1}^n u_{ij}^m D_{jk}}{\sum_{j=1}^n u_{ij}^m} - \frac{m \sum_{h=1}^n \sum_{j=1}^n u_{ij}^m u_{ih}^m D_{jk}}{2 \left[\sum_{j=1}^n u_{ij}^m \right]^2}. \quad (2.3)$$

The c-mean vectors, $V = [v_i]$, $1 \leq i \leq c$, are scaled n-tuples of memberships,

$$v_i = \frac{(u_{i1}^m, u_{i2}^m, \dots, u_{in}^m)^T}{\sum_{k=1}^n u_{ik}^m}. \quad (2.4)$$

The membership matrix, U , is initialized randomly. The number of clusters, c (> 1), and the fuzzifier, m (> 1), are fixed. The algorithm then iterates between equations 2.2 and 2.3, until the change in memberships in two successive iterations falls below a certain prefixed threshold, ϵ . Termination of the algorithm indicates that a local minima partition is achieved. In every iteration, the c-mean vectors are updated using equation 2.4 after all the membership values have been updated. After the algorithm converges, the membership information is defuzzified by assigning the conformation j to the cluster i if $u_{ij} > u_{kj}$ ($k \neq j$) for all $1 \leq i \leq c$, $1 \leq j \leq n$. The representative conformation is identified as the one with the highest membership value in that particular cluster. This process is carried out for a range of values for c . The clustering results are then evaluated by cluster validity analyses as described in the next section.

2.3 Fuzzy Cluster Validity Measures

Different fuzzy cluster validity indices and measures have been proposed in the literature to characterize the goodness of the partition. The simplest of these is the partition coefficient,⁷⁸ which describes the fuzziness of the partition. It is inversely proportional to the average fuzzy overlap between the clusters, and is given by

$$F = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2. \quad (2.5)$$

$F = 1$ indicates no overlap between clusters and is the case when FCM degenerates to hard c-means. On the other hand, $F \rightarrow 1/c$ is the extreme fuzzy case when all the entities are shared equally between all the clusters. Hence, the partition coefficient can take values between $1/c \leq F \leq 1$. Normalizing F as shown below can compensate for this dependence on c .

$$F' = \frac{cF - 1}{c - 1}. \quad (2.6)$$

A high value of F (and F') indicate a better partition, where clusters are compact and well separated, as opposed to a low value which indicates almost equal sharing of all entities among all the clusters.

The application of Shannon's entropy⁸⁸ to fuzzy clustering resulted in another cluster validity measure known as the partition entropy⁷⁸ and is given by

$$H = -\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij} \ln u_{ij}. \quad (2.7)$$

A good partition is characterized by a low value of H ; it can take values between $0 \leq H \leq \ln c$. Since H varies with $\ln c$, the monotonically decreasing tendency of H with c is not as severe as in the case of F and hence normalizing H has little beneficial effect.

Both the partition coefficient and entropy measure the amount of fuzziness from cluster membership information and do not consider geometric properties such as size, shape, and compactness of the clusters. Gath and Geva⁸⁹ proposed using fuzzy volume and fuzzy density of the clusters as cluster validity criteria; a good cluster is characterized by a high value of fuzzy partition density and an accompanying low value of fuzzy

hypervolume. The compactness criterion⁸² considers cluster compactness and separation as a measure of cluster validity. This criterion is also sometimes referred to as the Xie-Beni index and a modified version for use in relational clustering is given in terms of a_{ik} 's from equation 2.3 by

$$S = \frac{\sum_{i=1}^c \sum_{k=1}^n u_{ik}^2 a_{ik}}{n \left\{ \min_{1 \leq i, j \leq c, i \neq j} \|v_i - v_j\|^2 \right\}} \quad (2.8)$$

While the numerator describes the compactness of clusters in the partition, the factor in the denominator describes the separation of the clusters. A low value of S indicates a good partition.

2.4 Overview of Novel Feature Extraction Techniques

Careful feature extraction is a crucial first step in pattern recognition. In this work, two motivations guided the feature extraction process: (a) reduction of the feature space and (b) handling of the symmetry of each phenyl ring. Separate superposition-dependent and superposition-independent features were developed for use as input to the clustering algorithm. Chapter 3 will describe the two feature extraction methods more comprehensively.

The data set consisted of 728 conformations of **1**. The molecular structure of **1** is shown in Figure 2.1. If the Cartesian coordinates of each heavy (non-hydrogen) atom were taken as a feature, the dimensionality of the resulting feature space would be $N \times 3$, where N is the number of heavy atoms in the molecule. In general, reduced dimensionality of a large input data matrix is desirable for more easily-interpretable results. Moreover, some features are redundant and retaining redundant data not only

makes the feature space high dimensional and cluttered yet sparse, but also usually has periodicity that makes data classification and interpretation very difficult. Therefore, it is preferable to use the smallest possible feature vector by replacing the factor $N \times 3$ by the minimum number of features necessary to describe the molecule (i.e., the *minimal feature set*).

2.4.1 Superposition-Dependent Features

Clustering of superposition-dependent features depends on the superposition of all conformations in a common reference frame. Section 3.2 describes a novel superposition-dependent feature extraction method for obtaining a reasonably-sized feature vector for clustering. The section also investigates how different molecular superpositions can be used in combination with different minimal feature sets in order to focus the clustering on those parts of the molecule that contain important pharmacophore features. With the superposition-dependent features, the problem of phenyl ring symmetry was handled by using molecular planes as part of the feature set. Each of the two phenyl rings of **1** contains symmetry-equivalent atoms that have different atom labels. For example, the 2-position carbon in the P1-plane in Figure 2.1 is atom number C₁₂, whereas the 2'-position carbon is atom number C₁₆. Rotation of the phenyl ring of the P1-plane by 180° gives a molecular structure which is indistinguishable from the previous one, yet the labeled atoms are in different positions. Superposition of these two structures would show a perfect fit, yet calculations that are based on atom labels, such as the RMS distance between atoms, would show a large difference. As described in Section 3.2, consideration of the planes on which the phenyl rings lie provides an atom

label-independent solution to the symmetry problem. Atom label-independent description of the phenyl rings was achieved by using plane equations, which specify the planar orientations of the phenyl rings, and selected atomic coordinates (of the two fluorine atoms), which specify their exact location.

2.4.2 Superposition-Independent Features

Clustering of superposition-independent features does not require the superposition of all conformations in a common reference frame. This eliminates a time-consuming and often subjective process of superimposing a large number of conformations on a common substructure. Section 3.3 describes superposition-independent features derived from parameters obtained from the relative orientation of molecular planes. Each molecular conformation is described by a set of three translational and three rotational parameters. This set of six numbers (called the “planes parameters”) is calculated for each conformation and is used to define the proximity matrix. Thus, this feature set also fulfills the first requirement outlined above, that of feature space reduction. Further, it also meets the second requirement, that of handling the phenyl ring symmetry, in a way similar to that described above for the superposition-dependent features.

The planes parameters for the 728 conformations of **1** were calculated using the Planes program⁹⁰ developed in the Venanzi lab primarily by Deepa Pai and Rohan Woodley. The Planes program is a generalization of the 3DNA program,⁹¹ which is a versatile software package for the analysis, reconstruction, and visualization of three-dimensional nucleic acid structures. The 3DNA source code was made available to the Venanzi lab by Professor Wilma Olson.

2.4.2.1 Nucleic Acid Structure Analysis Programs. Nucleic acid structures are flexible macromolecules that can be bent, kinked, knotted, and unknotted, unwound and rewound by the proteins that interact with them. For example, of the several conformations that the DNA molecule can adopt, the most common is B-DNA, which is a right-handed double helix with a wide (major) and a narrow (minor) groove. A-DNA is another conformation of DNA and has a very deep major and a shallow minor groove. Other DNA conformations include the left-handed Z-DNA and asymmetric forms of DNA. Based on the paradigm that the function of a nucleic acid depends upon its structure, an understanding of DNA structure can aid in interpreting and predicting drug-DNA and protein-DNA interactions. Several popular approaches have been used to analyze nucleic acid structures. Comparative studies on some of these programs, such as: CEHS,^{92,93} CompDNA,^{94,95} Curves,^{96,97} FREEHELIX,⁹⁸ NGEOM,^{99,100} NUPARM,¹⁰¹ and RNA,¹⁰²⁻¹⁰⁴ have shown that the choice of reference frame rather than the mathematical calculation results in discrepancies in the parameters evaluated using different programs.^{105,106} Olson et al.¹⁰⁷ have recommended a set of standard base reference frames to describe the three-dimensional arrangements of bases and base-pairs in nucleic acid structures.

2.4.2.2 The 3DNA Program. The DNA bases adenine (A), guanine (G), thymine (T), and cytosine (C) are planar molecules. In DNA they form A-T and C-G base pairs which are separated by the phosphate backbone. The base pairs stack upon each other forming the DNA double helix and consecutive base pairs along the helical axis form a base step. The sequence-dependent structure of DNA is related to its biological function. Variation in the local structure of DNA affects DNA morphology and depends on the relative

orientation of the bases, which is affected by changes in the torsional angles of the DNA phosphate backbone.

Lu and Olson⁹¹ created the 3DNA program for the analysis, reconstruction, and visualization of three-dimensional nucleic acid structures. The program can be used with parallel and anti-parallel double helices, single-stranded nucleic acids, multi-stranded helices, and complex tertiary folding substructures common in both DNA and RNA. The program uses a coordinate reference frame for the description of nucleic acid base-pair geometry and a rigorous matrix-based algorithm to evaluate the local conformational parameters. The basic concepts, theorems, and proofs of the mathematics behind nucleic acid structure analysis are explained by Babcock et al.¹⁰⁴

A coordinate reference frame defined by the planar nucleotides (A, G, C, T, and uracil) is used to determine the base pair and base step parameters. Base pair parameters describe the position and relative orientation of one base with respect to its complementary base in a base pair. Similarly, base step parameters describe the position and orientation of consecutive bases along the helical axis of DNA. These rotational and translational parameters are rotations and displacements about the x-, y-, and z-axes of the reference frame and are illustrated in Figure 2.3. Thus, instead of defining the relative orientation of the bases and base pairs by using the multitude of backbone torsional angles, their relative orientation can be quite succinctly described by these rotational and translational parameters.

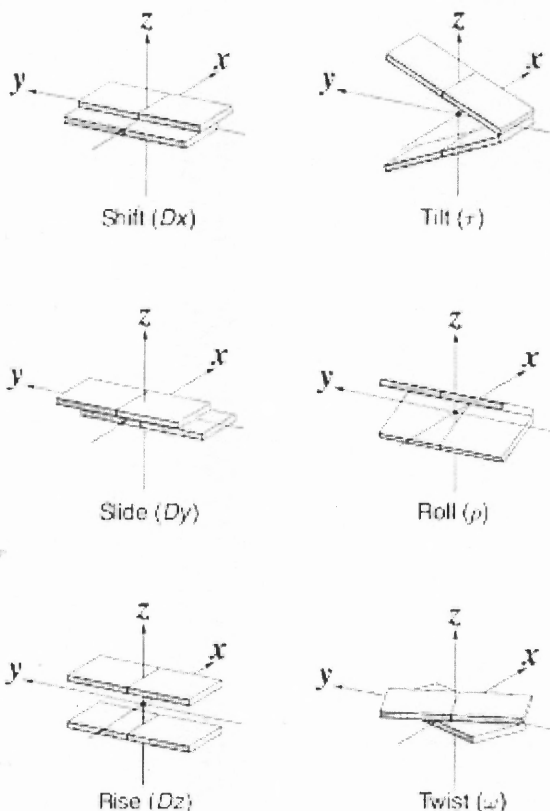


Figure 2.3 Base-step parameters used in nucleic acid structure analysis.⁹¹

2.4.2.3 The Planes Program. The Planes program⁹⁰ was developed by the Venanzi group as an extension of the 3DNA program by generalizing its scope from nucleic acids to any arbitrary molecule.

If an arbitrary molecule is viewed as a single strand of DNA, then the position and orientation of one molecular fragment with respect to another could be completely defined using six degrees of freedom: three angles and three displacements. Thus, rotational and translational parameters similar to those described for nucleic acids could be used to describe the relative orientation of any two molecular planes, such as the planes containing the piperazine and phenyl rings in the GBR 12909 analogues. Unlike DNA, which contains planes defined by either a purine or a pyrimidine ring, the planes in

an arbitrary molecule can be defined using any ring fragments. Due to the helical structure of DNA, the relative position and orientation of either a base pair or a base step is quite restricted. However, for a flexible molecule such as **1**, the relative position and orientation of any two molecular planes could encompass a much larger conformational space depending on the molecule's structural characteristics.

While the 3DNA program calculates base step parameters for only consecutive base pairs, the Planes program calculates the parameters of one plane relative to every other plane in the molecule. The Planes program uses the terminology for the base step (Shift, Slide, Rise, Tilt, Roll, and Twist) to characterize these parameters. (This is an arbitrary choice since, in the 3DNA program, the base step parameters are calculated in exactly the same way as the base pair parameters.) Section 3.3 describes the technique for extracting a superposition-independent feature vector based on planes parameters of the conformations of **1**.

CHAPTER 3

FUZZY CLUSTERING: METHODS

This chapter describes the method by which different feature vectors were generated and used as input into the FRC algorithm.

3.1 Conformational Analysis

The data set of conformations of **1** was obtained by random search of the conformational space using version 6.9 of the SYBYL molecular modeling package.¹⁰⁸ Ab initio quantum mechanical calculations at the HF/6-31G* level on GBR 12909 showed that protonation on N₄ (see Figure 2.1), the piperazinyl nitrogen distal to the bisphenyl group, was favored over protonation on N₁ (W.J. Skawinski, personal communication). Further, Dutta *et al.*⁴⁸ showed that N₄ is more essential for activity than N₁ in piperidine analogues of GBR 12909. Thus, the molecule was protonated on N₄ prior to random search. The piperazine ring was fixed as an aggregate and the eight torsional angles were randomly altered during the search. The piperazinyl side chains were both maintained in the equatorial position by checking for chirality so that the conformers were not reflected through a plane for comparison against conformers already found by the search. Symmetry was checked to reduce the number of bonds selected for rotation at each iteration. One thousand search iterations were carried out. At each step in the iteration, the eight torsional angles were randomly altered and the resulting structure was minimized using the Powell minimization method¹⁰⁹ and a convergence threshold of 0.05. The Tripos force field¹¹⁰ was used along with Gasteiger-Hückel charges¹¹¹⁻¹¹³

and a non-bonded distance cutoff of 8.0 Å. A minimized conformer was "accepted" as a new conformer if it met the following energy and RMS criteria: (1) Its RMS distance difference compared to all other conformers was at least 0.20 Å, and (2) Its energy was within 20 kcal/mol of the energy of the conformer identified to have the lowest energy at that particular step in the random search. The random search procedure ended after 1,000 steps and the "accepted" conformations were collected and used for the clustering study. The conformational analysis was performed on a SGI 500-MHz IP35 processor with 512 MB RAM on an IRIX release 6.5 operating system.

3.2 Superposition-Dependent Features

The GBR 12909 analogue, **1**, contains two pharmacophore features that are found in most dopamine reuptake inhibitors: a quaternary nitrogen (N₄ in Figure 2.1) in close proximity to an aromatic ring (here, the naphthalene ring). It also contains a bisphenyl group which has been shown to be necessary for good binding affinity by all GBR 12909 analogues.^{47,114-117} In order to aid in the feature extraction process, the molecule was conceptually divided as described below into regions containing these pharmacophore elements. Two different types of superpositions were applied to the data set of molecular conformations. Superposition 1 involved atoms in the region close to the DAT inhibitor pharmacophore elements, while Superposition 2 involved atoms in the region close to the bisphenyl group. Different *minimal features sets* were identified for each superposition. In this way the effect of clustering the conformations using a feature set defined for the molecule as a whole versus using feature sets defined for various fragments could be compared. Because the different superpositions affect the data dissimilarity matrix, **D**,

and the distance matrix, **A**, the clustering results are superposition-dependent. The superpositions and the related feature vectors are summarized in Table 3.1 and are described below.

3.2.1 Superposition 1

The data set of molecular conformations was superimposed by a rigid body superposition using atoms N₁, C₂, N₄, and C₅ in the piperazine ring. These four atoms form the central piperazine ring plane, C-plane, shown in Figure 2.1. The C-plane was fixed in the $y = 0$ plane for all structures. The molecules were translated in space so that N₁ was at the origin of the coordinate system. The molecule was divided into A- and B-sides around the C-plane as shown in Figure 2.1. The A-side and N₄ contain the DAT inhibitor pharmacophore elements. The B-side contains the bisphenyl group.

If the features were defined by the Cartesian coordinates of each heavy atom, the dimensionality of the resulting feature space would be $35 \times 3 = 105$. However, since the six heavy atoms in the ring have the same coordinates in every conformer in Superposition 1, they can be excluded from the coordinate data matrix. This results in a feature space of size $29 \times 3 = 87$, which is still quite large. Three different feature vectors were constructed using the novel feature extraction method described below in order to further reduce the size of the feature space and to compare the effects of clustering on the full molecule versus the A-side or the B-side.

3.2.1.1 Feature Extraction for A-side Clustering, Superposition 1. Examination of 1 indicates that the A-side of the molecule can be reconstructed using two sets of atom coordinates and one plane equation. The reconstruction sequence for the A-side, using

coordinates of atoms C_{23} and C_{29} and the plane equation for the N-plane, is illustrated in Figure 3.1. Starting with the known position in space of a single atom, C_{29} , it is possible to use bond length and bond angle information to construct the rest of the naphthalene fragment in the plane specified by the known plane equation of the N-plane. Once an arbitrary orientation of the naphthalene fragment is obtained, it is rotated about C_{29} within the N-plane such that C_{23}' , the arbitrary location of C_{23} , coincides with the true known coordinates of C_{23} . The resulting fragment fully specifies the A-side. The coordinates of atoms C_{23} and C_{29} and the plane equation for the N-plane form the minimal feature set for the A-side because these features contain the minimum information needed to completely specify the A-side of each conformation. The A-side feature vector used as the input to the fuzzy clustering algorithm was derived from the minimal feature set and consists of coordinates of C_{23} and C_{29} and the angle between the N-plane and C-plane, as summarized in Table 3.1. Since the C-plane is fixed in the $y = 0$ plane for all conformations, it is excluded from the definition of the minimal feature set and only the equation for the N-plane need be included. The two atoms and the two planes that define the angle between planes are labeled in Figure 1. The dimensionality of the feature space for A-side-only clustering is thus reduced to $[2 \times 3 \text{ coordinates} + 1 \text{ angle}] = 7$.

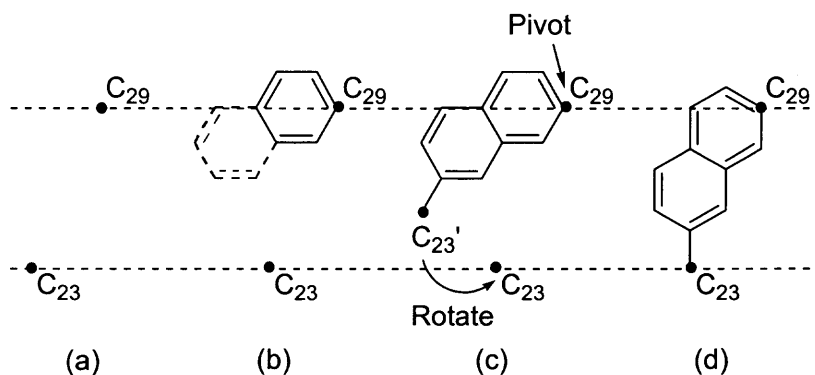


Figure 3.1 Reconstruction sequence for the A-side.

Table 3.1 Summary of Feature Vectors Used for the Two Superpositions

<i>Superposition^a</i>	<i>Clustering Side</i>	<i>Feature Vector</i>		<i>Best Result^c</i>
		<i>Atoms</i>	<i>Angle Between Planes^b</i>	
1	A	C ₂₃ , C ₂₉	N/C	c = 3, c = 6
1	B	C ₇ , C ₈ , O ₉ , C ₁₀ , F ₅₀ , F ₅₅	P1/C, P2/C	c = none
1	full molecule	C ₇ , C ₈ , O ₉ , C ₁₀ , C ₂₃ , C ₂₉ , F ₅₀ , F ₅₅	P1/C, P2/C, N/C	c = none
2	B'	C ₁₀ , F ₅₀ , F ₅₅	P1/O, P2/O	c = 9

^a 1: Atoms defining the C-plane (N₁, C₂, N₄, C₅). 2: Atoms defining the O-plane (C₇, C₈, O₉).

^b X/Y denotes angle between X-plane and Y-plane.

^c c = none: no natural groups detected.

3.2.1.2 Feature Extraction for B-side Clustering, Superposition 1. Examination of 1 indicates that the B-side of the molecule can be reconstructed using six sets of atom coordinates and two plane equations. The reconstruction sequence for the B-side begins with known coordinates for atoms F₅₀ and F₅₅ and known equations of the P1- and P2-planes. The two phenyl rings are constructed within the P1- and P2-planes using bond angle and bond length information for atoms in a phenyl ring. Once arbitrary positions for each phenyl ring are obtained, they are rotated about F₅₀ and F₅₅ within the P1- and P2-planes, respectively, such that C_{10'}, the arbitrary location of C₁₀, coincides with the true known coordinates of C₁₀. Further inclusion of the known coordinates of atoms O₉, C₈, and C₇ then completely specifies the B-side of each conformation. Thus, the coordinates of atoms C₇, C₈, O₉, C₁₀, F₅₀, and F₅₅ and the equations of P1- and P2-planes form the minimal feature set for the B-side. The feature vector for B-side clustering derived from this minimal feature set is summarized in Table 3.1 and the required atoms and planes are labeled in Figure 2.1. The dimensionality of the feature space for B-side-only clustering is thus reduced to $[6 \times 3 \text{ coordinates} + 2 \text{ angles}] = 20$.

3.2.1.3 Feature Extraction for Full-Molecule Clustering, Superposition 1.

Examination of **1** indicates that the combination of the minimal feature sets for the A- and B-sides leads to the minimal feature set for the entire molecule. Since, for all conformations of **1**, the piperazinyll nitrogen N₁ was fixed at the origin and the C-plane was fixed in the $y = 0$ plane, reconstruction of the entire molecule can be fully described using the minimal feature sets of the A- and B-sides. Thus, the molecule can be reconstructed using the known coordinates of eight atoms and three known plane equations. These eight atoms and three planes are labeled in Figure 2.1 and the feature vector derived from this minimal feature set is summarized in Table 3.1. Compared to a dimensionality of 87 based only on atom coordinates, the dimensionality of the feature space obtained here is significantly reduced to $[8 \times 3 \text{ coordinates} + 3 \text{ angles}] = 27$.

3.2.2 Superposition 2

In order to focus the clustering on the side of the molecule containing the bisphenyl group, **1** was divided into an A'- and a B'-side as shown in Figure 2.2. The molecular conformations of **1** were superimposed on the O-plane formed by atoms C₇, C₈, and O₉. For all structures, this O-plane was fixed in the $z = 0$ plane with the oxygen atom at the origin.

3.2.2.1 Feature Extraction for B'-side Clustering, Superposition 2. Examination of **1** indicates that the B'-side of the molecule can be reconstructed using three sets of atom coordinates and two plane equations. The minimal feature set for the B'-side consists of coordinates of atoms C₁₀, F₅₀, and F₅₅ and the equations of the P1- and P2-planes. The reconstruction sequence for the B'-side begins with known coordinates for atoms F₅₀ and

F_{55} and known equations of the P1- and P2-planes. The two phenyl rings are constructed within the P1- and P2-planes as above. After arbitrary positions for each phenyl ring are obtained, the rings are rotated about F_{50} and F_{55} within the P1- and P2-planes, respectively, such that C_{10}' , the arbitrary location of C_{10} , coincides with the true known coordinates of C_{10} . Since, for all conformations, the O_9 atom is fixed at the origin and the O-plane is fixed in the $z = 0$ plane, atom O_9 and the O-plane are excluded from the definition of the minimal feature set for the B'-side. The feature vector for B'-side clustering is summarized in Table 3.1 and the required atoms and planes are labeled in Figure 2.2. The dimensionality of the feature space for the B'-side is $[3 \times 3 \text{ coordinates} + 2 \text{ angles}] = 11$.

3.2.3 Determining the Angle between Two Planes

The angle between two planes is an important part of each feature set. The following protocol describes the procedure for calculating this angle. If two non-parallel planes intersect, their angle of intersection can be characterized by the acute angle θ , where $0 < \theta < \pi/2$, or the obtuse angle ϕ , where $\phi = \pi - \theta$. Whether the acute or obtuse angle is used to define the angle between two planes, each containing a fragment of molecular substructure, depends on which side of the planes the fragments lie. Figure 3.2 illustrates the two cases of the angle between planes containing molecular fragments. Since the equations of the two intersecting planes, P1 and P2, are known, the line of intersection of the two planes can be determined. Choosing any arbitrary point p on the line of intersection determines vectors \mathbf{px} and \mathbf{py} , where x and y are points on the desired side of the planes P1 and P2, respectively. The dot product of the vectors \mathbf{px} and \mathbf{py} is $\mathbf{m} =$

$\mathbf{px} \cdot \mathbf{py}$. For non-parallel planes, $\mathbf{m} \neq 0$; for positive \mathbf{m} the desired angle of intersection is the acute angle θ and for negative \mathbf{m} it is the obtuse angle ϕ . This scheme ensures a consistency in the way the feature vector is built for all conformations in the data set.

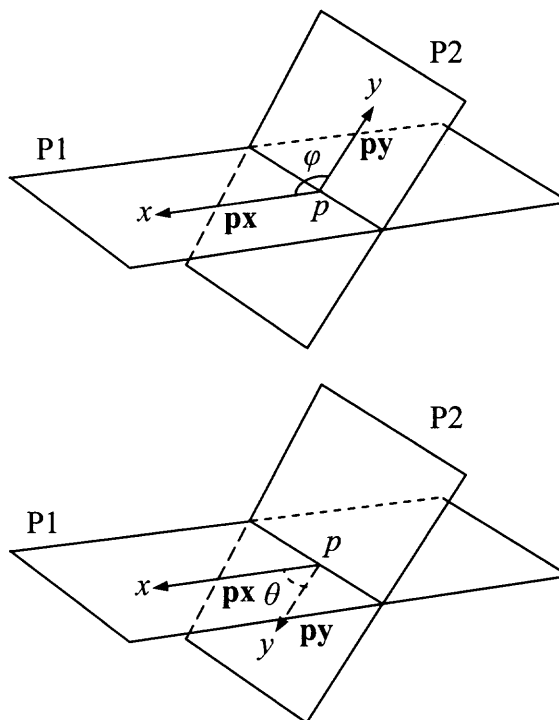


Figure 3.2 Determination of the angle between two planes.

3.2.4 Fuzzy Clustering

Each feature vector was converted into the inter-conformer dissimilarity matrix \mathbf{D} and used as the input for FRC. For each value of c ($2 \leq c \leq 12$), the clustering routine was run 10 times with a different random initialization and the median value of membership was considered for cluster assignments. Conformers were assigned to a cluster based on the largest value of their memberships over the c clusters. The representative structure for each cluster was defined as the conformation with the highest membership value in

that cluster. The two user-defined parameters used for FRC were $m = 2$ and the termination condition ϵ (change in memberships of successive iterations) = 10^{-5} . (The clustering results, however, are not very sensitive to these parameters.) The output of the clustering was used as input to the validity procedures. The FRC and cluster validity procedures were implemented by C++ programs developed by Amit Banerjee on a Sun Blade 1500 workstation running a 1-GHz 64-bit Ultrasparc III processor.

3.3 Superposition-Independent Features

The planes parameters calculated for a particular molecular conformation are based only upon the relative orientation and position of the planes defined for that conformation. In other words, the conformation can be located anywhere in space and oriented in any manner (i.e., it can be described by an arbitrary reference frame), but the planes parameters will have the same values in every location and orientation. Thus, the need to superimpose all conformations of **1** on a common substructure (such as Superpositions 1 or 2 above) is eliminated. This results in savings of time and effort and makes the overall clustering process simpler.

3.3.1 Calculation of Planes Parameters

The planes parameters that were used in this study were calculated by Deepa Pai and Liang-Yu (Lydia) Shih of the Venanzi group.

3.3.1.1 Definition of Standards. The conformations of **1** that were generated by conformational analysis as described in Section 3.1 may contain i. aromatic ring planes that are not exactly planar and ii. piperazine rings that deviate abnormally from the usual

chair configuration. For this reason, the determination of the planes parameters⁹⁰ for a conformation of **1** begins with the definition of standard (or ideal) structures for the planes contained in the conformation. As identified in Figure 2.1, four planes were defined for **1**: the two phenyl ring planes (P1- and P2-planes, defined by six atoms each), the central piperazine ring plane (C-plane, defined by four atoms), and the naphthalene plane (N-plane, defined by ten atoms). Two different standards were constructed, a piperazine standard for the C-plane and a benzene standard for the other three planes. (Besides **1**, there are at least 250 other analogues of GBR 12909, many of which have a benzene-like substituent in place of the naphthalene ring of **1**. In order to make the Planes program as general as possible and to compare the planes parameters of different GBR 12909 analogues in future studies, only the benzene-like ring on the N-plane of **1** that is proximal to the piperazine plane was used in the calculation of the planes parameters. Thus, it was not necessary to build a separate standard for the N-plane since the benzene standard was used for it instead.) As in the 3DNA program, each standard was superimposed onto its corresponding ring plane by performing a unit quaternion-based least-squares fitting procedure, first introduced by Horn,¹¹⁸ on the atoms that lie on the planes. This procedure yields the origin and the reference coordinate frame of the planes for that conformation.

3.3.1.2 Calculation of Translational and Rotational Parameters. The Planes program calculates the translational parameters (Shift, Slide, and Rise) and rotational parameters (Tilt, Roll, and Twist) for every possible combination of pairs of planes defined in the conformation. For each pair of planes, the vector product of the z-axes provides the *hinge axis* (or the line of intersection between the two planes). The dot product of the z-

axes provides the *net bending angle* between the z-axes of the two planes. Each plane is rotated about the hinge axis by half of the net bending angle and in opposite directions. Thus, the z-axes of the two planes are now aligned and the two rotated planes are parallel to each other. As in the 3DNA program, the x-, y-, and z-axes of the crucial *middle frame* are then obtained by averaging the x-, y-, and z-axes of the rotated planes. The origin of the middle frame is obtained by taking the geometrical average of the origins of the rotated planes. Parameters defined from the middle frame will be internally consistent (or “absolute”) in that they will be independent of an external frame of reference and therefore of superposition. The three translational parameters are the projections of the vector between the origins of the two rotated planes on to the x-, y-, and z-axes of the middle frame. The rotational parameter, Twist, is the angle from the y-axis of one rotated plane to that of the other rotated plane. The angle between the hinge axis and the y-axis of the middle frame is termed the *phase angle*. The remaining two rotational parameters are obtained as follows: Roll is the product of the net bending angle and the cosine of the phase angle, and Tilt is product of the net bending angle and the sine of the phase angle. The Planes program, along with explanatory figures and equations as well as verification of the code, is described in greater detail in Deepa Pai’s master’s thesis.⁹⁰

For each conformation of **1**, there are six pairs of planes: (N x P1), (N x P2), (C x P1), (C x P2), (C x N), and (P1 x P2). The planes parameters were calculated for each pair of planes and for each conformation of **1**. Thus, a total of 728 x 6 x 6 parameters were calculated.

3.3.2 Feature Vectors

Due to the flexibility of **1**, the relative orientation of the above pairs of planes can assume a much wider range of values than can consecutive bases in a base step in the much more rigid DNA structure. In order to investigate the sensitivity of the clustering results to the translational and rotational features, three types of clustering studies were carried out for each pair of planes using feature vectors defined by: (1) all six translational (T) and rotational (R) parameters, (2) only the three translational parameters, and (3) only the three rotational parameters. These studies can be identified by a shorthand notation for the proximity matrix (see Section 3.3.3) constructed in each case. Using the A-side clustering case as an example, the notation becomes: (1) $[N \times C]_{T+R}$, (2) $[N \times C]_T$, and (3) $[N \times C]_R$, respectively.

3.3.3 Proximity Matrices

As with the feature sets used in the superposition-dependent cluster analyses, the feature sets in the superposition-independent analyses also consists of “mixed” features: a set of three translational parameters (measured in Angstroms) and three rotational parameters (measured in degrees). As mentioned in Section 2.2, an object-space-based clustering technique such as FCM could be used directly on this feature set. Alternatively, the feature set could be first transformed into a proximity matrix, which relates pairwise dissimilarity between conformations, followed by the use of a relational clustering scheme to cluster conformations over this relational space. Here, as in the superposition-dependent case, a relational clustering scheme was chosen as the partitioning

methodology. Converting the data in the planes feature space to a proximity distance matrix would also provide better understanding of the interconformational similarities.

For each pair of planes, three different types of proximity matrices were constructed: (a) proximity defined by all six planes parameters, (b) proximity defined by the three translational parameters, and (c) proximity defined by the three rotational parameters. Each proximity matrix defines a distinct feature vector. The results of clustering using the three different types of proximity matrices were compared in order to evaluate the separate contributions of the translational and rotational components to the observed clustering. For proximity matrices involving mixed features, the distance between any two conformations k and j is defined as

$$D_{kj} = \left[\sum_{p=1}^3 (t_{pk} - t_{pj})^2 + s \sum_{p=1}^3 (r_{pk} - r_{pj})^2 \right]^{1/2}, \quad (3.1)$$

where t_{pk} and t_{pj} are the three translational parameters for k and j respectively, and r_{pk} and r_{pj} are the three rotational parameters for k and j respectively, with $1 \leq p \leq 3$. The scaling factor, s , is a constant chosen according to the range of the translational parameters relative to that of the rotational parameters. A judicious choice for the scaling factor, s , is the ratio of the absolute squared differences between the maximum and minimum of the translational parameters and the rotational parameters over the entire data set,

$$s = \frac{(t_{\max} - t_{\min})^2}{(r_{\max} - r_{\min})^2}. \quad (3.2)$$

Such a scaling scheme is known as *range-based scaling*.¹¹⁹ This was done prior to computing the proximities using the Euclidean distance norm. For feature sets consisting of only the translational or the rotational parameters, no scaling was required.

3.3.4 Fuzzy Clustering

The clustering routine for every proximity matrix was performed for $2 \leq c \leq 14$ and for every value of c , the routine was run 20 times with a different random initialization of memberships. The partition that minimized the FRC objective functional⁶⁸ J , which is shown in equation 3.3, was used for membership and cluster assignments.

$$J = \sum_{i=1}^c \frac{\sum_{j=1}^n \sum_{k=1}^n u_{ik}^m u_{ij}^m D_{jk}}{2 \sum_{t=1}^n u_{it}^m} . \quad (3.3)$$

Conformations were assigned to a cluster based on the largest value of their memberships over the c clusters. The representative structure for each cluster was defined as the conformation with the highest membership value in that cluster. The two user-defined parameters used for FRC were $m = 2$ and the termination condition ε (change in memberships of successive iterations) $= 10^{-5}$. (The clustering results, however, are not very sensitive to these parameters.) The output of the clustering was used as input to the validity procedures. Amit Banerjee performed all clustering calculations using the same FRC and cluster validity programs that were used in the superposition-dependent studies.

CHAPTER 4

SUPERPOSITION-DEPENDENT CLUSTERING RESULTS

4.1 Conformational Analysis

A total of 728 unique conformations were found in the range 11.2 - 27.9 kcal/mol. Every conformation was found at least once. The measure of completeness of the random search is given by¹²⁰

$$\text{probability of finding all conformers} = 1 - (0.5)^n,$$

where n is the number of times each conformation was found. This suggests that there was 50% chance that all possible conformations were found during the random search conformational analysis.

4.2 Clustering

The optimal number of clusters found for each feature vector and superposition is given in the last column of Table 3.1. The optimal number of clusters was determined as that value of c for which the cluster validity indices have the following relationship: partition coefficient (F) - high, normalized partition coefficient (F') - high, partition entropy (H) - low, and compactness index (S) - low.

4.2.1 Full-Molecule Clustering, Superposition 1

The flexibility of **1** ensured that a large conformational space was covered by the random search protocol, as can be seen by superposition of all 728 conformations in Superposition 1, Figure 4.1. Clustering of the conformations using the whole-molecule

feature vector outlined in Table 3.1 indicated the absence of natural groupings according to the behavior of the cluster validity indices (not shown). This is perhaps not surprising, given the wide range of positions occupied by the atoms of the B-side in Superposition 1. Figure 4.1 shows more clearly-defined groups on the A-side of the superimposed conformations due to more limited positions available to the naphthalene ring. Since the piperazine and naphthalene rings contain the pharmacophore features that are found in most DAT inhibitors, the next clustering study used a feature vector defined only in terms of the A-side in order to focus on these pharmacophore features.

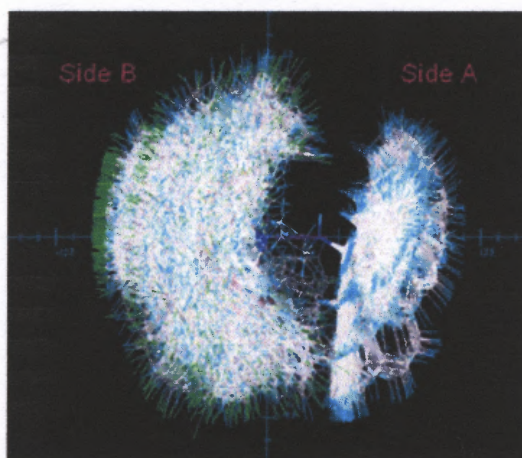


Figure 4.1 Side view of the 728 conformations of **1**, Superposition 1.

4.2.2 A-Side Clustering, Superposition 1

The cluster validity results for the A-side partitions for Superposition 1 are shown in Figure 4.2. All four validity indices attain their first inflexion point and their respective optima at $c = 3$ suggesting a good three-cluster partition. The compactness index, S , indicates good partitioning for $c = 6$ through $c = 9$ with the other three indices either monotonically increasing or decreasing over that range. This suggests a good second level partitioning at the lower bound, $c = 6$.

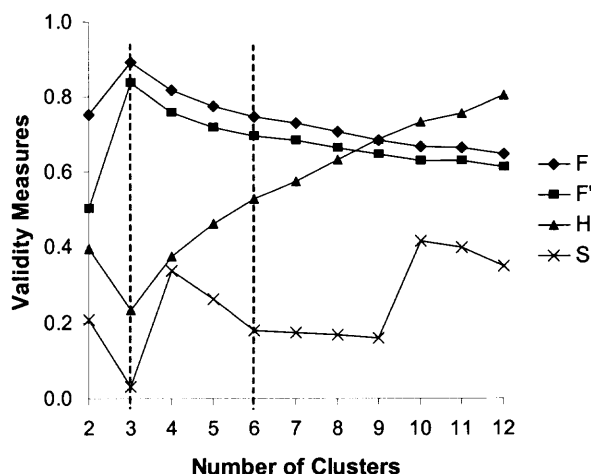


Figure 4.2 Cluster validity plots for partitions on the A-side.

It is not possible to visualize the clusters in a feature vector space that consists of only an angle (between the C- and N-planes) and two points (C_{23} and C_{29}). A qualitative way of visualizing the clusters is to show their relationship to easily-identifiable physical features of the molecule, such as the A1 and A2 torsional angles (see Figure 4.3(a)). These angles are important because they determine the relative orientation of the C- and N-planes. However, it should be noted that because the feature vector for A-side clustering was not defined specifically in terms of A1 and A2, representation of the clusters in (A1, A2) space is not equivalent to representation of the clusters in feature vector space. As a result, clusters which are well-separated in feature vector space may appear to overlap somewhat in (A1, A2) space. This is similar to the effect seen if clusters that are well-separated along the x-, y-, and z-axes in three-dimensional Cartesian space are projected onto a plane (such as defined by the x- and y-, x- and z-, or y- and z-axes). In other words, that the clusters appear to mix and have poor separation could be an artifact of viewing the results in (A1, A2) space.

Torsional angles A1 and A2 have physical significance because they determine how the conformations form natural groups on the A-side of the molecule. The 728 conformations of **1** were identified by the Random Search procedure to be minima on the conformational potential energy surface of the molecule. This means that each conformation has values of A1 and A2 (as well as B1 - B6) that were determined by minimizing the conformational energy for rotation around these angles. The A1 and A2 torsional angles contain $\text{N}(\text{sp}^3)\text{-C}(\text{sp}^3)$ and $\text{C}(\text{sp}^3)\text{-C}(\text{sp}^2)$ bonds, respectively. Therefore, the A1 and A2 torsional angles output by the Random Search technique should be close to the values of the torsional angles found in staggered conformations of compounds such as aminomethane and methylbenzene, which can be considered to be models for the A1 and A2 torsional angle rotational barriers, respectively.

In order to visualize these relationships, the 728 conformations were plotted in (A1, A2) torsional angle space with each conformation color-coded by the color of the cluster to which it was assigned by the FRC procedure. Figure 4.3(b) is a scatter plot of the 728 conformations for the $c = 3$ cluster level. The figure shows that the three clusters (identified by green, blue, and red points) are located about 120° apart on the A1 axis (at approximately $\text{A1} = 60^\circ, 180^\circ, \text{ and } 300^\circ$, respectively). The location of these clusters corresponds to rotational minima around the $\text{N}_4(\text{sp}^3)\text{-C}_{23}(\text{sp}^3)$ bond in A1, and is typical of the rotational minima in aminomethane. The three clusters appear to be well-defined by differences in the A1 values of the conformations. This shows that the fuzzy clustering technique has identified the natural groups based on the minimum in the conformational energy for the A1 torsional angle.

Figure 4.3(b) also shows that most of the conformations are found clustered along the A2 axis at approximately $A2 = 90^\circ$ and 270° . Several conformations are also found "spread out" along the A2 axis at intermediate values of the angle. The location of these clusters corresponds to rotational minima around the $C_{23}(sp^3)-C(sp^2)$ bond in A2. This complex pattern of rotational minima is due to the effect of large substituent groups on the carbons in the $C_{23}(sp^3)-C(sp^2)$ bond in A2.

Figure 4.3(c) shows that the six clusters identified for $c = 6$ (shown in green, yellow, blue, cyan, magenta, and red) are directly related to the three clusters of Figure 4.3(b). The yellow and green clusters of Figure 4.3(c) contain the same points as the green cluster of Figure 4.3(b). A similar relationship holds for the blue and cyan/blue and magenta and red/red clusters of Figure 4.3(c)/4.3(b), respectively. Figure 4.3(c) shows that the clusters are well-separated by their A1 values and fairly well-separated by their A2 values. There is, however, some apparent mixing between the groups of colored points based on the value of A2. Some green points are found in the yellow cluster centered around $(A1 = 60^\circ, A2 = 260^\circ)$. Similarly some yellow points are found in the green cluster located near $(A1 = 60^\circ, A2 = 60^\circ)$. Similar mixing is seen in the blue and cyan clusters, as well as the magenta and red clusters. Since the cluster validity indices in Figure 4.2 indicate good cluster separation, the apparent mixing is probably due to viewing the scatter plot in $(A1, A2)$ space rather than feature vector space.

The FRC results suggest that grouping the conformations by the value of the A1 angle provides a clear separation into three clusters, if only the A-side of the molecule is considered. However, separation of the conformations into a larger number of clusters appears to be more complex than simply basing the grouping on the value of A2. The

results of the cluster validity tests show that while separation at the $c = 3$ level is obvious, separation at the $c = 6$ level is less so. This indicates the complexity involved in separating conformations of a very flexible molecule into many clusters. However, Figure 4.3(c) shows that at cluster level $c = 6$ the FRC technique has identified natural groups related to minima in the conformational energy for combinations of the A1 and A2 torsional angles.

Figure 4.4 shows the molecular conformations that correspond to the scatter plots in Figure 4.3. The view depicted is a 90° clockwise rotation about the central plane of Figure 4.1 such that the A-side naphthalene rings are presented frontally (the piperazine ring and B-side are not shown). The molecules are oriented in a somewhat off-center view along the $C(sp^3) - C(sp^3)$ bond of the A1 torsional angle. Figure 4.4(a) shows three clusters of 229, 270, and 229 conformations each, which correspond to the red, blue, and green clusters, respectively, of Figure 4.3(b). Figure 4.4(b) shows the representative structure for each cluster identified as random search conformations #62 (red; $A1 = 305^\circ$, $A2 = 96^\circ$), #251 (blue; $A1 = 174^\circ$, $A2 = 268^\circ$), and #96 (green; $A1 = 69^\circ$, $A2 = 101^\circ$). Thus, the representative structures have A1 values that differ by about 120° . The clusters at the $c = 3$ level appear to separate well in terms of the A1 torsional angle.

Figure 4.4(c) shows six clusters of 77, 153, 128, 142, 82, and 146 conformations each, which correspond to the magenta, red, blue, cyan, yellow, and green clusters, respectively, of Figure 4.3(c). The figure illustrates the apparent cluster mixing noted in Figure 4.3(c). As in Figures 4.3(b) and 4.3(c), comparison of the conformations in the clusters in Figures 4.4(a) and 4.4(c) shows that, in general, the $c = 6$ magenta and red

clusters form the $c = 3$ red cluster, the $c = 6$ blue and cyan clusters form the $c = 3$ blue cluster, and the $c = 6$ yellow and green clusters form the $c = 3$ green cluster.

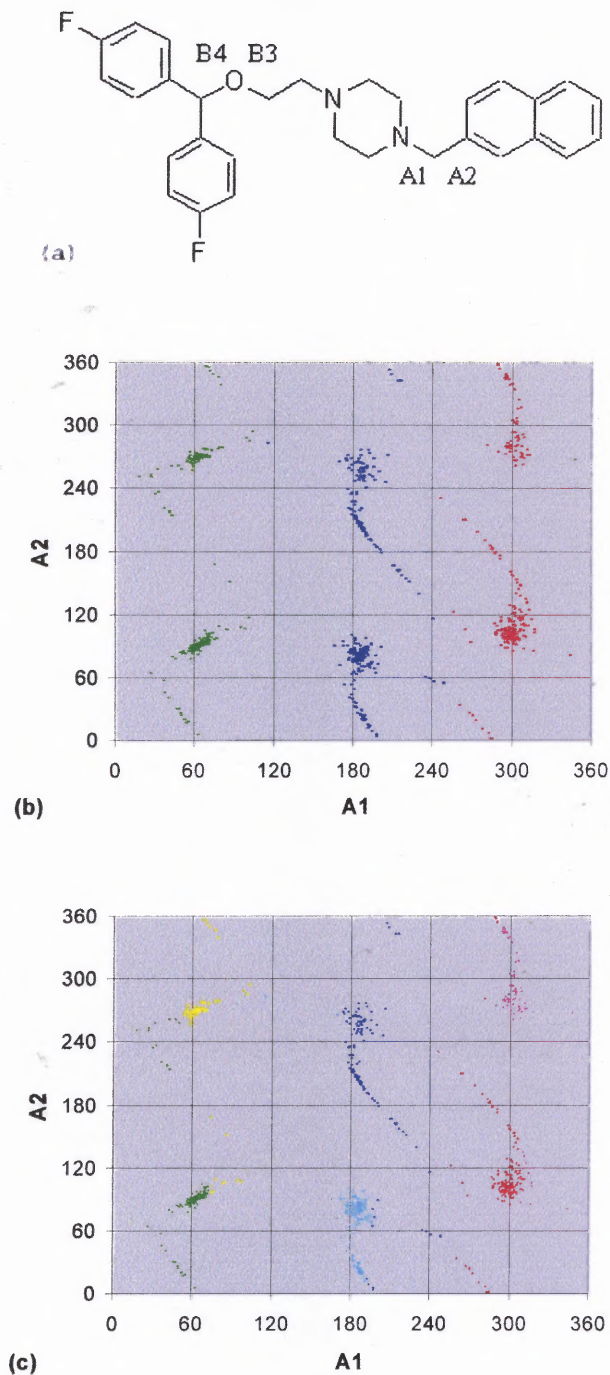


Figure 4.3 Torsional angles definitions and conformations of **1** in (A1, A2) space.

Figure 4.4(d) shows the representative structure for each cluster identified as random search conformations #154 (magenta; $A1 = 301^\circ$, $A2 = 293^\circ$), #531 (red; $A1 = 298^\circ$, $A2 = 128^\circ$), #428 (blue; $A1 = 192^\circ$, $A2 = 191^\circ$), #248 (cyan; $A1 = 183^\circ$, $A2 = 68^\circ$), #232 (yellow; $A1 = 62^\circ$, $A2 = 265^\circ$), and #177 (green; $A1 = 62^\circ$, $A2 = 90^\circ$), respectively. Note that the $A1$ values of the representative structures of the color-related (yellow/green, red/magenta, blue/cyan) clusters at the $c = 6$ cluster level have very similar values to those in the color-related representative structure at the $c = 3$ level. For example, the $A1$ value of the representative structure for the $c = 3$ red cluster (305°) is close to those of the representative structures for the $c = 6$ magenta (301°) and red (298°) clusters. Figure 4.4(d) illustrates that the representative structures at the $c = 6$ and $c = 3$ levels have similar $A1$ values that differ by about 120° . Figure 4.4(d) also shows that the $A2$ values of the color-related representative structures differ by various amounts: 165° (magenta/red), 123° (blue/cyan), and 175° (yellow/green). The fact that the $A2$ values of the blue and cyan representative structures differ by much less than 180° could be due to incomplete searching of the conformational space. As noted above, there is only a 50% chance that all possible conformations were found during the random search conformational analysis.

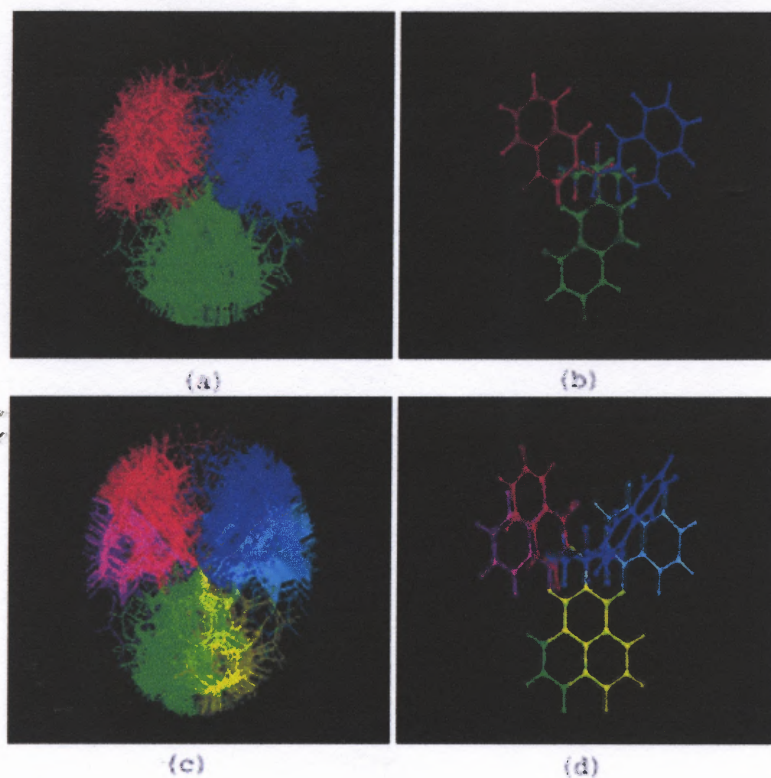


Figure 4.4 Results for the A-side clustering at $c = 3$ and $c = 6$, Superposition 1.

4.2.3 B-Side Clustering, Superposition 1

Clustering using the B-side feature vector indicated the absence of natural groups. This is consistent with the fact that the B-side of **1** is much more flexible than the A-side due to the presence of six rotatable bonds on the B-side versus two on the A-side. The B-side of **1** can access a much wider range of conformational space than the A-side, as shown in Figure 4.1. None of the validity indices provides a reason to believe that there is an underlying structure on the B-side (Figure 4.5). The compactness index is not plotted because the results were not considered to be sufficiently consistent, indicating a lack of substructure. The normalized partition coefficient, F' , takes values very close to zero ($cF \rightarrow 1$) and hence the results at all levels of clustering are too fuzzy to be of any significance.

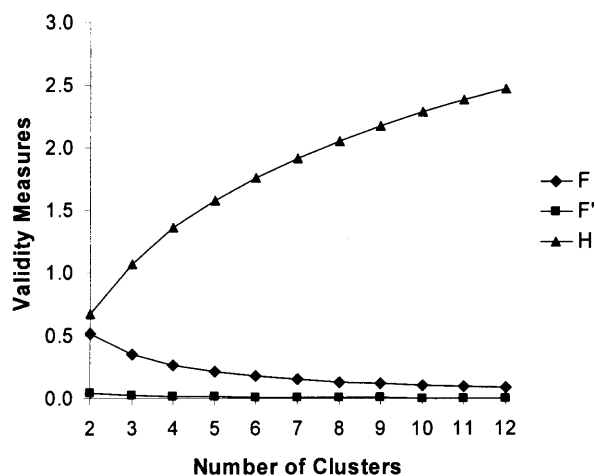


Figure 4.5 Cluster validity plots for partitions on the B-side.

4.2.4 B'-Side Clustering, Superposition 2

The cluster validity indices plotted in Figure 4.6 suggest nine optimal clusters for the B'-side. The compactness index, S, has its lowest value for $c = 9$. The other indices support this partition, indicating well-separated and compact clusters.

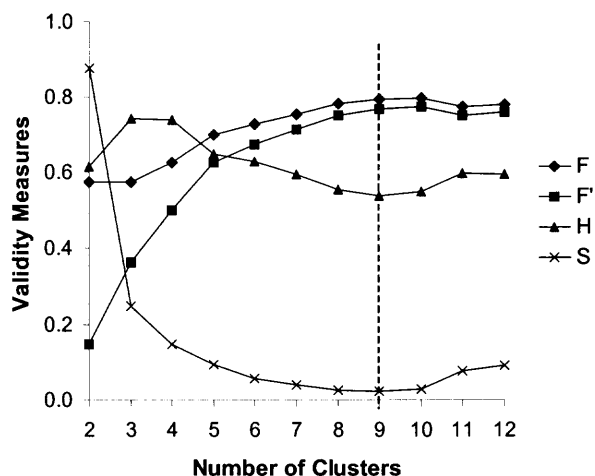


Figure 4.6 Cluster validity plots for partitions on the B'-side.

As above, one way to visualize the clusters is by scatter plots of the conformations in torsional angle space. Comparison of Figure 2.2 with Figure 2.1 shows

that the B'-side of **1** contains only three rotatable bonds instead of the six bonds on the B-side. Comparison of Figures 4.3(a) and 2.2 shows that torsional angles B3 and B4 in part control the orientation of the bisphenyl group with respect to the O-plane and the rest of the molecule. Since Superposition 2 is based on the O-plane, it allows the grouping on the B'-side of the superimposed conformations to be observed. One way of visualizing the clusters is to plot the conformational minima in (B3, B4) space. However, it should be emphasized here, as above, that the B'-side feature vector was not defined directly in terms of B3 and B4 but rather in terms of the related angles between the O- and P1- and P2-planes. Therefore, plotting the data in (B3, B4) space, although an obvious way to visualize the physical data, is not exactly equivalent to plotting the results in the B'-side feature vector space.

Figure 4.7 shows the nine distinct clusters on the scatter plot of the conformations in (B3, B4) torsional angle space. This corresponds to the nine rotational minima that result from a combination of the three rotational minima for staggered conformations around the C(sp³)-O(sp³) bond in B4 with the three rotational minima for staggered conformations around the O(sp³)-C(sp³) bond in B3. As shown in the figure, the clusters are located at about 120° intervals along the B3 and B4 axes. Apparent mixing of some clusters is due to representation of the clusters in (B3, B4) space rather than feature vector space.

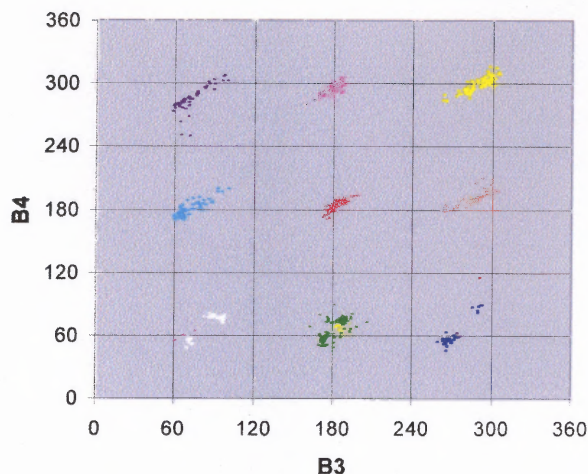


Figure 4.7 Conformations of **1** in (B3, B4) space.

Figure 4.8 shows these nine B'-side clusters as well as the representative conformations from each cluster. Each cluster is formed by the bisphenyl group on the B'-side (the A'-side is not shown). Each phenyl ring of the bisphenyl group in a cluster occupies two different regions in space. For example, three clusters (blue, green, and white) have both of their phenyl rings located out on the edge and six clusters (red, magenta, purple, cyan, orange, and yellow) have one phenyl ring located out on the edge and the other located in the center, coming out of the plane of the figure. Since no two colors appear in the same region, the clusters are distinct. For example, while one phenyl ring of both the orange and the yellow clusters seems to be overlapping in the center, the other phenyl ring of the orange cluster lies on the bottom left and that of the yellow cluster lies on the top left. Thus, the orange and yellow clusters are distinct and do not overlap.

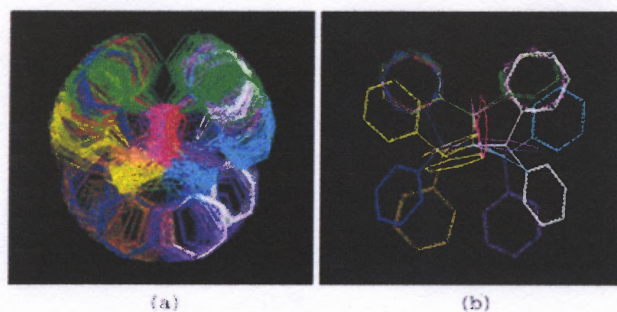


Figure 4.8 Clustering results for the B'-side at $c = 9$.

4.3 Identification of Full-Molecule Representative Structures

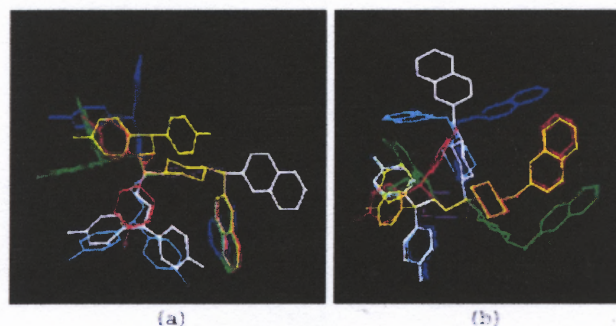
Since the full-molecule clustering suggested the absence of natural groups, the superposition-based and region-specific clustering results obtained above (for the A- and B'-sides) were used to identify representative structures for **1**. The A1 and A2 values of the six cluster representatives from the A-side clustering were combined with the B3 and B4 values of the nine representatives from the B'-side clustering to construct 54 “ideal” combinations of the four torsional angles. Then a search through the dataset of 728 conformations using a tolerance of $\pm 2.5^\circ$ on each torsion angle produced six matches as listed in Table 4.1. The table ranks the structures by their energy relative to that of the GEM conformation. It is interesting to note that the GEM conformation is not one of the representative structures. Since ligands have been shown to bind to proteins in conformations that have energies over 10 kcal/mol above the GEM,⁵⁹ the representative structures appear to be reasonable in terms of energy.

Table 4.1 Torsional Angles^a and Relative Energies^b of Full-Molecule Representatives

<i>Color</i>	<i>Conf. #^c</i>	<i>A1</i>	<i>A2</i>	<i>B1</i>	<i>B2</i>	<i>B3</i>	<i>B4</i>	<i>B5</i>	<i>B6</i>	<i>Relative Energy</i>
White	723	303	290	266	62	181	296	319	26	3
Blue	232	62	265	71	62	177	294	137	199	6
Yellow	638	63	92	65	295	180	182	163	49	6
Green	391	61	89	75	196	292	297	116	175	7
Red	394	60	90	61	295	266	57	266	350	8
Cyan	710	61	89	221	42	181	295	128	10	10

a. In degrees.
b. In kcal/mol and relative to the global energy minimum.
c. Number of the conformation in the random search output.

These six conformations were aligned using Superposition 1 in Figure 4.9(a) and using Superposition 2 in Figure 4.9(b). They form the final set of representative structures of **1** that will be used as templates for future CoMFA studies of GBR 12909 analogues. The conformations appear to be representative of the regions of space occupied by **1**.

**Figure 4.9** Full-molecule representative structures.

4.4 Generalization of the Feature Extraction Method

Although the feature extraction method was developed keeping the structure of **1** in mind, the method can be generalized to other molecules. Intelligent selection of features combined with a general adherence to these guidelines can produce a useful feature set,

consisting of atom coordinates and planes, which can be used to reconstruct the molecule. The associated feature vector, consisting of atom coordinates and angles between pairs of planes, can then be used as input for fuzzy clustering of molecular conformations.

The process begins with the identification of planes that can be used in reconstructing the molecule. Symmetric planar rings (such as phenyl rings) could be selected to avoid symmetry-related problems, as described in Section 3.2. Planes that contain important structural moieties, such as pharmacophore elements, would be chemically sensible. In order to reconstruct the molecule, one atom per selected plane needs to be included in the feature set (for example, atom C₂₉ on the N-plane in **1**, see Figure 3.1). Thus, selection of large planar rings, like naphthalene, enables significant reduction in feature space. Selection of planes at the extremities of the molecule could also be useful in the reconstruction process. The clustering technique being superposition-dependent, a substructure should be identified that will be used to superimpose all conformations. Since this substructure will be common to all conformations, atoms that lie on it need not be included in the feature set. The guidelines for constructing a feature vector for a general molecule are summarized below:

1. Identify planes.
2. Identify all (heavy) atoms that do not lie on the planes.
3. Identify a superimposable substructure.
4. Remove from the set identified in step 2 any atom that is part of the substructure identified in step 3.
5. Include in the set one suitable atom on each plane (not necessary if plane is part of the superimposable substructure).

6. Select angles between pairs of planes such that the overall structure of the molecule can be “captured”.

Applying the guidelines to **1**:

1. Planes C, N, P1, and P2.
2. Atom set: Atoms C7, C8, O9, C10, and C23.
3. Superimposable substructure: Piperazine.
4. Atom set: Atoms C7, C8, O9, C10, and C23 (remove atoms: none).
5. Atom set: Atoms C7, C8, O9, C10, C23, C29, F50, and F55 (include atoms F50 (P1-plane), F55 (P2-plane), and C29 (N-plane)).
6. Angles between pairs of planes: N/C, P1/C, and P2/C.

Applying the guidelines to cocaine (see Figure 4.10):

1. Planes X, Y, and Z.
2. Atom set: Atoms 2, 3, 4, 6, 7, 8, 9, 10, and 11.
3. Superimposable substructure: Tropane ring.
4. Atom set: Atoms 2, 3, 4, 6, and 7 (remove atoms 8, 9, 10, and 11).
5. Atom set: Atoms 1, 2, 3, 4, 5, 6, and 7 (include atoms 1 (plane X) and 5 (plane Z)).
6. Angles between pairs of planes: X/Y and Z/Y.

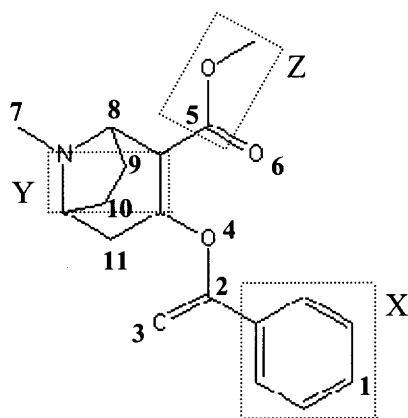


Figure 4.10 Identification of reduced feature set for cocaine.

This example shows that the novel feature extraction techniques can be easily generalized to other molecules.

4.5 Comparison to Hierarchical Clustering Using XCluster

In a separate publication,²⁹ Gilbert and Venanzi present the results of hierarchical clustering of the same data set of 728 conformations of **1** using XCluster.⁷⁰ In that study, several feature vector and superposition options were explored and are summarized in Table 4.2. They range from full molecule clustering (options *a* and *b*) to A-side clustering (options *c*, *d*, and *e*) to B-side clustering (options *f* - *j*). Since XCluster allows for clustering using Cartesian coordinates of atoms *or* torsional angles, whereas the novel feature extraction procedure of the present work uses atomic coordinates and angles between planes, the feature vectors are not exactly the same. However, both techniques allow the user to focus on features related to full molecule, A-side or B-side clustering with different superpositions, so the results of the studies are comparable. For example, the "center ring" superposition of Table 4.2 is the same Superposition 1 of Table 3.1. The "oxygen and neighboring carbons" superposition of Table 4.2 is slightly different

than Superposition 2 of Table 3.1 because it involves C₁₀, O₉, and C₈, whereas Superposition 2 uses O₉, C₈, and C₇ (see Figure 2.1). The "all heavy atom" feature vector in Table 4.2 (options *a* and *b*) is related to the "full molecule" feature vector in Table 3.1. Similarly the A-side feature vectors in options *c*, *d*, and *e* of Table 4.2 are closely related to the A-side feature vector in Table 3.1. Of the B-side feature vectors in Table 4.2, option *j* is the closest to the B-side feature vector of Table 3.1.

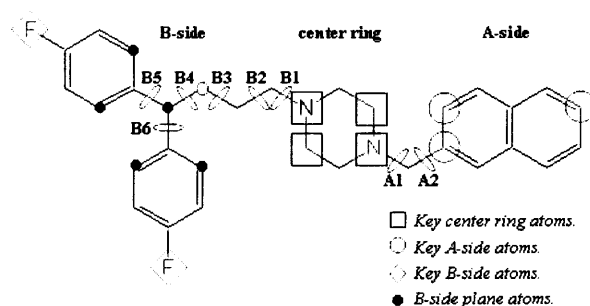


Figure 4.11 Features selected in XCluster studies.²⁹

The last column of Table 4.2 summarizes the appearance of the XCluster distance maps for the various feature vector/superposition options and can be compared to the last column of Table 3.1. For the full-molecule clustering, the FRC and XCluster techniques give the same results because no obvious large clusters are detected. Both methods also agree for A-side clustering. XCluster options *c* - *d* all give six large clusters, which is the same as the FRC result. Display of the XCluster A-side results (not shown) shows the same type of separation noted in Figures 4.3(c) and 4.4(c). This is because the cluster memberships are a result of natural groupings determined by the (A1, A2) values that correspond to minima in the conformational potential energy surface for rotation around the N(sp³)-C(sp³) and C(sp³)-C(sp²) bonds in A1 and A2.

Table 4.2 Summary of XCluster Studies

<i>Option</i>	<i>Feature Vector^a</i>	<i>Superposition^b</i>	<i>Distance Map Appearance</i>
A	All heavy atoms	□ Center ring ^c	Many small clusters
B	All heavy atoms	All heavy atoms	Many small clusters
C	□ Center ring ○ A-side key atoms	□ Center ring ^c	Five or six large clusters
D	□ Center ring ○ A-side key atoms	□ Center ring ○ A-side key atoms	Six or seven large clusters
E	A1 and A2	Five heavy atoms defining A1 and A2 ^d	Six large clusters
F	□ Center ring ◇ B-side key atoms	□ Center ring ^c	Many small clusters
G	□ Center ring ◇ B-side key atoms	□ Center ring ◇ B-side key atoms	Two large, many small clusters
H	□ Center ring ◇ B-side key atoms	Oxygen and neighboring carbons ^c	Two large clusters and seven small clusters
I	□ Center ring ◇ B-side key atoms • B-side plane atoms	Oxygen and neighboring carbons ^c	Fifteen small clusters
j	B1 through B6	Nine heavy atoms defining B1 through B6 ^d	Many small clusters

^a Feature vector: The atoms or angles used to calculate the intermolecular distances. Key atoms (identified by symbols noted) and torsional angles are shown in Figure 4.11. RMSD calculations for the distance matrix were carried out on the feature set atoms in the corresponding alignment.

^b Superposition performed by XCluster unless otherwise noted.

^c Superposition performed in SYBYL.

^d Superposition of conformations for the torsional angle studies is not necessary to calculate RMSD values, but was carried out to properly visualize clusters.

XCluster options *f*, *g*, and *j* all result in many small clusters and agree with the FRC result shown in Figure 4.5 that there is no underlying structure on the B-side using these superpositions. Although the "oxygen and neighboring carbons" superposition of option *h* is slightly different than Superposition 2 of the B'-side clustering in Table 3.1, both these FRC and XCluster studies result in detection of nine clusters. This is because both feature vectors focus on the region near (B3, B4). As discussed in Section 4.2, this is because cluster memberships are the result of natural groupings due to the combination of the three conformational minima for rotation around the C(sp³)-O(sp³) bond in B4 with the three conformational minima for rotation around the O(sp³)-C(sp³) bond in B3 (see Figure 4.7). In summary, the novel feature extraction technique presented here combined with the fuzzy relational clustering methodology gives the same results as the hierarchical clustering approach implemented in XCluster when similar feature vectors and superposition options are used.

4.6 Use of Torsional Angles as Feature Vectors?

Since Figure 4.3 shows that the conformations appear to cluster well in (A1, A2) space, the question may arise as to why torsional angles were not used in the FRC feature vector. While this might work well in cases like A-side clustering, Superposition 1, where only two torsional angles can completely specify the conformations allowed for the A-side of the molecule, its application is perhaps less useful in describing regions of the molecule that have many rotatable bonds. The Venanzi group's alternative analysis of the conformations of **1** by singular value decomposition (SVD) of all eight torsional angles has been described elsewhere.¹²¹ The novel feature extraction technique was

developed for the express purpose of going beyond torsional angle analysis to a more global approach that employs features of the whole molecule or molecular fragment as a basis for classification. The purpose of the novel feature extraction process was to identify a *minimal feature set* that could be used to classify the conformations into groups. The usefulness of that approach has been demonstrated in A-side and B-side clustering which focused on the important pharmacophore features of 1.

4.7 Different Superpositions

Whereas an analysis such as one based on torsional angles would be independent of superposition of the conformations, the technique presented here is superposition-dependent. Different superposition might necessitate a redefinition of the minimal feature set from which the input feature vector is derived. However, even if the input feature vector is found to be the same for a new superposition, the new positions of the conformations would lead to a change in the distance matrix. This necessitates careful application of this technique when selecting optimal alignment rules for conformations. On the other hand, using different superpositions enables one to focus on specific pharmacophore regions of the molecule and provides clustering results that may not be uncovered without such data reduction. The next chapter contains the results from cluster analyses using superposition-independent feature vectors introduced in Section 3.3.

CHAPTER 5

SUPERPOSITION-INDEPENDENT CLUSTERING RESULTS

5.1 Clustering

The results for the superposition-independent clustering studies are summarized in Table 5.1 using the notation defined in Section 3.3.2. Since the behavior of the two phenyl rings in **1** has been found to be coupled,¹²² clustering studies involving the P1-plane do not give significantly different information from those involving the P2-plane. For this reason, only the P2-plane results are presented here. Additional information on the P1-plane results is given in Appendix A. Visualization of the clustering results is possible in those cases where the features consist of *either* translational *or* rotational planes parameters. In such cases, the clustering is shown in two-dimensional or three-dimensional translational (Slide, Shift, Rise) or rotational (Roll, Tilt, Twist) space, using the planes parameters as coordinate axes. Note that visualization in six-dimensional translational plus rotational space is not possible.

Table 5.1 Superposition-Independent Clustering Results

<i>Type of Clustering</i>	<i>Proximity Matrix</i>	<i>Optimal Number of Clusters, c</i>
Full Molecule	$[N \times P2]_{T+R}$	5
Full Molecule	$[N \times P2]_T$	5
Full Molecule	$[N \times P2]_R$	13
B-side	$[C \times P2]_{T+R}$	3
B-side	$[C \times P2]_T$	4
B-side	$[C \times P2]_R$	8
A-side	$[N \times C]_{T+R}$	9
A-side	$[N \times C]_T$	7
A-side	$[N \times C]_R$	8

5.1.1 Full-Molecule Clustering

In the case of the superposition-dependent analysis (Section 4.2.1) where the full molecule was clustered using a proximity matrix derived from eight atom locations and three sets of angles between planes, natural groups were not located. In contrast, partitions produced for the $[N \times P2]_{T+R}$ proximity matrix indicate the presence of five clusters. This is confirmed by the Xie-Beni index S , though the other validity measures are inconclusive (see Figure 5.1): S takes its lowest value over the range $2 \leq c \leq 14$ at $c = 5$, H is seen to be monotonically increasing, and F and F' are monotonically decreasing over the entire range.

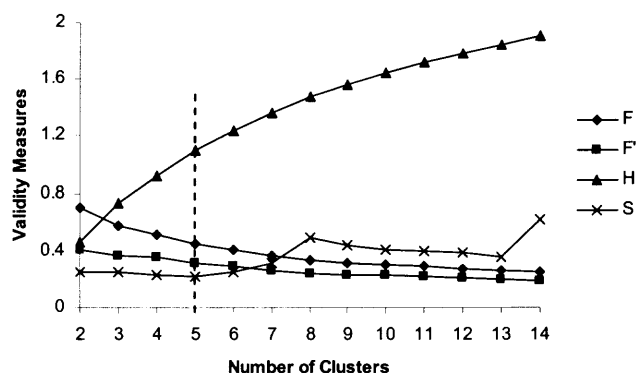


Figure 5.1 Cluster validity plots for the $[N \times P2]_{T+R}$ proximity matrix.

This separation into five clusters is best visualized in the three-dimensional translational space (Slide, Shift, Rise) and in the two-dimensional (Slide, Rise) plane as shown in Figures 5.2 and 5.3 respectively. In these and all subsequent plots each point represents a conformation, the conformations are color-coded by cluster, the translational parameters are given in Angstroms (\AA), and the rotational parameters in degrees.

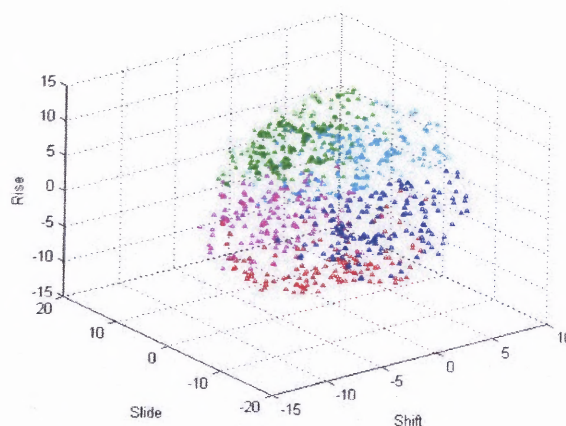


Figure 5.2 (Slide, Shift, Rise) space for $[NXP2]_{T+R}$, $c = 5$.

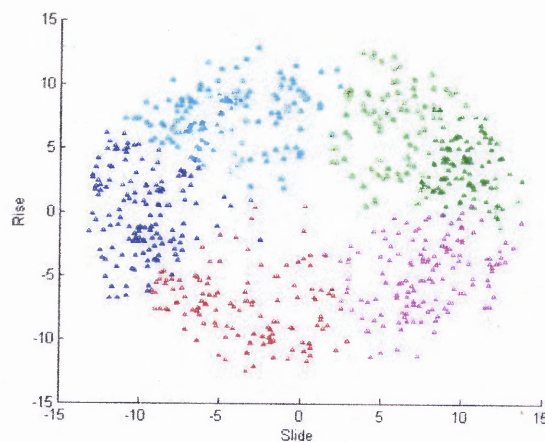


Figure 5.3 (Slide, Rise) space for $[NXP2]_{T+R}$, $c = 5$.

The partition produced for the pure translational proximity matrix $[NXP2]_T$ is very similar to the one produced for the $[NXP2]_{T+R}$ clustering and also identifies five clusters (Figure 5.4). Figure 5.5 shows the plot of conformations in (Slide, Shift, Rise) space for $[NXP2]_T$ that is similar to the plot in Figure 5.2. This seems to indicate that the translational parameters may be the chief determinant of separation in full-molecule clustering. This is supported by two additional observations. First, the plot (Figure 5.6)

of conformations in (Roll, Tilt, Twist) rotational space for $[N \times P2]_{T+R}$ shows no separation of conformations into clusters. Also, the validity plot (Figure 5.7) for clustering over the rotational proximity matrix, $[N \times P2]_R$, identifies 13 clusters, which does not agree with the $[N \times P2]_{T+R}$ and $[N \times P2]_T$ results, shown in Table 5.1, that suggest that the optimal number of clusters is five.

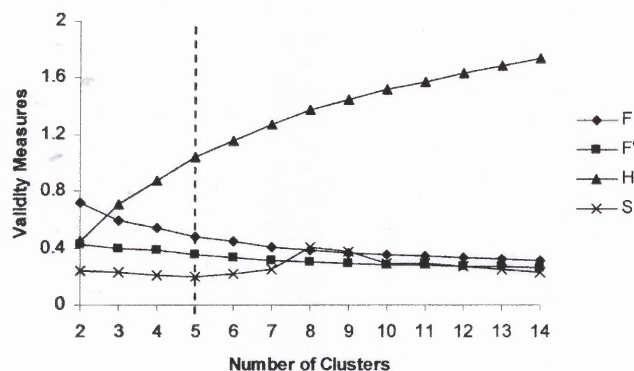


Figure 5.4 Cluster validity plots for the $[N \times P2]_T$ proximity matrix.

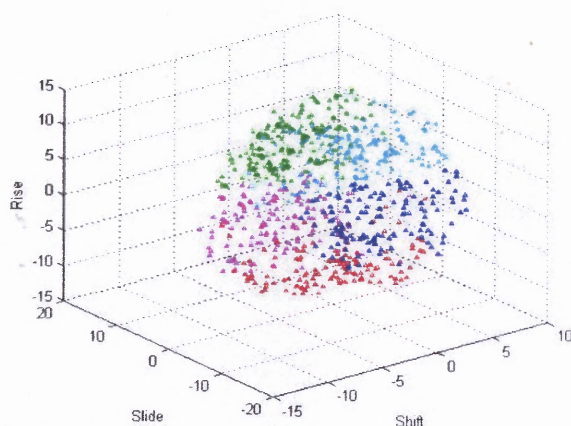


Figure 5.5 (Slide, Shift, Rise) space for $[N \times P2]_T$, $c = 5$.

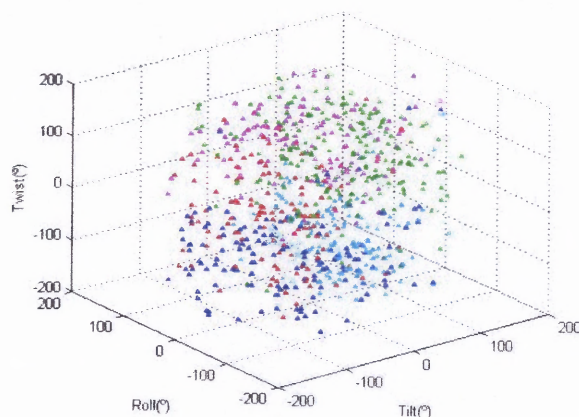


Figure 5.6 (Roll, Tilt, Twist) space for $[NXP2]_{T+R}$, $c = 5$.

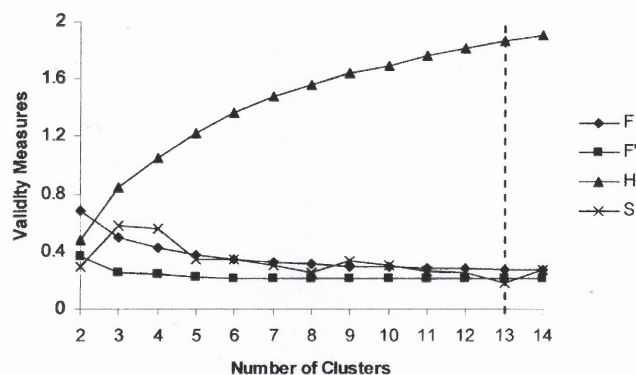


Figure 5.7 Cluster validity plots for the $[NXP2]_R$ proximity matrix.

The cluster validity plots for the $[NXP1]_{T+R}$ (Figure A.6 in Appendix A) and $[NXP1]_T$ (Figure A.8) proximity matrices indicate $c = 4$ as the optimum number of clusters (compare with Figures 5.1 and 5.4, respectively, for the $(NXP2)$ proximity matrix that suggest $c = 5$ as the optimum value). However, as the two phenyl rings in the bisphenyl moiety do not rotate freely due to the coupling of the rotational energy profile of one to that of the other,¹²² the partitions produced by the $(NXP1)$ proximity matrices would be expected to be similar to the ones produced by their $(NXP2)$ counterparts.

When the (NxP1) proximity matrices were explored further by using the value of $c = 5$ as indicated in Table 1 for the corresponding (NxP2) proximity matrices, the resulting plots were similar to the (NxP2) plots. This is illustrated in Figure A.7 which is very similar to Figure 5.2.

5.1.2 B-Side Clustering

As shown in Figure 5.8 for the $[C_{xP2}]_{T+R}$ proximity matrix, the cluster validity measures for the B-side clustering are not as conclusive as those for the full-molecule clustering. S behaves well over $2 \leq c \leq 6$, after which it takes very large values, which indicates an infinitesimally small distance between the closest prototype centers for all $c > 6$. In other words, good clusters are arbitrarily subdivided into artificial, overlapping clusters for all $c > 6$. This prompted a search for a good partition in the range $2 \leq c \leq 6$. In this range, S attains its lowest value at $c = 3$ and the normalized partition coefficient, F' , also attains its maximum value. The other two indices, F and H , are non-indicative for $2 \leq c \leq 6$.

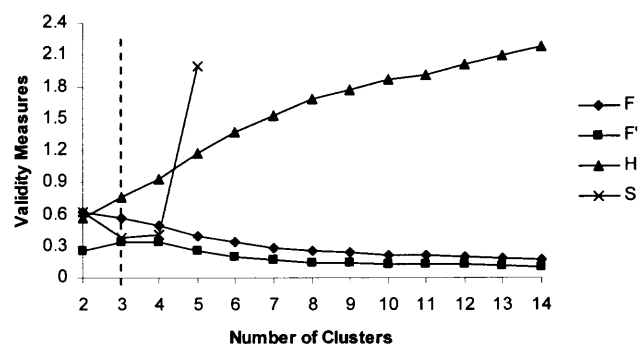


Figure 5.8 Cluster validity plots for the $[C_{xP2}]_{T+R}$ proximity matrix.

Figure 5.9 shows the conformations plotted in (Slide, Shift, Rise) translational space for $c = 3$ for the $[C_xP2]_{T+R}$ proximity matrix clustering. The conformations separate well in translational space and this can be seen particularly in the (Shift, Rise) and (Slide, Rise) plots of Figures 5.10 and 5.11, respectively.

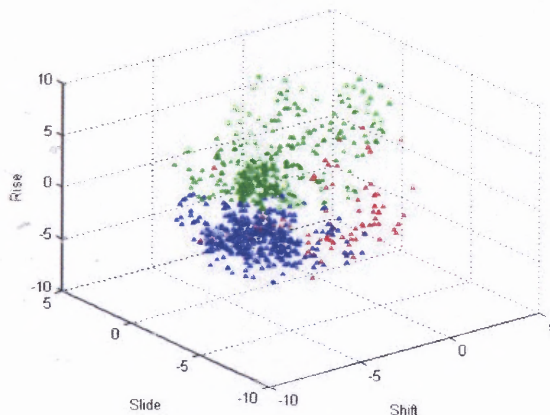


Figure 5.9 (Slide, Shift, Rise) space for $[C_xP2]_{T+R}$, $c = 3$.

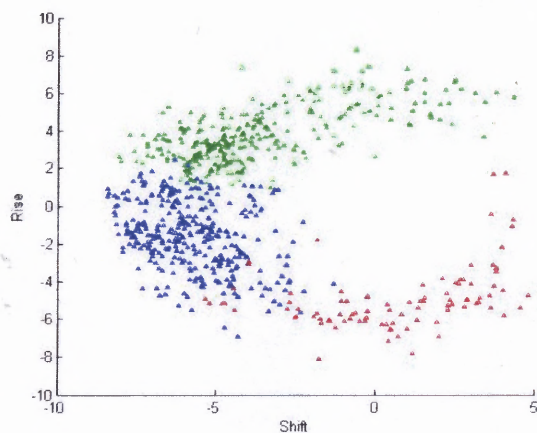


Figure 5.10 (Shift, Rise) space for $[C_xP2]_{T+R}$, $c = 3$.

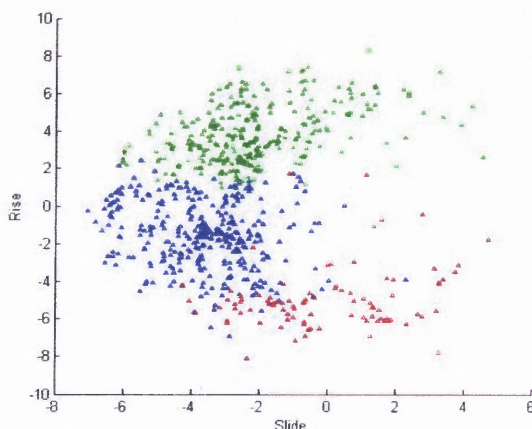


Figure 5.11 (Slide, Rise) space for $[CxP2]_{T+R}$, $c = 3$.

The cluster validity plots for $[CxP2]_T$ (Figure 5.12) and $[CxP2]_R$ (Figure 5.13) identify $c = 4$ and $c = 8$, respectively, as the optimal number of clusters. However, when $c = 3$ was used instead for the $[CxP2]_T$ clustering then the plot of conformations in (Slide, Shift, Rise) space looked as shown in Figure 5.14 and this partition is similar to the one shown in Figure 5.9 for $[CxP2]_{T+R}$. Note, therefore, that a sub-optimal three-cluster partition for $[CxP2]_T$ has been compared to the optimal three-cluster partition for $[CxP2]_{T+R}$. Similar results were found for proximity matrices involving P1 (plots provided in Appendix A). The plot of conformations in (Slide, Shift, Rise) space for $[CxP1]_{T+R}$ is given in Figure A.15 and is very similar to Figure 5.9 for $[CxP2]_{T+R}$. Since the optimal number of clusters based on a proximity matrix defined by only the translational features is very similar to that defined by both translational and rotational features, this seems to indicate that the translational, rather than the rotational, parameters determine the B-side clustering. This is similar to the full-molecule clustering case and is possibly again due to the fact that the C- and P1- or P2-planes are relatively far apart.

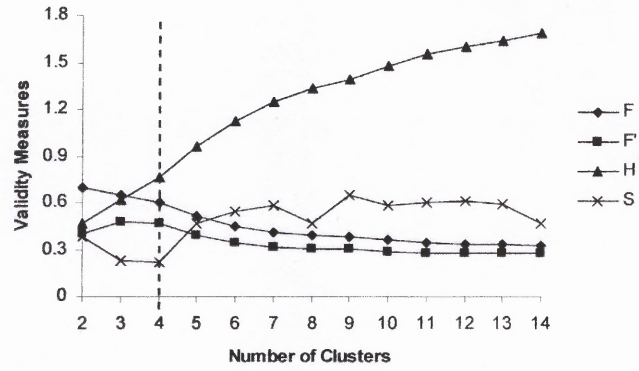


Figure 5.12 Cluster validity plots for the $[CxP2]_T$ proximity matrix.

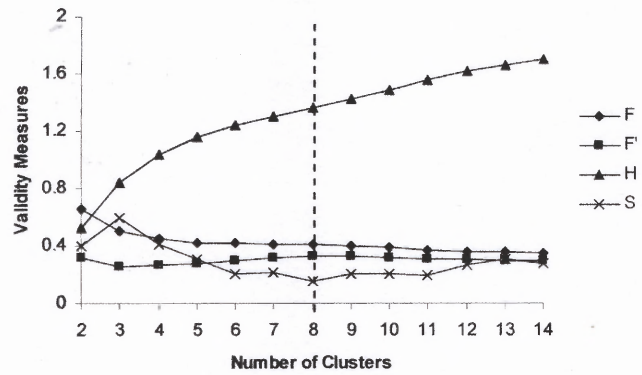


Figure 5.13 Cluster validity plots for the $[CxP2]_R$ proximity matrix.

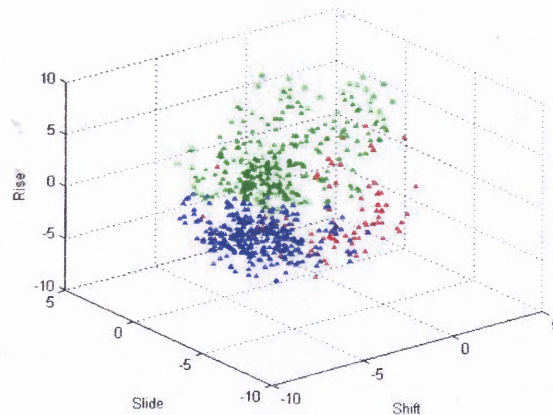


Figure 5.14 (Slide, Shift, Rise) space for $[CxP2]_T$, $c = 3$.

5.1.3 A-Side Clustering

The A-side is described by the proximity matrix constructed from the $[NxC]_{T+R}$ feature space. As in the full molecule case, separate studies were carried out for proximities resulting from the $[NxC]_T$ and $[NxC]_R$ feature spaces. Figure 5.15 shows the cluster validity plots for the A-side for $[NxC]_{T+R}$ over the range $2 \leq c \leq 14$. At $c = 9$, F and F' take their maximum values and H and S take their minimum values. Thus, unlike the full-molecule partitions, all four validity measures seem to be in agreement in this case. Conformations plotted in the (Slide, Shift, Rise) translational and (Roll, Tilt, Twist) rotational space are shown in Figures 5.16 and 5.17, respectively. The separation of conformations into nine clusters is clearly visible in both translational and rotational space. In contrast to the full-molecule and B-side clustering results, for A-side clustering both the translational and rotational parameters appear to play a role in separating the conformations into clusters. This may be because N- and C-planes are, for most of the conformations in this study, much closer in space than the N- and P1- or P2-planes or the C- and P1- or P2-planes.¹²² A complete analysis of the conformational profile of **1** will be given in a separate publication.¹²² The proximity of the N- and C-planes indicates that their relative rotation as well as their relative separation is important to the clustering. For planes that are far apart (such as the N- and P1- or P2-planes), their relative rotational orientation seems to be of lesser significance to clustering than their distance of separation.

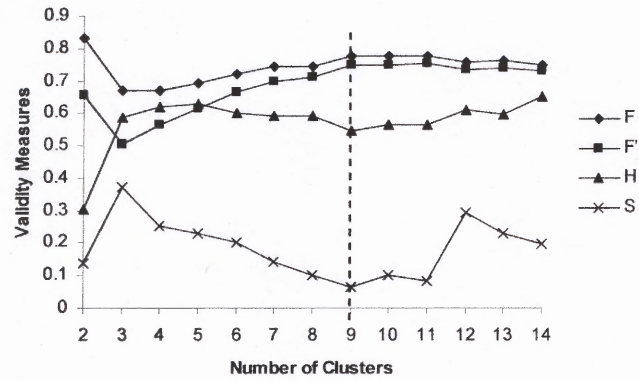


Figure 5.15 Cluster validity plots for the $[NxC]_{T+R}$ proximity matrix.

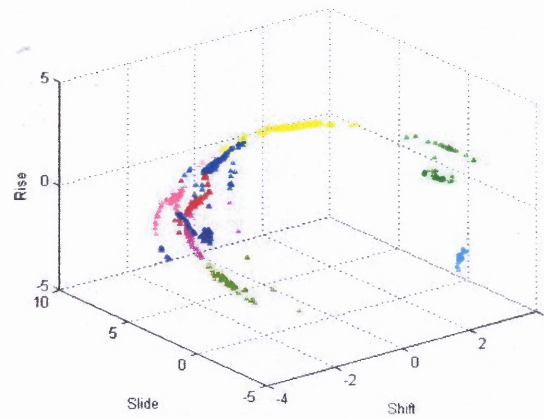


Figure 5.16 (Slide, Shift, Rise) space for $[NxC]_{T+R}$, $c = 9$.

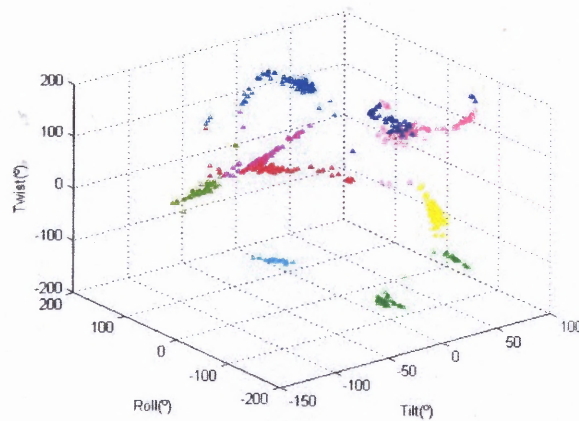


Figure 5.17 (Roll, Tilt, Twist) space for $[NxC]_{T+R}$, $c = 9$.

5.3 Discussion

The superposition-independent cluster analyses presented here were performed with the aim to propose an intuitive and generalizable protocol for feature extraction and clustering. A conformation (or a fragment in it) is represented by orientational parameters derived from a pair of planes. A numerical notion of dissimilarity between two conformers is then based on these parameters. Such pair(s) of planes can be easily identified in any flexible large molecule and orientational parameters can be calculated. Since the feature extraction procedure in its present form only considers a pair of planes (planes located at the structural extremities of the molecule), the process is less sensitive to structural singularities, especially when compared to the molecular-reconstruction-based feature extraction technique for the superposition-dependent cluster analyses (Chapter 4). However, the protocol can be generalized to include more than just a pair of planes. The full-molecule feature set can also be alternatively defined by 12 orientational parameters, six from the (NxC) pair and six more from the (CxP2) pair. In other words, the full molecule can be considered to be a combination of the A-side and the B-side feature sets in their entirety.

The superposition-independent feature set developed here is fundamentally different from the superposition-dependent one in Chapter 4. Although the underlying data set is the same in both cases, the feature sets, and therefore the feature vectors (and also the representative structures identified), are not directly comparable. This might explain why, for superposition-dependent A-side clustering, $c = 3$ is considered the best partition, while $c = 9$ is a natural choice for the same side here. It might also explain why the superposition-dependent feature set was not able to detect clusters for the case of the

full-molecule clustering, while the superposition-independent feature set did. A powerful feature vector may be constructed by combining the two approaches. In the molecular-reconstruction-based approach, the rotational aspect for a pair of planes was represented by a single parameter – the angle at which the planes intersect. In addition, specific heavy atom locations were also considered as part of the feature vector, making the feature vector dependent upon superposition. In the superposition-independent analysis, however, a pair of planes was represented by three rotational parameters in addition to three translational parameters. Thus, an alternative, more comprehensive superposition-dependent feature vector could combine atom coordinates and the six planes parameters.

For the B-side and the full-molecule clustering, where the extreme planes were located further apart, it was seen that the three translational parameters played a predominant role in creating a partition. This is true even though the rotational parameters were scaled by an order of magnitude when the proximity (dissimilarity) measure was calculated. The fact that translational parameters outweigh the rotational parameters makes intuitive sense, and this was captured by the cluster analysis as well. The construction of a feature set based on molecular planes parameters also enables the user to separate the relative translation and rotation effects and study the contributions of each of these individually.

In almost all the visualization plots shown here, with the exception of 3-D plots for the A-side clustering, the clusters (and the data) appear to be continuous (as opposed to being discrete and widely separated). A natural question is – why search for clusters in a continuous data set? However, the data being continuous could be an artifact of the visualization space. Clustering results for a higher dimensional space must be viewed in

a lower dimensional visualization space, where only half (or one-third) of the dimensions used in clustering are represented. In other words, translational or rotational parameters individually may appear to be random and continuous in three-dimensional space but this does not necessarily mean that in the higher six-dimensional space the data is continuous as well. Even if the data is continuous in the higher dimensional clustering space, the objective of clustering is to locate widely dissimilar representative patterns. The clusters identified cover a distinct region within this continuous space and hence a representative conformer (a mean-located conformer within such a distinct region) is truly dissimilar to other identified representatives.

CHAPTER 6

QSAR OF METHYLPHENIDATE ANALOGUES: METHODS

6.1 Introduction

The preceding chapters focused on a challenging aspect of ligand-based drug design: identification of useful conformations of a small molecule in the absence of structural information about its protein receptor. In the present case, **1** is a GBR 12909 analogue that binds to the DAT, blocking the reuptake of dopamine into the presynaptic dopaminergic nerve terminal leading to its pharmacological effects. As the three-dimensional structure of the DAT remains as yet undiscovered, computational efforts that target this system include strategies based on ligand information. The number of ligands in the GBR 12909 analogue data set for which structure and binding affinity data is available is well over 200 and includes some stereochemical information as well.^{47,49,50,114-117,123-126} In the case of the MP analogue data set, this number is 80.^{37,39,43,44} As the previous chapters indicate, however, consideration of the three-dimensional structure of even a single, flexible analogue such as **1** can be expensive. For a data set with a large number of analogues, some of which are even more flexible than **1** and can thus assume many more conformations, three-dimensional studies may be prohibitively expensive. The purpose of cluster analysis of the conformations of **1** was to identify putative bioactive conformations that could be used in 3-D QSAR analyses such as CoMFA studies. Yet, three-dimensional studies like CoMFA are prone to extensive considerations of specific conformations. If the identification of these conformations is

performed in the absence of credible structural information, then it would limit the amount of faith that could be placed in even the numerically good models that are generated.

On the other hand, “two-dimensional” studies that consider only the connectivity (or chemical graph) of the atoms in a molecule do not need conformational information. Two-dimensional structural descriptors are non-empirical numerical properties calculated from the connection table of a molecule (for example, from elements, formal charges, and bonds but not from atomic coordinates). Such 2-D descriptors are, therefore, well suited for studies involving large data sets. For this reason, the QSAR studies in this work included several studies performed on the MP analogue data set using 2-D descriptors.

The Molconn-Z module of SYBYL from Tripos¹⁰⁸ and the Molecular Operating Environment¹²⁷ (MOE) from the Chemical Computing Group were used to calculate separate descriptor spaces for each of the neutral and the protonated data sets. Molconn-Z descriptors include a wide range of 2-D topological indices of molecular structure, such as molecular connectivity Chi indices;¹²⁸⁻¹³⁰ Kappa shape indices;^{131,132} electrotopological state (E-state) and hydrogen E-state indices;¹³³ atom type and bond type electrotopological state indices,¹³³ topological equivalence indices and total topological index,⁶⁹ several information indices,⁶⁹ and counts of graph paths, atoms, atom types, and bond types. A total of 524 Molconn-Z descriptors can be calculated. The use of these 2-D descriptors in QSAR analyses has been reviewed recently.⁶⁹ The MOE package calculates three types of descriptors: 2-D, internal 3-D (descriptors that depend upon conformation but not on orientation in space), and external 3-D (descriptors that depend upon conformation as well as on orientation in space). This package can calculate 184 2-

D, 57 internal 3-D, and 10 external 3-D descriptors. There is some overlap with the Molconn-Z 2-D descriptors set. Internal 3-D descriptors from MOE include potential energy descriptors; surface area, volume, and shape descriptors; and conformation-dependent charge descriptors. The difference between the internal and external 3-D descriptors can be understood by using the dipole moment as an example. While the overall magnitude of the dipole moment depends upon the three-dimensional conformation, it does not depend upon the specific orientation in space of that conformation (that is, on an absolute reference frame). Thus, the overall dipole moment is an internal 3-D descriptor because it depends on internal coordinates only. However, if the dipole moment is resolved into its three Cartesian components then each of these components would be an external 3-D descriptor, as it would depend on both the conformation as well as its specific Cartesian coordinates. Since for the MP analogue data set no information about the 3-D structure of the DAT is available, external 3-D descriptors would not be useful and hence were not calculated. Previous work²²⁻²⁵ on this data set proposed the MP GEM conformation as a putative bioactive MP conformation. For this reason, the QSAR studies in this work also included studies performed using internal 3-D descriptors from MOE. In addition, 3-D QSAR (CoMFA) studies based on the GEM conformation were also performed here as a comparison. As in the previous studies,²²⁻²⁵ the neutral and protonated species of the analogues were treated separately.

The presence of a common 14-atom scaffold in all molecules in the MP analogue data set (see Figure 6.1) allows the calculation of atom-level E-state indices for these 14 atoms. The E-state is a combination of electronic and topological information obtained at the atom level.¹³³ The E-state index for an atom in a molecule is a measure of both the

electron accessibility and the topological accessibility of that atom. These indices are correlated with Mulliken-Jaffe electronegativity¹³³ and in modeling studies have been found to be correlated with (among others): ¹⁷O NMR frequencies¹³⁴⁻¹³⁷ for ethers, aldehydes, and ketones; binding of indolealkylamines binding to 5-HT₂ receptors;¹³⁴ binding of barbiturates to beta-cyclodextrin;¹³⁰ binding of corticosteroids;¹³⁸ and inhibition of flu virus by benzimidazoles.¹³³ Thus, the rationale for the calculation of the E-state indices for each of the 14 atoms in the scaffold is that the variations in the scaffold atoms might be correlated with the variations in structure due to substitutions on the scaffold atoms. Furthermore, these structural variations might be correlated with the binding affinity of the analogues with the DAT. In addition, as will be seen below, the E-state indices were used in a novel procedure for determining the constitution of the test set used for model validation.

The procedure for developing a QSAR model can be divided into three steps: data preparation, data analysis, and model validation. In the first stage (data preparation), a suitable data set comprising of a series of molecules is selected, the molecular descriptors for each molecule in the series are calculated, and appropriate statistical methods for data analysis and correlation are chosen. The second stage (data analysis) includes the construction of statistical models that correlate the values of the descriptors, which constitute the independent variables, with the value of the biological activity (or bioactivity), which serves as the dependent variable. In the third stage (model validation), the reliability of the models is assessed by testing their ability to reproduce the bioactivity of similar molecules.

This chapter describes the procedure for developing QSAR models for the MP analogue data set. Section 6.2 describes the data preparation stage. Sections 6.3 and 6.4 describe the data analysis stage; exploratory models that were developed using only the E-state indices for the atoms of the scaffold are covered in Section 6.3 while more rigorous models using a variety of descriptors are explained in Section 6.4. Section 6.5 describes the model validation stage.

After data preparation, the first task was to determine the adequacy of the E-state indices of the scaffold atoms for use in modeling the MP analogue data set. Exploratory regression models were developed for this purpose. Thus, forward stepwise regression analysis (Section 6.3.1) was successfully used to see whether the E-state indices could correctly characterize the substitution pattern observed due to the substitutions at known sites on the scaffold. Since this meant that the E-state indices could encode structural information for this data set, they were used to identify a representative subset of analogues that could be used as a test set in model validation studies (Section 6.3.2). *The use of E-state indices for determining test sets is a novel aspect of the present work.* The training set identified during this process was used to develop a preliminary “all possible subsets” regression model (Section 6.3.3) using E-state indices only. When this model was used to predict the test set analogues, most of the predictions were good. However, a few residuals were not meaningful suggesting that this preliminary model was not adequate. This set the stage for developing more rigorous and robust partial least squares regression models (Section 6.4). These models used 2-D and 3-D descriptors other than the E-state indices (see above) in various combinations. The best performing models

from these rigorous models were selected for validation using the test set developed earlier (Section 6.5).

Chapter 7 presents the results of these QSAR studies and includes the data analysis and model validation stages.

6.2 Data Preparation

The MP analogue data set includes 80 molecules all of which share an important feature: a common scaffold consisting of 14 atoms (Figure 6.1; the 14th atom is included in R3). The analogues with their substituents and DAT binding affinity are listed in Table B.1 in Appendix B. MP is analogue **39** with R1 = H, R2 = H, and R3 = CO₂CH₃. There are three possible sites of substitutions on MP which produce the other 79 analogues in the table: 1) the phenyl ring (R1), 2) the piperidinyll nitrogen (R2), and 3) the side chain (R3), which includes atom 14 as the carbon atom attached to atom 7. It should be noted in Table B.1 that the MP analogues are numbered 1 through 80; the numbering scheme is distinct from that of the GBR 12909 analogues in the first part of this work.

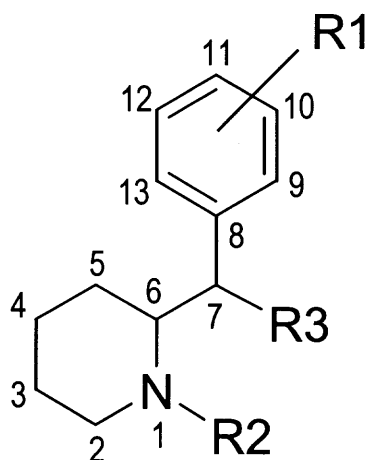


Figure 6.1 The scaffold for the MP analogue data set.

The respective GEM conformations²⁴ for neutral and protonated MP were selected and all 80 analogues were constructed in these conformations to give two data sets: the neutral MP (MPN) data set and the protonated MP (MPP) data set. The site of protonation is the nitrogen atom (atom 1) in the piperidine ring.

6.2.1 MPN Data Set

For each analogue, the piperidine ring was maintained in the chair configuration and its substituents were constructed in the equatorial positions. The analogue was minimized using the MMFF94¹³⁹ force field and charges. Systematic search was performed on all torsional angles in the analogue and the resulting conformations were ranked by energy. As implemented in SYBYL,¹⁰⁸ systematic search increments each specified torsional angle by a specified increment, examines the resulting conformations for van der Waals contacts, and calculates the energy of each conformation without geometry optimization. The lowest energy conformation from the systematic search was selected and minimized. If this new, minimized conformation had a lower energy than the minimized conformation prior to systematic search, then the new conformation was selected for the data set, otherwise the original was kept. The minimization and search parameters are provided in Appendix C. The resulting analogues were aligned on the five central atoms (atoms 6, 7, 8, 14 and the hydrogen on atom 7). This alignment is important for the CoMFA studies that were performed on the data sets. The template MPN GEM conformation is shown in Figure 6.2. The four main torsional angles for this conformation (T1, T2, T3, and T4) are provided in Appendix D.

There was one exception to the above procedure. For analogue **48** (2-OH MPN), the above procedure led to hydrogen bonding between the oxygen and the hydrogen on the piperidinyl nitrogen (2.06 Å). This distorted the analogue significantly from the template MPN GEM conformation. Since some of the QSAR models to be developed include 3-D descriptors, this distortion could cause a model to be unreliable. For this reason, it was decided to rotate the phenyl ring in this analogue by 180°, causing the 2-OH substituent to move out of the way of the piperidine ring while still maintaining the original MPN GEM conformation.

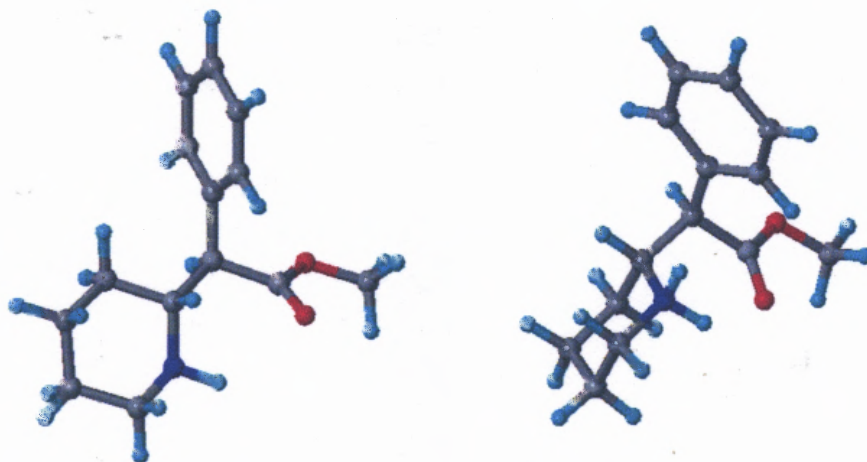


Figure 6.2 The MPN (left) and MPP (right) GEM conformations.²⁴

6.2.2 MPP Data Set

The procedure outlined above for the MPN data set (analogue construction, minimization, systematic search, and selection) was also followed for the MPP data set with several exceptions. Since the presence of the proton in the MPP GEM conformation affects the rotational barrier around T1 and makes this conformation a little more

constrained than the MPN GEM conformation,²⁴ some MPP analogues behaved differently from the others. These were analogues **2**, **4**, **5**, **25**, **28**, **66**, **70**, **71**, and **75** (see Table B.1 for structures). When the procedure outlined above was followed for these nine analogues, it resulted in final geometries that were significantly distorted from the template MPP GEM conformation. The reason behind the distortion was the selection by systematic search of a conformation in which the torsional angle T1 was significantly different from that in the GEM conformation (see Appendix D). In the case of the MPP data set, since there is a proton on the piperidinyll nitrogen for all 80 analogues, hydrogen bonding can and does occur between a side chain oxygen and a hydrogen on the nitrogen. However, in the case of the distorted analogues (all except analogue **75**), such hydrogen-bonding *did not* occur. Analogue **75** has two oxygen atoms in the side chain that may hydrogen-bond. For this analogue, the original procedure produced both distortion in T1 as well as a different side chain orientation because of a different hydrogen-bonding pattern. In all other non-distorted analogues, both T1 and the side chain were positioned in the same hydrogen-bonding pattern as in the GEM conformation. Thus, for the nine distorted analogues, the systematic search was repeated but, to minimize distortion, torsional angle T1 was not included in the search and the side chain was oriented such that it assumed the “correct” hydrogen-bonding orientation (i.e., that of the GEM conformation). The selection of the new non-distorted conformations for these nine analogues and the alignment of the MPP data set analogues then followed as for the MPN data set above. The nine non-distorted analogues that were finally selected had the following hydrogen bond distances: **2** (1.81 Å), **4** (1.79 Å), **5** (1.79 Å), **25** (1.79 Å), **28** (1.82 Å), **66** (1.83 Å), **70** (1.84 Å), **71** (1.87 Å), and **75** (1.90 Å). The template MPP

GEM conformation is shown in Figure 6.2 and its T1-T4 torsional angles are listed in Appendix D.

6.3 Data Analysis I: Exploratory Models

6.3.1 Forward Stepwise Regression Analyses and Scaling

To get some idea about the adequacy of the E-state indices for the predictive models developed in the present work, exploratory models using forward stepwise regression¹⁴⁰ were developed. The forward stepwise regression technique progressively includes variables in a model until a satisfactory regression equation is achieved. The E-state indices for the analogues were regressed on the binding affinity of the analogues. The result of this process is the identification of those E-state indices that are the best correlated with the binding activity values.

The size of the data matrix was 14 descriptors x 80 analogues. Each descriptor corresponds to the E-state value of an atom in the scaffold. Thus, the E-state indices were named ES01, ES02, ..., ES14 corresponding to atom 1, atom 2, ..., atom 14 in the scaffold. A total of 16 (eight for each data set) forward stepwise regression models were examined using either scaled or unscaled raw descriptor values or the principal components of the scaled or unscaled descriptors. The scaling methods used were 1) range scaling, 2) standard deviation-based normalization, or 3) mean absolute deviation-based normalization. Range scaling was performed using the formula

$$Z_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}, \quad (6.1)$$

where z_i is the scaled descriptor, x_i is the unscaled descriptor, and x_{\min} and x_{\max} are the respective minimum and maximum values for this descriptor in the data set. Since range scaling is sensitive to outliers, other methods for scaling were also explored. These included using either the standard deviation (sd) or the mean absolute deviation (mad) of the descriptor column for normalization. The sd-based normalization is less sensitive to outliers but still retains squared dependence on the deviations from the mean. Using mad-based normalization removes the squared dependence on the deviation from the mean, making this method least sensitive to outliers. For sd-based normalization,

$$z_i = \frac{x_i - \bar{x}}{\text{sd}}, \text{sd} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}, \quad (6.2)$$

where n is the number of analogues or data points and \bar{x} is the mean of the descriptor column. For mad-based normalization, sd in Equation 6.2 is replaced by mad which is given by

$$\text{mad} = \frac{1}{n} \sum |x_i - \bar{x}|. \quad (6.3)$$

All the preceding calculations were performed in Matlab.¹⁴¹ The Matlab commands used are listed as Appendix F.

As will be seen in the results of these forward stepwise regression analyses in Chapter 7, the raw and unscaled E-state indices performed satisfactorily and were used directly in the all possible subsets regression analyses of Section 6.3.3 and in constructing the test set (Section 6.3.2.4).

6.3.2 Identification of Test Set Analogues

The identification of a test set of analogues is an important part of a model validation strategy.^{142,143} While the PLS analyses performed in Section 6.4.2 employ an internal validation method as well as activity-shuffling (progressive scrambling) to assess the reliability of the models, external validation of a model by predicting the binding affinity of a suitable test set of compounds is more desirable. External validation offers a direct assessment of a predictive model. However, care must be taken to select the analogues for constructing the test set. The test set analogues should span both the structure space as well as the activity space.

The present work developed a novel method to determine the optimal members of the test set by using the E-state indices of the atoms of the scaffold. Three different test sets, developed from a combination of different descriptor selection criteria (see Sections 6.3.2.2 and 6.3.2.4 below), were examined. Each test set was identified using a dissimilarity-based compound selection algorithm.^{144,145} After correlation analysis (see Section 6.3.2.4), select E-state indices corresponding to some of the 14 atoms in the scaffold were used to construct a dissimilarity matrix, *D*. The elements of *D* are the complements of pairwise Tanimoto coefficients.¹⁴⁶ This dissimilarity matrix was then used along with a sphere-exclusion algorithm^{147,148} for identifying a deterministic test set. This ensures that this procedure will always identify the same test set given the same input parameters.

6.3.2.1 Tanimoto Coefficient. The Tanimoto coefficient is frequently used in the determination of intermolecular similarities and database searching applications.^{143,145,149-}

¹⁵¹ The Tanimoto coefficient, S , for the similarity between two molecules, M_i and M_j , is given by

$$S(M_i, M_j) = \frac{\sum_{k=1}^m E_{ik} E_{jk}}{\sum_{k=1}^m E_{ik}^2 + \sum_{k=1}^m E_{jk}^2 - \sum_{k=1}^m E_{ik} E_{jk}}, \quad (6.4)$$

where M_i and M_j are the i^{th} and j^{th} analogues for which the pairwise Tanimoto coefficient is being sought; $i, j = 1, 2, \dots, n$, with $n = 80$ for the MP data set; $k = 1, 2, \dots, m$ with $m = 14$ for the 14-atom E-state indices descriptor set. The E 's in equation 6.4 are the E-state indices for the molecules being used in the current calculation. Note that for $i = j$, $S = 1$.

The dissimilarity matrix, D , is then the complement of the Tanimoto coefficients obtained above in the similarity matrix, S , and is given by

$$D(M_i, M_j) = 1 - S(M_i, M_j). \quad (6.5)$$

Thus, for $i = j$, dissimilarity $D = 0$. The complement of the Tanimoto coefficient is also known as the Soergel distance.

6.3.2.2 Sphere-Exclusion Algorithms. Dissimilarity-based compound selection (DBCS) algorithms have been used effectively for selecting structurally-diverse subsets of a data set of molecules.^{144,145,152} The underlying idea behind using automated compound selection algorithms stems from the assumption that a set of compounds that spans structural space (defined here in terms of E-state indices for the 14 atoms of the scaffold) will also span the biological activity space.¹⁵³ A number of DBCS algorithms have been proposed and include maximum-dissimilarity algorithms¹⁵⁴ and sphere-exclusion (SE) algorithms.¹⁴⁷ All algorithms are initialized by selecting a first test set compound based on some algorithm-specific criterion. Thereafter, maximum-

dissimilarity algorithms rely upon determining the most dissimilar compound in the remaining data set in each iteration of compound selection. Sphere-exclusion algorithms, on the other hand, use a predefined value of a “threshold dissimilarity”, t , and in each iteration reject all compounds in the remaining data set that have a dissimilarity less than t with respect to a compound in the current test set. Thus, t defines the radius of a sphere in the descriptor space and at each stage compounds lying within this sphere are excluded from further consideration as candidates for the test set (see Figure 6.3). A larger value of t would generate a smaller test set and vice versa. In Figure 6.3, the closed circles at the center of each sphere represent compounds that have been selected for the test set. The open circles that lie within a sphere will be excluded from consideration for the test set. The algorithm will execute as long as there remain open circles that lie outside all current spheres.

The sphere-exclusion algorithm can be outlined as follows: 1) Define a threshold dissimilarity value, t ; 2) Initialize the test set by selecting a compound, C , from the data set; 3) Reject compounds that have a dissimilarity with C of less than t ; and 4) Repeat from step 2 until all compounds in the data set have been analyzed. Several variants are possible depending upon the value of t in step 1 and the selection criterion adopted in step 2. A value of 0.15 has been suggested for t .^{153,155} Snarey et al¹⁴⁵ discuss three such variants: the SE-MinSum, SE-MinMin, and the SE-MinMax algorithms. In all three, the first test set compound selected is one that has the smallest sum of dissimilarities relative to the rest of the data set. In the SE-MinSum algorithm, the compound with the smallest sum of dissimilarities is selected in step 2 every time; in the SE-MinMin algorithm, the compound with the smallest minimum dissimilarity is selected; and in the SE-MinMax

algorithm, the compound with the smallest maximum dissimilarity is selected. Thus, all sphere-exclusion algorithms start with the same compound that is located near the center of the data set and progressively move outwards.

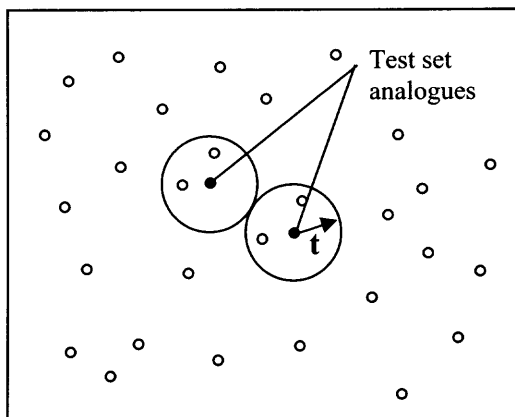


Figure 6.3 Schematic for sphere-exclusion algorithms.

6.3.2.3 D-SIM Version 1.0. A C++ program was developed in this study to assist in fast and automated test set identification based on the above discussion. This program permits the user to input a variable value of t ($0 < t < 1$) and select one of the above sphere exclusion algorithms. By default, $t = 0.15$ and the algorithm used is SE-MinMax. The program uses as input a text file containing the number of rows and the number of columns on the first row and the input matrix of descriptors (see Section 6.3.2.4) following that. The input matrix must have the descriptor values range-scaled so that they lie between 0 and 1. This is to ensure that the Soergel distance calculated above is also well-behaved. The program outputs the row numbers of the compounds selected for the test set. The complementary set of compounds in the data set becomes the training set. The program code for D-SIM version 1.0 is provided as Appendix G.

6.3.2.4 Unsupervised Forward Selection of Redundant Descriptors. Correlation analysis was performed on the 14 raw, unscaled E-state indices of the scaffold atoms to select an ideal set of descriptors before calculating the test set using D-SIM. A program developed by Whitley and coworkers¹⁵⁶ (available from <http://www.cmd.port.ac.uk>), which implements an unsupervised forward selection (UFS) algorithm¹⁵⁶ was used to select the least redundant (or most orthogonal) descriptors. This algorithm first selects the two descriptors that have the smallest pairwise correlation coefficient. Next, it rejects each of the remaining descriptors whose pairwise correlation coefficient with the first two descriptors exceeds a user-specified value, $r_{\max}^2 < 1$. The algorithm then iterates until all descriptors are either selected or rejected. A descriptor is selected if it has the smallest squared multiple correlation coefficient with the previously selected columns. All descriptors with squared multiple correlation coefficients with currently selected descriptors greater than r_{\max}^2 are rejected. Since the selection of additional descriptors is based upon their multiple correlation with those already chosen, the algorithm builds a subset of descriptors that is as orthogonal as possible.

6.3.3 All Possible Subsets Regression Analyses

The 14 raw, unscaled E-state indices were used to develop preliminary all possible subsets regression models. The PROC RSQUARE function of version 9.1 of the SAS System for Windows¹⁵⁷ finds subsets of descriptors that best predict the binding affinity by linear regression. The RSQUARE method was used to perform all possible subsets regression and calculate the models in decreasing order of r^2 magnitude within each subset size. These analyses were done for both the neutral and protonated data sets.

These analyses were done to evaluate the performance of the training set as well as to gauge the adequacy of the E-state indices. For this reason, the size of the input matrix was (66 training set analogues, see Section 7.2) x (14 raw, unscaled E-state indices).

6.4 Data Analysis II: Robust Models

6.4.1 Calculation of Descriptors

The Molconn-Z module in SYBYL was used to separately calculate 524 2-D topological descriptors separately for the MPN and MPP data sets. Similarly, MOE was used to calculate 184 2-D and 57 internal 3-D descriptors for the two data sets. CoMFA steric and electrostatic 3-D descriptors were also calculated for the data sets using SYBYL. The input parameters for the CoMFA calculations are listed in Appendix E.

6.4.2 PLS Analyses

The partial least squares (PLS) module in SYBYL was used to analyze the data using descriptors other than the 14 E-state indices for the MP scaffold. PLS has been shown to work effectively for over-square matrices such as those encountered in QSAR studies.¹⁵⁸⁻

¹⁶¹ All non-CoMFA descriptors (that is, the Molconn-Z and MOE descriptors) were autoscaled and all CoMFA (steric and electrostatic) descriptors used CoMFA standard scaling. The leave-one-out crossvalidation method was used to calculate the crossvalidated r^2 (or q^2), the standard error of prediction (SDEP), and the optimum number of components. The maximum number of components allowed in a model was six. The column filtering value for crossvalidated CoMFA models was set to 2.0

kcal/mol. Non-crossvalidated PLS models used the optimum number of components identified in the crossvalidated run.

In the leave-one-out crossvalidation procedure, one analogue (or row) in the data set was deleted and a QSAR equation was derived from the remaining analogues (or rows). The binding affinity of the deleted analogue was predicted using the derived equation and the deviation from its actual value (residual) was calculated. The procedure continued until every analogue was deleted once and its residual was calculated. The sum of all the squared residuals was calculated as the PRESS (predictive residual sum of squares) statistic. The PRESS statistic was used to calculate the crossvalidated correlation coefficient (q^2) and the crossvalidated standard error of prediction (SDEP). The q^2 is given by

$$q^2 = 1 - \left(\frac{\text{PRESS}}{\text{SS}} \right), \quad (6.6)$$

where SS is the sum of squared deviations from the mean.

6.5 Model Validation

The QSAR models were validated using both activity shuffling (progressive scrambling) and external test set validation.

6.5.1 Progressive Scrambling

Progressive scrambling¹⁶² is recommended for large, redundant data sets that have been used to generate models employing crossvalidation techniques such as leave-one-out crossvalidation that was used in this work. Consider as an example a data set with points that are clustered near each other in hyperspace. Since leave-one-out crossvalidation

works by eliminating one point from the data set, when a point within such a cluster is eliminated during crossvalidation, it will not cause any loss in information. Thus, the predictions obtained from crossvalidation might generate a false sense of confidence in the q^2 value. Progressive scrambling helps to address this problem and works by first partitioning the rows (sorted by activity) in the data set into bins and then shuffling or scrambling the activities (the Y values) a user-specified number of times. Three main statistics are generated: 1) cSDEP (which is the estimated crossvalidated standard error at a user-specified “critical point”), 2) Q^2 (which is the expected value of q^2 at the specified critical point), and 3) dq^2/dr_{yy}^2 (which is the slope, at the critical point, evaluated with respect to the correlation of the original activities versus the scrambled activities). Given a redundant data set, the value of Q^2 is considered conservative such that a fairly low value of Q^2 might indicate a robust original model. A value greater than 1.2 for the slope is considered to indicate an unstable model, that is, a model that changes greatly with small changes in the underlying activity values.

Progressive scrambling was performed in SYBYL with 50 scramblings using a maximum of 10 bins to a minimum of 2 bins. The critical point was specified at the default value of 0.85 and the random seed for every scrambling analysis was set to 123456.

6.5.2 Test Set Validation

External validation of selected models was performed using the test set identified in Section 7.2. The activities of the test set analogues were predicted using these models and the residuals were noted.

CHAPTER 7

QSAR OF METHYLPHENIDATE ANALOGUES: RESULTS

7.1 Forward Stepwise Regression

The results of the forward stepwise regression (see Section 6.3.1) are tabulated in Table 7.1. Sixteen models were developed as described below using only the E-state indices of the 14 scaffold atoms. A model was named according to the type of input matrix used. The following scheme was used: the first three letters of the model identify the data set (MPN for the neutral data set and MPP for the protonated data set); the fourth letter indicates whether the E-state indices were used “as is” (R, for raw values) or their principal components (P) were used; the last letter indicates the type of scaling used (U for unscaled, R for range-scaled, S for sd-based normalization, and M for mad-based normalization). For example, MPN-RS indicates that this model was developed using an input matrix that contained raw E-state indices for the MPN data set that were scaled using sd-based normalization. The table lists the statistics obtained for each regression model. The r^2 statistic is a measure of the extent to which the model explains the variance in the data. The adjusted r^2 statistic is used as an alternative to r^2 and is adjusted for the number of descriptors included in the stepwise regression model. While r^2 always increases if more variables are included in the model, the adjusted r^2 statistic can fall if adding a variable does not contribute to explaining the variance. It is, therefore, sometimes a more appropriate statistic than r^2 . The F-value tests the null hypothesis that none of the descriptors has any effect on the binding affinity (that is, all the regression coefficients obtained in a model are zero). The p-value (or the probability value) can

help decide whether to reject the null hypothesis. The p-value is the probability, given the null hypothesis, of obtaining a chance F-value greater than the calculated F-value. In other words, a large F-value along with a small p-value suggests rejection of the null hypothesis. The smaller the p-value the greater the evidence against the null hypothesis. The root mean square error (RMSE) is the value that is minimized in regression analysis. Thus a good model would tend to have a high value for r^2 , adjusted r^2 , and the F-value, and a low value for RMSE and the p-value.

Table 7.1 Forward Stepwise Regression Results

<i>Model</i>	<i>Important Atoms</i>	r^2	<i>Adjusted r^2</i>	<i>RMSE</i>	<i>Intercept</i>	<i>F-value</i>	<i>p-value</i>
MPN-RU	9, 10, 11, 12	0.388	0.347	0.734	4.987	11.883	1.57E-07
MPN-RR	9, 10, 11, 12	0.388	0.347	0.734	5.625	11.883	1.57E-07
MPN-RS	9, 10, 11, 12	0.388	0.347	0.734	6.760	11.883	1.57E-07
MPN-RM	9, 10, 11, 12	0.388	0.347	0.734	6.760	11.883	1.57E-07
MPN-PU	1, 9, 10	0.499	0.458	0.668	6.760	14.738	5.13E-10
MPN-PR	1, 10, 11	0.477	0.442	0.678	6.760	17.075	5.39E-10
MPN-PS	1, 9, 11, 12	0.487	0.445	0.676	6.760	14.023	1.23E-09
MPN-PM	9, 11, 12	0.505	0.465	0.664	6.760	15.081	3.40E-10
MPP-RU	9, 10, 11	0.347	0.313	0.752	6.093	13.491	3.81E-07
MPP-RR	9, 10, 11	0.347	0.313	0.752	6.105	13.491	3.81E-07
MPP-RS	9, 10, 11	0.347	0.313	0.752	6.760	13.491	3.81E-07
MPP-RM	9, 10, 11	0.347	0.313	0.752	6.760	13.491	3.81E-07
MPP-PU	7, 9, 10, 14	0.496	0.463	0.665	6.760	18.481	1.32E-10
MPP-PR	10, 11	0.441	0.412	0.696	6.760	20.016	1.17E-09
MPP-PS	1, 7, 10, 11, 12, 14	0.476	0.441	0.679	6.760	17.056	5.49E-10
MPP-PM	7, 10, 11, 12, 14	0.523	0.470	0.661	6.760	11.268	1.50E-09

In Table 7.1, the column marked “Important Atoms” lists the atoms that were found by the model to be the most important for bioactivity. For the raw E-state models, these are simply the atoms whose descriptors were selected in the corresponding forward stepwise regression analysis. For the principal components-based models, the atoms listed in the table had the highest coefficients (with absolute value greater than or equal to 0.3) in the principal component that had the maximum correlation with the binding

affinity. This principal component was invariably *not* the component with the highest associated variance. Table 7.1 suggests that all models identified the phenyl ring (consisting of atoms 8, 9, 10, 11, 12, and 13, see Figure 6.1) as being important for bioactivity. Most models based on the principal components of the MPN data set identified both the phenyl ring and the piperidinyll nitrogen (atom 1) as important. Most models based on the principal components of the MPP data set identified both the phenyl ring and the side chain (atom 14) as being important for bioactivity. This could be because in the protonated data set, hydrogen-bonding between an oxygen atom in the side chain and a hydrogen atom on the piperidinyll nitrogen occurs in *all* analogues (see Section 6.2.2). One model (MPP-PS) identified all three possible sites of substitutions as being important. The unscaled models also performed satisfactorily (with r^2 of 0.35 or above) suggesting that E-state indices could be used without scaling. The raw E-state models were not affected by scaling while the principal components-based models were. Since scaling “shifts” the data points from their original positions and since the principal components were obtained *after* scaling, the principal components-based models would be expected to be sensitive to scaling.

In general, the principal components-based models had higher r^2 , adjusted r^2 , and F-values and lower RMSE values than the raw E-state models. The highest r^2 was achieved for the MPP-PM model but its F-value was relatively low. The model MPP-PS, which identified all three regions of substitution as important, had fairly high r^2 and F-values. The p-values for all models were very low suggesting a significant probability of the existence of a relationship between E-state indices and binding affinity (pIC_{50}). The forward stepwise analyses suggest that E-state indices could be used for building more

elaborate predictive models. However, this aspect was not pursued further. Instead, the relationship between the E-state indices of the scaffold atoms and the DAT binding affinity of the analogues was used to derive a novel procedure for identifying a representative test set of analogues.

7.2 Test Set Identification

A new method to determine the test set by using E-state indices was developed and is an important aspect of the present work. While there are three main regions of substitution (see Figure 6.1), the analogues can be grouped based on all possible combinations of substitutions with respect to the parent compound, MP (compound **39** in Table B.1 with $R1 = H$, $R2 = H$, and $R3 = CO_2CH_3$). The data set contains 29 analogues with substitutions only on the phenyl ring (R1), 23 analogues with substitutions only on the piperidinyll nitrogen (R2), five analogues with substitutions only in the side chain (R3), seven analogues with substitutions at both the phenyl ring and the nitrogen (R1+R2), four analogues with substitutions at both the phenyl ring and the side chain (R1+R3), nine analogues with substitutions at both the nitrogen and the side chain (R2+R3), and two analogues with substitutions at all three sites (R1+R2+R3). This is summarized in the row marked “Data Set” in Table 7.2, where the total number of analogues, including MP, is 80. Table B.1 also lists this information for each analogue in the column titled “Group”. It is also possible to consider the data set such that the substitutions occur *at least* at R1 (this would group analogues for R1, R1+R2, R1+R3, and R1+R2+R3), *at least* at R2 (R2, R1+R2, R2+R3, and R1+R2+R3), and *at least* at R3 (R3, R1+R3,

R2+R3, and R1+R2+R3). These additional groupings are shown in Table 7.2 under R1*, R2*, and R3*, respectively, and the analogues are listed in Table B.2 in Appendix B.

Table 7.2 Test Set Identification Results

	R1	R2	R3	R1+R2	R1+R3	R2+R3	R1+R2+R3	Total ^a	R1*	R2*	R3*	Test Set Analogues
Data Set	29	23	5	7	4	9	2	79	42	41	20	
TS1	8	1	0	2	1	1	1	14	12	5	3	9, 26, 27, 32, 33, 34, 45 53, 56, 57, 60, 67, 71, 80
TS2	4	1	1	1	1	0	1	9	7	3	3	11, 27, 32, 37 56, 60, 68, 71, 80
TS3	6	1	1	3	1	1	1	14	11	6	4	11, 26, 27, 30, 32, 33, 34 41, 56, 57, 62, 67, 71, 80
^a For Data Set, total refers to the number of analogues in each category. For the test sets, total refers to the total number of test set analogues.												

Three test sets were examined using different combinations of r_{\max}^2 for selecting least redundant of the 14 raw, unscaled E-state indices prior to test set identification (see Section 6.3.2.4) and the threshold radius, t , for sphere-exclusion (see Section 6.3.2.2). All test sets were constructed using the SE-MinMax algorithm. The first test set, TS1, was developed using $r_{\max}^2 = 0.99$ and $t = 0.10$. The suggested^{153,155} value for t is 0.15. However, a smaller value was used for TS1 so as to obtain a larger test set that would be more likely to span the structural variation in the data set. For this test set, the UFS algorithm identified nine E-state indices (corresponding to atoms 1, 3, 7, 9, 10, 11, 12, 13, and 14) that had a squared multiple correlation coefficient $< r_{\max}^2 = 0.99$. These nine descriptors were range-scaled and used as input into D-SIM. Range-scaling of D-SIM input descriptors such that each descriptor value is between 0 and 1 is essential otherwise the complement of the Tanimoto coefficient (see equations 6.4 and 6.5) may not be meaningful (it might be outside the interval [0, 1] and could assume negative values). The resulting test set of 14 analogues, which is 18% of the data set, is listed in Table 7.2.

This result shows that this test set could not identify an analogue with only R3 (side chain-only) substitution. Also, the distribution is skewed toward R1* analogues even though both R1* and R2* have a nearly equal number of analogues (42 and 41, respectively).

For the second test set, TS2, r_{\max}^2 was reduced to 0.85. The UFS algorithm rejected a larger number of correlated descriptors and selected only six least redundant descriptors (corresponding to atoms 1, 9, 10, 11, 12, and 14). This is a better result than before as the corresponding atoms are exactly the sites of substitutions for all analogues. These six descriptors were range-scaled and input into D-SIM using the suggested^{153,155} value of 0.15 for t . Expectedly, a larger value of the threshold radius produced a smaller test set with only nine analogues (11% of the data set). As shown in Table 7.2, this test set did not include an analogue with substitutions on both the piperidinyll nitrogen atom and the side chain (R2+R3). The skewness toward R1* also remained.

The third test set, TS3, was developed using the same r_{\max}^2 value as for TS2 (0.85) but retained the value of t that was used for TS1 (0.10). Since r_{\max}^2 was the same as for TS2, the same set of six least redundant descriptors was obtained. Table 7.2 shows the result for TS3 to be the best: all possible combinations of substitutions were represented in this test set and the skewness toward R1* was also reduced slightly. Examination of the 14 analogues (18% of the data set) identified by this test set also suggests that the actual substitutions are wide-ranging and representative. In terms of binding affinity values also, this test set performs very well with activity ranging from 5.33 to 8.77 log units, which is a difference of greater than 3 log units. This test set was selected for all model validation studies.

The training set corresponding to the test set TS3 contains the complementary 66 analogues (80 - 14). This training set was selected to develop further exploratory models in Section 7.3. It is used in Section 7.4 to develop more rigorous predictive QSAR models. Some of these models were externally validated by predicting the binding affinity values for the analogues forming TS3. These validation results are described in section 7.5.

7.3 All Possible Subsets Regression

The results of all possible subsets for the MPN and MPP training sets are attached as Appendix H. Note that each training set contains the same 66 analogues that are neutral for the MPN study and protonated for the MPP study. For these analyses, the r^2 statistic was used as the selection criterion. Thus, the results display the best models for a given subset size in order of decreasing r^2 . Some interesting observations on the resulting models follow. For both data sets, the phenyl ring substitutions have the most effect on progressive improvement in r^2 . For instance, ES09, ES10, and ES11 (corresponding to atoms 9, 10, and 11 in Figure 6.1) together raise the r^2 to 0.45 in the three-descriptor models for both data sets. For the MPP training set, at least two six-descriptor models (in bold type in Appendix H) identified all three substitution sites (the phenyl ring, N1, and C14) as important. These models have r^2 values of 0.49 and 0.48, which are comparable to that of the other six-descriptor models. The full 14-descriptor models for both the MPN and MPP data sets have r^2 of 0.60. Table 7.3 lists the residuals obtained for the full models. While 12 of the analogues were well predicted by the models, two analogues (**26** and **27**) were very poorly predicted. Indeed, their predicted activity is not meaningful as

it is predicted as a negative log number. Appendix H shows that while these two models have fairly high r^2 values, the standard errors for individual coefficients are very large, possibly implying that the model is not appropriate for predictions. As will be seen in Section 7.5.2, a more rigorous model was able to significantly improve the residuals for analogues **26** and **27**.

Like any other statistical method, while this method can serve as a useful exploratory model-building tool, it cannot be relied upon to determine the true functional form of a model. From observations like those listed above, it is possible to get a general intuition about E-state indices and to surmise that they can be broadly informative. Together with the initial stepwise regression models described in Section 7.1, the results in this section allow a measure of confidence in the utility of these descriptors. In the next section, the training set of 66 analogues identified in Section 7.2 was used to develop predictive QSAR models. This training set is complementary to the test set (TS3) that was identified based upon the structural dissimilarity between the analogues in the data set. That the structural dissimilarity was encoded using E-state indices and yet identified a representative test set of analogues, provides further confidence in the suitability of the resulting training set for modeling purposes.

Table 7.3 All Possible Subsets Regression^{a,b}: Test Set Residuals

<i>Analogue</i>	<i>Actual</i>	<i>MPN</i>		<i>MPP</i>	
		<i>Predicted</i>	<i>Residual</i>	<i>Predicted</i>	<i>Residual</i>
11	6.18	6.05	0.14	6.03	0.16
26	5.33	-9.28 *	14.61 *	-9.03 *	14.36 *
27	7.18	-11.22 *	18.40 *	-11.04 *	18.22 *
30	6.35	7.39	-1.04	7.13	-0.78
32	6.21	6.43	-0.23	6.55	-0.35
33	5.92	6.87	-0.96	6.75	-0.83
34	6.31	7.84	-1.53	7.62	-1.31
41	6.79	6.53	0.26	6.44	0.35
56	5.85	4.52	1.33	4.41	1.44
57	7.39	5.56	1.83	5.72	1.67
62	8.77	7.91	0.86	8.04	0.72
67	7.75	6.77	0.98	6.86	0.89
71	8.38	7.90	0.48	8.00	0.38
80	6.21	8.60	-2.39	8.18	-1.97
^a Training set: 66 analogues complementary to those in TS3; Descriptors: raw, unscaled E-state indices for the 14 scaffold atoms. ^b All numbers are indicated as negative log of IC ₅₀ . * Not meaningful.					

7.4 Partial Least Squares Analyses

Various PLS models were examined using either 2-D or 3-D descriptors individually and in combination. The models are listed in Table 7.4 that shows the name of a model and which descriptors were used for it. The models suffixed with “all” were developed using all 80 analogues in a data set, while those suffixed with “trn” were developed using the training set of 66 analogues that was identified in Section 7.2. The neutral models are listed on the left in Table 7.4 and the protonated models, which follow the same protocol, are listed on the right.

Table 7.4 Description of PLS Models

<i>Neutral Models</i>	<i>Descriptor Set</i>	<i>Number of Analogues</i>	<i>Analysis Type</i>	<i>Protonated Models</i>
mpn_c1_all	CoMFA	80	CoMFA	mpp_c1_all
mpn_z1_all	Molconn-Z	80	2D	mpp_z1_all
mpn_z2_all	CoMFA + Molconn-Z	80	CoMFA + 2D	mpp_z2_all
mpn_e1_all	MOE	80	2D	mpp_e1_all
mpn_e2_all	MOE	80	3D	mpp_e2_all
mpn_e3_all	MOE	80	2D + 3D	mpp_e3_all
mpn_e4_all	CoMFA + MOE	80	CoMFA + 2D	mpp_e4_all
mpn_e5_all	CoMFA + MOE	80	CoMFA + 3D	mpp_e5_all
mpn_e6_all	CoMFA + MOE	80	CoMFA + 2D + 3D	mpp_e6_all
<i>Neutral Models</i>	<i>Descriptor Set</i>	<i>Number of Analogues</i>	<i>Analysis Type</i>	<i>Protonated Models</i>
mpn_c1_trn	CoMFA	66	CoMFA	mpp_c1_trn
mpn_z1_trn	Molconn-Z	66	2D	mpp_z1_trn
mpn_z2_trn	CoMFA + Molconn-Z	66	CoMFA + 2D	mpp_z2_trn
mpn_e1_trn	MOE	66	2D	mpp_e1_trn
mpn_e2_trn	MOE	66	3D	mpp_e2_trn
mpn_e3_trn	MOE	66	2D + 3D	mpp_e3_trn
mpn_e4_trn	CoMFA + MOE	66	CoMFA + 2D	mpp_e4_trn
mpn_e5_trn	CoMFA + MOE	66	CoMFA + 3D	mpp_e5_trn
mpn_e6_trn	CoMFA + MOE	66	CoMFA + 2D + 3D	mpp_e6_trn

The results for the PLS models for the MPN data set are given in Table 7.5 and for the MPP data set in Table 7.6. In these two tables, q^2 is the crossvalidated r^2 statistic, SDEP is the standard error of prediction after leave-one-out crossvalidation, and the optimal number of components is listed under “Components”. The non-crossvalidated models were obtained for the number of components specified and, for each model, the r^2 statistic, the standard error of estimate (SEE), and the F-value are listed. The p-values for all non-validated runs are not shown in the tables because they were always very small (0.000 to three decimal places). Thus, for all these models, the probability of ($r^2 = 0$) is zero (or very small), signifying the existence of a relationship between the descriptors used in a model and the binding affinity. Also listed in Tables 7.5 and 7.6 are the three

statistics produced by progressive scrambling of the models (see Section 6.5.1). These three statistics will be discussed in the next section on model validation.

Table 7.5 Results of PLS Analyses for the MPN Data Set

<i>model</i>	q^2	<i>SDEP</i>	<i>Components</i>	r^2	<i>SEE</i>	<i>F-value</i>	Q^2	<i>cSDEP</i>	$dq^2/dr^2_{yy'}$
mpn_c1_all	0.412	0.729	6	0.865	0.349	77.841	0.176	0.862	0.782
mpn_z1_all	0.295	0.793	5	0.700	0.517	34.482	0.216	0.835	0.674
mpn_z2_all	0.351	0.746	2	0.551	0.620	47.316	0.182	0.837	0.326
mpn_e1_all	0.289	0.781	2	0.402	0.715	25.924	0.185	0.835	0.300
mpn_e2_all	0.246	0.804	2	0.375	0.731	23.138	0.135	0.860	0.346
mpn_e3_all	0.290	0.780	2	0.420	0.705	27.828	0.182	0.837	0.311
mpn_e4_all	0.302	0.773	2	0.405	0.714	26.235	0.187	0.834	0.302
mpn_e5_all	0.259	0.797	2	0.385	0.726	24.115	0.141	0.857	0.353
mpn_e6_all	0.299	0.775	2	0.422	0.704	28.072	0.183	0.836	0.312
<i>model</i>	q^2	<i>SDEP</i>	<i>Components</i>	r^2	<i>SEE</i>	<i>F-value</i>	Q^2	<i>cSDEP</i>	$dq^2/dr^2_{yy'}$
mpn_c1_trn	0.556	0.631	6	0.924	0.261	119.396	0.407	0.727	0.617
mpn_z1_trn	0.287	0.792	5	0.716	0.500	30.289	0.167	0.856	0.637
mpn_z2_trn	0.302	0.784	5	0.720	0.497	30.855	0.170	0.854	0.648
mpn_e1_trn	0.140	0.856	3	0.398	0.716	13.657	0.093	0.879	0.161
mpn_e2_trn	0.131	0.860	3	0.447	0.686	16.710	0.083	0.884	0.494
mpn_e3_trn	0.152	0.844	2	0.362	0.731	17.905	0.108	0.864	0.165
mpn_e4_trn	0.157	0.848	3	0.401	0.714	13.840	0.095	0.878	0.166
mpn_e5_trn	0.173	0.839	3	0.456	0.681	17.346	0.087	0.882	0.506
mpn_e6_trn	0.167	0.836	2	0.364	0.730	18.012	0.109	0.864	0.167

From Table 7.5, for neutral MP analogues, the best models are the CoMFA models derived using the full data set (mpn_c1_all) and the training set (mpn_c1_trn). Of these two, the training set CoMFA model is better: it has the highest q^2 , r^2 , and F-value obtained for any neutral MP model, with the lowest associated errors of prediction and estimation. This model explains 92.4 % of the variance in the training data set. It has a high F-value implying that there is a good chance that a relationship exists between the CoMFA predictors (steric and electrostatic field values) and bioactivity measured as pIC₅₀. The CoMFA model obtained using the full data set has slightly poorer statistics

but is comparable to that obtained using the training set. The same trend is true for the protonated MP data sets in Table 7.6. The best protonated model was mpp_c1_trn, which explained 91.8 % of the original variance and had the highest q^2 .

Table 7.6 Results of PLS Analyses for the MPP Data Set

<i>model</i>	q^2	<i>SDEP</i>	<i>Components</i>	r^2	<i>SEE</i>	<i>F-value</i>	Q^2	<i>cSDEP</i>	$dq^2/dr^2_{yy'}$
mpp_c1_all	0.525	0.655	6	0.888	0.319	96.014	0.411	0.728	1.072
mpp_z1_all	0.167	0.862	5	0.703	0.514	35.053	0.168	0.860	0.515
mpp_z2_all	0.325	0.761	2	0.522	0.640	42.025	0.190	0.833	0.346
mpp_e1_all	0.335	0.755	2	0.433	0.697	29.388	0.204	0.825	0.342
mpp_e2_all	0.363	0.739	2	0.445	0.689	30.869	0.238	0.807	0.494
mpp_e3_all	0.382	0.727	2	0.486	0.664	36.359	0.240	0.807	0.419
mpp_e4_all	0.337	0.753	2	0.434	0.696	29.531	0.205	0.825	0.342
mpp_e5_all	0.365	0.738	2	0.448	0.688	31.227	0.240	0.806	0.494
mpp_e6_all	0.377	0.730	2	0.487	0.663	36.483	0.240	0.806	0.419
<i>model</i>	q^2	<i>SDEP</i>	<i>Components</i>	r^2	<i>SEE</i>	<i>F-value</i>	Q^2	<i>cSDEP</i>	$dq^2/dr^2_{yy'}$
mpp_c1_trn	0.568	0.622	6	0.918	0.271	109.644	0.537	0.644	0.759
mpp_z1_trn	0.288	0.792	5	0.719	0.497	30.757	0.159	0.860	0.626
mpp_z2_trn	0.294	0.788	5	0.721	0.496	30.999	0.161	0.859	0.631
mpp_e1_trn	0.159	0.846	3	0.425	0.700	15.293	0.099	0.876	0.261
mpp_e2_trn	0.156	0.848	3	0.435	0.694	15.922	0.091	0.880	0.503
mpp_e3_trn	0.210	0.814	2	0.397	0.711	20.725	0.152	0.843	0.257
mpp_e4_trn	0.177	0.837	3	0.427	0.699	15.377	0.100	0.875	0.261
mpp_e5_trn	0.171	0.840	3	0.439	0.691	16.188	0.093	0.879	0.507
mpp_e6_trn	0.217	0.811	2	0.397	0.711	20.770	0.152	0.843	0.256

In general, the training set models had slightly poorer crossvalidated statistics than the full data set models but had comparable non-crossvalidated statistics. The difference in their crossvalidated statistics is because the full data set models have a larger structure and activity space available for deriving the QSAR equation. Thus, in the case of the full data set models, there is a greater chance of redundancy so that during leave-one-out crossvalidation, the model will accumulate lower errors of prediction. In other words, the PRESS value will be reduced. Then, by equation 6.6, a lower PRESS

value will lead to a higher q^2 for these full data set models. The similar r^2 values, on the other hand, suggest that the training set models explain as much variance as the full data set models. This is an indication of the adequacy of this training set, which was identified in Section 7.2 using E-state indices for encoding structural dissimilarity, for the derivation of the QSAR models developed here.

For both the neutral and protonated data sets, the models that used Molconn-Z descriptors had significantly higher r^2 values than those for the models that used MOE descriptors. This suggests that the Molconn-Z descriptors were better able to explain the variance in these models. This would indicate that for the analyses in this work, the Molconn-Z descriptor set was more suitable or adequate than the MOE descriptor set.

The introduction of CoMFA descriptors into the full data set models that used the Molconn-Z 2-D descriptors only, raised the q^2 somewhat (from 0.295 for mpn_z1_all to 0.351 for mpn_z2_all and from 0.167 for mpp_z1_all to 0.325 for mpp_z2_all) but lowered the corresponding r^2 significantly (from 0.700 to 0.551 and from 0.703 to 0.522, respectively). A different trend is observed for the corresponding training set models: both q^2 and r^2 remained nearly unchanged. This suggests that the combination of CoMFA steric and electrostatic descriptors and Molconn-Z 2-D descriptors decreased the explanatory power of the full data set models but had no such effect on the training set models. In other words, this provides another indication of the adequacy of the training set of analogues used for model development.

No significant differences were observed in the statistics of models derived from various combinations of MOE descriptors. The results were similar also for corresponding models for the neutral and protonated data sets that used these descriptors.

7.5 Model Validation

The models that were constructed using the full data set of 80 analogues (that is, the models with the suffix “all”) were validated using only progressive scrambling (see Section 6.5.1). The training set models (that is, the models with the suffix “trn”) were validated using progressive scrambling as well as test set activity prediction.

7.5.1 Progressive Scrambling Results

The results of progressive scrambling are listed in Table 7.5 for the neutral MP data set models and Table 7.6 for the protonated MP data set models. The three progressive scrambling statistics are described in Section 6.5.1. A stable or robust model is one with high Q^2 , low cSDEP, and a slope near 1.0. Such a model will be affected in proportion to the magnitude of the change in the underlying activity values. The Q^2 and cSDEP statistics tend to be inversely related. The Q^2 value is a conservative statistic and will usually be low because it is based on randomized or noisy activity values. In Table 7.5, the CoMFA models mpn_c1_all and mpn_c1_trn can be compared for stability. For mpn_c1_all, the Q^2 is low at 0.176 but its slope is the closest to 1.0 ($dq^2/dr_{yy}^2 = 0.782$) amongst all neutral models. This model may not be stable because of its very low Q^2 ; however, the somewhat more important slope statistic suggests otherwise. The mpn_c1_trn model is stable because its Q^2 is large (0.407) and its slope is fairly good (0.617).

The CoMFA models for the protonated data set in Table 7.6 are both robust. With Q^2 of 0.411 and slope almost 1.0, the mpp_c1_all model is the most stable of all

PLS models calculated. The model mpp_c1_trn has the highest Q^2 value of all models and a slope very near 1.0 ($dq^2/dr^2_{yy} = 0.759$).

For other models, in general those based on the full data set were more stable than those based on the training set. The protonated MP data set models were more stable than the corresponding neutral MP data set models. Models that included Molconn-Z descriptors were more stable than those that included MOE descriptors.

7.5.2 External Validation

The test set (TS3) used for activity predictions was developed in Section 7.2. From each of the neutral and the protonated MP data sets, the model with the best crossvalidated and explanatory statistics was selected for external validation. For each data set, this model was the one that had the best crossvalidated and explanatory statistics. Thus, models selected for validation were mpn_c1_trn and mpp_c1_trn, that is, the CoMFA models based on the training set. The prediction was carried out in SYBYL and the predicted activity values as well as the residuals for the two models are listed in Table 7.7. As shown in the table, for the mpn_c1_trn model, all but two analogues (**26** and **32**) were predicted with residuals less than 2 units. For the mpp_c1_trn model, the residuals were even better behaved with just one analogue (**34**) having a residual greater than 2 units. This is an improvement over the predictions of the preliminary all possible regression model that did not give meaningful results for analogues **26** and **27**.

Table 7.7 Test Set (TS3) Activity Prediction^a

<i>Analogue</i>	<i>Actual</i>	<i>mpn_c1_trn</i>		<i>mpp_c1_trn</i>	
		<i>Predicted</i>	<i>Residual</i>	<i>Predicted</i>	<i>Residual</i>
11	6.18	6.50	-0.32	5.95	0.23
26	5.33	7.45	-2.13	7.02	-1.69
27	7.18	8.21	-1.02	7.89	-0.71
30	6.35	6.59	-0.24	6.62	-0.28
32	6.21	4.01	2.20	6.58	-0.38
33	5.92	6.69	-0.78	6.72	-0.80
34	6.31	7.83	-1.52	8.36	-2.05
41	6.79	7.16	-0.36	6.66	0.14
56	5.85	6.21	-0.36	6.18	-0.33
57	7.39	7.08	0.31	6.73	0.66
62	8.77	8.54	0.23	8.69	0.08
67	7.75	7.71	0.04	7.25	0.50
71	8.38	9.17	-0.79	8.32	0.06
80	6.21	7.71	-1.50	7.27	-1.05

^a All numbers are indicated as negative log of IC₅₀.

7.6 Data Interpretation and Predictions

Contour maps were calculated for the best neutral and protonated CoMFA models in Tables 7.5 (mpn_c1_trn) and 7.6 (mpp_c1_trn), respectively. The maps (Figure 7.1) are a way of visualizing the relative differences in the steric and electrostatic energies of the 66 analogues in the training set. Although the maps were calculated from the training set analogues, they are displayed with some of the TS3 test set analogues as an aid in interpreting the poor residuals of some of those analogues (Table 7.7). For the mpn_c1_trn model, the maps shown in Figure 7.1 include all 14 test set analogues. For the mpp_c1_trn model, only two test set analogues (**11** and **32**) are displayed for clarity.

For each model, the steric contour maps are displayed on the left while the electrostatic contour maps are displayed on the right. In the steric maps, the green regions enclose volumes within which addition of bulkier groups would lead to a better

binding affinity value. On the other hand, the yellow regions enclose volumes within which a reduction in steric bulk would produce a better binding affinity value. Similarly, in the electrostatic maps, the blue and red regions enclose volumes within which more positive charge and more negative charge, respectively, would give a better binding affinity value.

The maps for the two models are qualitatively similar except that for the protonated model, there is a preponderance of blue regions around the piperidinyl nitrogen indicating that more positive charge in these regions would give a better binding affinity value. The maps may be understood by using analogue **79** (3, 4-benzo MP, which has a benzene ring fused at the 3- and 4-positions of the phenyl ring of MP) that has a good binding affinity value (11 nM, see Table B.1). In this analogue, the bulky 3, 4-benzo group extends into the favorable green region off the 3- and 4-positions on the phenyl ring. In addition, in the corresponding electrostatic maps, the pi-electron cloud above and below the plane of the aromatic moiety of this group lies in the red regions (where higher negative charge is correlated with better binding affinity). The presence of sterically unfavorable yellow regions off the plane of the phenyl ring in MP, as for analogue **40** (4-*t*-butyl MP, with a poor IC₅₀ value of 13,450 nM), further constrains the choice of phenyl ring substituents when making predictions for a new compound based on the results in this and previous sections. Thus, predicted substitutions at the phenyl ring of MP that are favorable for binding affinity must have bulky, electron-rich atoms or groups at the 3- and 4-positions.

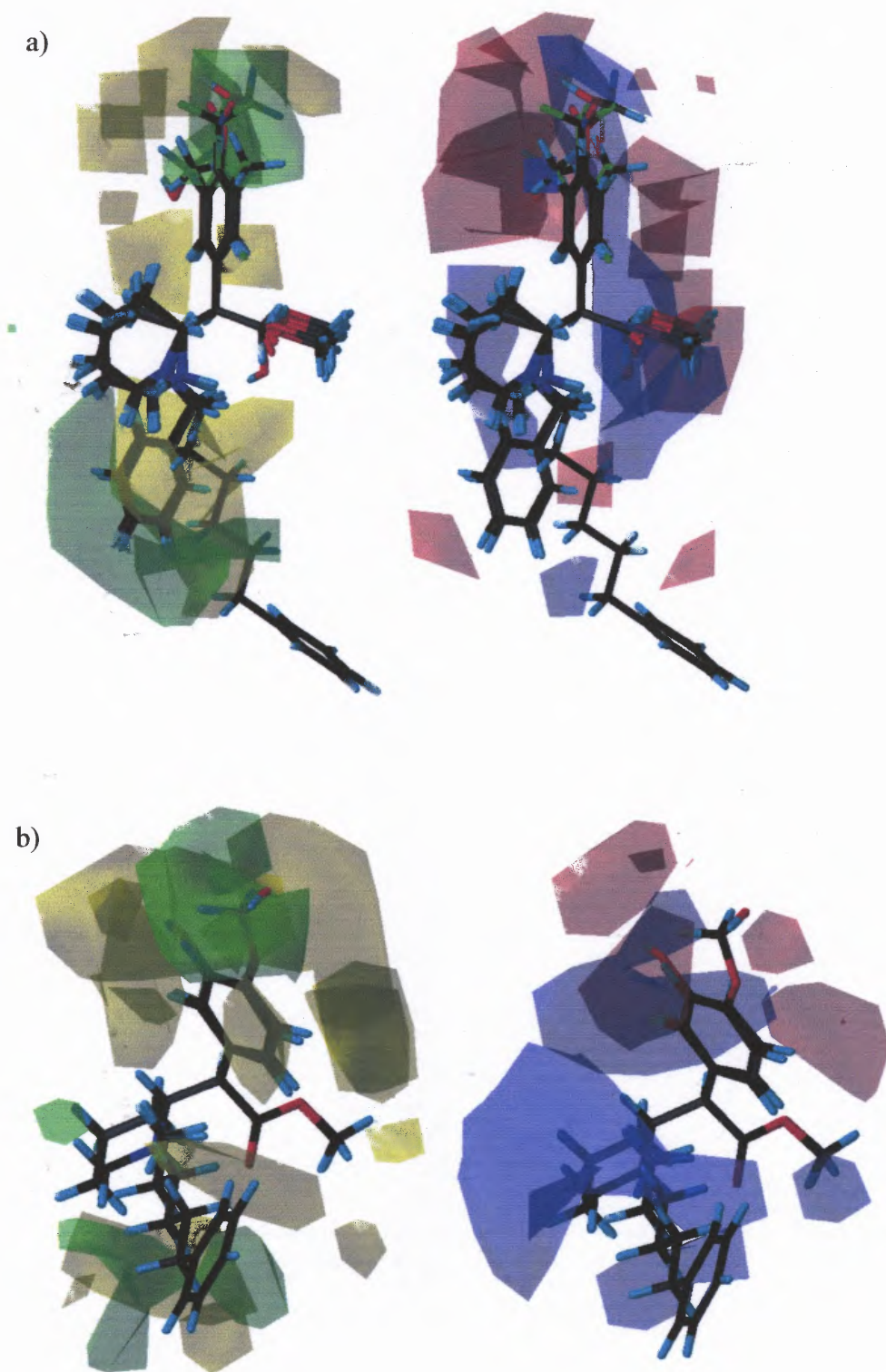


Figure 7.1 CoMFA maps for (a) mpn_c1_trn and (b) mpp_c1_trn models.

One compound that can be predicted in this way is 3, 4-dibromo MP. It has the relatively bulky, electron-rich bromine atoms at the positions associated with better binding affinity. Note that the related analogue **52** (3, 4-dichloro MP) has a very good binding affinity value (5.3 nM, see Table B.1). Yet, analogue **53** (3, 4-dimethoxy) has a relatively poorer affinity (810 nM). Since the methoxy groups are free to rotate they can occupy the unfavorable yellow regions off the plane of the phenyl ring (as for analogue **40** above). This indicates that the bulky, electron-rich atoms or groups at the 3- and 4-positions of the phenyl ring must be restricted from extending out of the plane of the phenyl ring, as in the case of the 3, 4-benzo MP analogue. Indeed, further analogues that could be explored are those that have substitutions (such as halogen or methoxy substitutions) on the 3, 4-benzo MP analogue.

For substitutions at the piperidinyl nitrogen, the steric maps indicate that a substituent with a longer chain may be a better DAT-binding ligand than one with a shorter chain (such as benzyl group, $-\text{CH}_2\text{Ph}$). This is indicated by the presence of sterically unfavorable yellow regions closer to the piperidine ring and sterically favorable green regions away from the ring. This interpretation could be disputed by giving the example of analogues like **70** and **71** that have very good binding affinity values (2.7 nM and 4.2 nM, respectively, see Table B.1). However, it should be noted that in such analogues there are also substitutions at the 3- and 4-positions of the phenyl ring that are favorable as discussed above. For example, both **70** and **71** have bulky, electron-rich chlorine atoms at the 3- and 4- positions.

From the results in this and previous sections, no conclusive claims can be made about the effect of the side chain variations on the binding affinity of the set of DAT-binding ligands used in this study.

From the above results and from the trends in binding affinity values listed Table B.1, it appears that the main factor in improving binding affinity values is substitution at phenyl ring (R1 in Figure 6.1), especially at the 3- and 4-positions. For example, all analogues that have the same 3,4-dichloro substitution at R1 (e.g., analogues **52**, **61**, **62**, and **72**) but a different substituent at R3 ($-\text{CO}_2\text{CH}_3$ in **52**, $-\text{COH}$ in **61**, $-\text{COCH}_3$ in **62**, and $-\text{CONH}_2$ in **72**), have very good binding affinity values (5.3 nM, 4.2 nM, 1.7 nM, and 16.4 nM, respectively). However, the corresponding analogues without a substitution at R1 but with the above substituents at R3, i.e., analogues **39** (83 nM), **30** (447.5 nM), **68** (97.1 nM), and **76** (1728 nM), all have relatively poorer binding affinity values.

7.7 Discussion

The results in this chapter indicate that, when applied to the MP analogues, the three-stage process of developing a useful QSAR model (Section 6.1) succeeded in rationally and incrementally identifying useful models. The presence of the 14-atom scaffold in these analogues led to the idea of using E-state indices to characterize the substitution pattern on the scaffold. The E-state indices were successfully tested for adequacy in encoding structural information about the analogues in the data set. The robust CoMFA models developed in this study permit useful predictions to be made for the development of new related compounds that may be better DAT-binding ligands.

Preliminary models with E-state indices alone yielded encouraging results and were able to identify the three substitution sites on the scaffold. The principal component-based forward stepwise regression models for the neutral and protonated data sets had reasonable differences in their characterization of the substitution pattern on the scaffold. While the stepwise regression models for the neutral data set identified the phenyl ring and the nitrogen atom as the most important, those for the protonated models identified the phenyl ring and the side chain as such. It would seem that this difference could be due to the fact that the additional hydrogen on the piperidinyll nitrogen of *all* the protonated analogues permits hydrogen-bonding with an oxygen atom in the side chain. This would make the side chain as well as the nitrogen atom important for all protonated analogues. Such hydrogen-bonding would not be observed in the case of those neutral analogues that have a substituent on the nitrogen atom. Thus, for the overall neutral data set, the side chain would not appear to be as important as the nitrogen atom. While Table 7.1 suggests such a reasoning, it should be noted that the E-state indices are based upon molecular connectivity only and *not* upon 3-D conformation. Though the 3-D conformation for the protonated MP analogues differs from that of the neutral analogues due to hydrogen-bonding in the former case, this difference in conformation would not be reflected in the E-state index. The only structural difference between the neutral and protonated MP analogues that would be considered in the calculation of the E-state indices is the presence of the additional hydrogen on the piperidinyll nitrogen atom. Thus, unless the E-state indices are capable of encoding the aforementioned hydrogen-bonding in an indirect way, the justification of the differences based on hydrogen-bonding may not be tenable. However, since the E-state index for an atom encodes

“electrotopological” information, meaning both electronic and topological structure information from *all other atoms* in the molecule,¹³³ the E-state indices calculated here might be able to account for hydrogen-bonding. It should be noted here, though, that there is one model for the protonated data set, MPP-PS (principal component- and standard deviation normalization-based), which identifies all three sites of substitution as important and another, MPP-PR (range scaling-based), which identifies neither the nitrogen nor the side chain.

The results of the forward stepwise regression models give a general idea about the utility of the E-state indices. Taken collectively, these models identified those atoms that are directly involved in substitutions. Even atom 7 that is identified in the models for the protonated analogues is directly connected to atom 14 in the side chain (see Figure 6.1). Thus, these results from exploratory models were taken to indicate the adequacy of the E-state indices for encoding structural information about the MP analogues.

The use of E-state indices as in the present work has some other advantages. The direct use of E-state indices in the models in Table 7.1 indicates an important feature of these descriptors. The direct correspondence between an E-state index and the atom for which it is calculated means that E-state indices can be easily and directly interpreted. The use of these indices in the regression analyses of Sections 7.1 and 7.3 shows that since the “units” of these indices are the same, additional scaling considerations may not be required during model development. Furthermore, since the number of atoms that comprise a scaffold in a given series of analogues would tend to remain small, the dimensionality of input matrices that use E-state indices for scaffold atoms only would also be low. This may offer many advantages in QSAR model development where

dimensionality reduction is frequently a primary concern. In addition, as shown in this work, once a suitable test set has been identified using E-state indices for select atoms of the scaffold, the test set can be used for validating other QSAR models developed for the series of analogues. For example, in this work the same test set was used to validate the models that were developed using other 2-D and even 3-D descriptors. Indeed, these models used descriptor sets that were calculated from different software packages but it was possible to use the same test set for validation purposes. It should be noted that in the case of a series of analogues without a common scaffold, it would not be possible to use E-state indices the way they are used in the present work.

The test set identification was based upon the selection of the most orthogonal E-state indices. The structural information encoded by the E-state indices allowed the identification of a truly representative subset of analogues that span not only the structure space but also the activity space. Since the particular test set identification technique used in this work is deterministic, that is, for a given set of input parameters, it identifies the same test set for a given series of analogues, it may eliminate the need for generating a whole collection of test and training sets and repeating model development on each different combination. Yet, the input parameters for calculating a test set, such as the threshold radius, t , and the particular DBCS algorithm used, may be varied to suit the modeler's preferences. The program D-SIM version 1.0 that was developed in this work (see Appendix G), allows the user to vary the value of t and offers a choice of the three sphere exclusion algorithms mentioned in Section 6.3.2.2. The program output includes a log file with a record of the order in which the analogues are selected in the test set. This information can give some idea about the distribution of the analogues as points within

the descriptor space. Thus, analogues that are selected in succession would be closer to each other in the descriptor space. Also, analogues that are selected earlier would lie toward the center of the descriptor space. As an example, in this work, the order of selection for TS3 was **11, 80, 67, 41, 56, 32, 71, 27, 33, 57, 62, 30, 26, and 34**. It could be imagined that, as it nears completion, the sphere exclusion algorithm would select the analogues located near the extremes of the descriptor space (See Figure 6.3 for the schematic representation). If the descriptor space is sparse (due to the series of analogues being “incomplete”), it is likely that the analogues that are included late in the test set might be outliers. Note that analogue **34** in TS3 was the last to be selected and had the largest residual from the best overall PLS model (mpp_c1_trn). Analogue **26**, which was selected just before analogue **34**, also had consistently large residuals.

The results of the PLS analyses in Tables 7.5 and 7.6 suggest a general trend for the crossvalidated q^2 to remain low. This is especially true for descriptors other than the CoMFA descriptors. While this suggests that these PLS models were not predictive, it should not be taken to mean that the descriptor sets are not useful. One way to improve the results would be to apply thorough and selective scaling of the non-CoMFA descriptors used. It should be noted that the CoMFA steric and electrostatic descriptors are automatically scaled in SYBYL using a scaling method that is appropriate for such descriptors. However, the non-CoMFA descriptors used in this work were simply autoscaled. Since several descriptors are related to each other, block scaling techniques could be used that would scale several descriptors together.

Another important way to improve the results would be to use a judicious descriptor selection (variable selection) scheme during the model development process.

This is the most important aspect of modeling that is missing from the current work. It includes an examination of each descriptor (or each subset of descriptors) and making considered judgments about its utility in modeling. In this way, only those descriptors that may be the most meaningful in the context of the modeling would be included in any analysis and hence would significantly reduce noise in the underlying data. If the concern is that the descriptor space might be inadequate then descriptor sets other than those from Molconn-Z and MOE could also be considered.

Yet another way to improve the results would be to consider techniques other than PLS. Since PLS assumes linearity between the descriptors and the activity values, it may not be the most suitable choice for modeling. Other linear or non-linear techniques, such as genetic algorithms and neural networks, could be explored in developing even better QSAR models.

CHAPTER 8

CONCLUSIONS

The absence of information about the DAT structure forces important constraints upon the scope of modeling that can be performed on DAT binding ligands such as the GBR 12909 analogues and the MP analogues. In this work, some of these constraints were highlighted in two sets of studies. In the first set of studies, the selection of a few representative structures as putative binding conformations from a large collection of conformations of a flexible GBR 12909 analogue was demonstrated by cluster analysis. Novel structure-based features were identified for the analogue and used for clustering. These features were shown to be useful in this work and are easily generalizable to other molecules. Since the feature space may or may not be Euclidean, a recently-developed fuzzy relational clustering algorithm capable of handling such data was used. Both superposition-dependent and superposition-independent features were used along with region-specific clustering that focused on separate pharmacophore elements in the molecule. Separate sets of representative structures were successfully identified for the superposition-dependent and superposition-independent analyses. The cluster analyses carried out in this work are thus a useful way of analyzing the conformations of flexible molecules used in ligand-based drug design.

The second set of studies included the development of QSAR models for the MP data set of 80 compounds. The E-state indices for the 14 atoms of the scaffold common to all 80 compounds were successfully used in a novel way to develop an effective test set that spanned both the structure space as well as the activity space. The utility of E-state indices in modeling a series of analogues with a common scaffold was

demonstrated. Several models were developed using various combinations of 2-D and 3-D descriptors. In terms of the predictive and explanatory capability and stability, the best models were those that were derived using CoMFA steric and electrostatic descriptors. Validation of all models by using progressive scrambling indicated several stable CoMFA models. External validation of these models by predicting the activity of test set analogues produced reasonable residuals. Further improvements in all models could be expected by using judicious scaling and variable selection techniques. The results of this work permit predictions of new compounds in the series of MP analogues, which may be better DAT-binding ligands. Substitutions in the phenyl ring of MP, especially at the 3- and 4-positions, were found to be the most important. It was found that for better DAT-binding the substituents at these positions should be relatively bulky, electron-rich atoms or groups. New compounds such as 3,4-dibromo MP or analogues of 3,4-benzo MP are predicted to bind well to the DAT.

APPENDIX A

ADDITIONAL RESULTS FOR SUPERPOSITION-INDEPENDENT CLUSTERING

These additional plots are provided in the order of the analysis performed. Thus, results for the [N_xP₂] and [N_xP₁] (full-molecule) analyses are first, followed by results for the [C_xP₂] and [C_xP₁] (B-side) analyses, with results for the [N_xC] (A-side) analyses last.

[N_xP₂]_{T+R}:

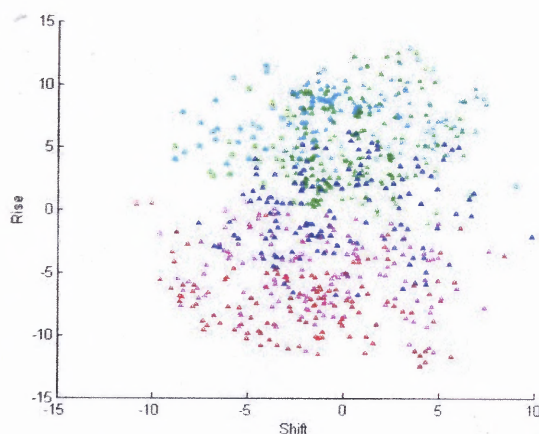


Figure A.1 (Shift, Rise) space for [N_xP₂]_{T+R}, $c = 5$.

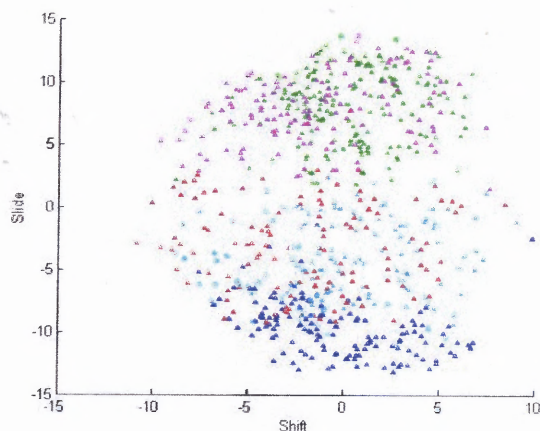


Figure A.2 (Shift, Slide) space for [N_xP₂]_{T+R}, $c = 5$.

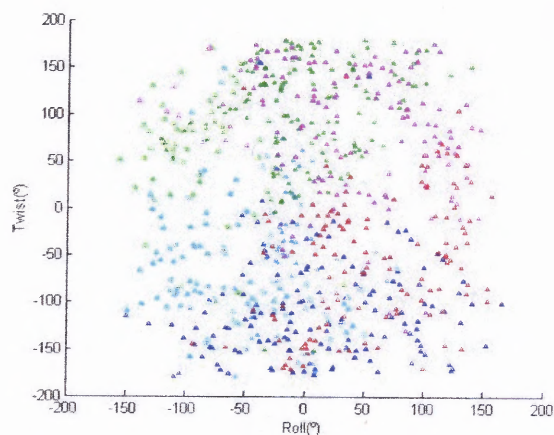


Figure A.3 (Roll, Twist) space for $[NxP2]_{T+R}$, $c = 5$.

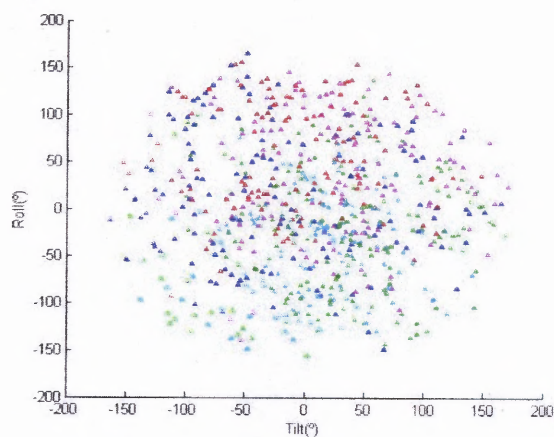


Figure A.4 (Tilt, Roll) space for $[NxP2]_{T+R}$, $c = 5$.

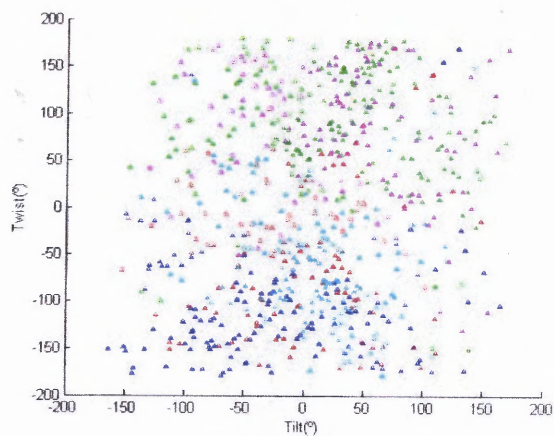


Figure A.5 (Tilt, Twist) space for $[NxP2]_{T+R}$, $c = 5$.

$[N \times P1]_{T+R}$:

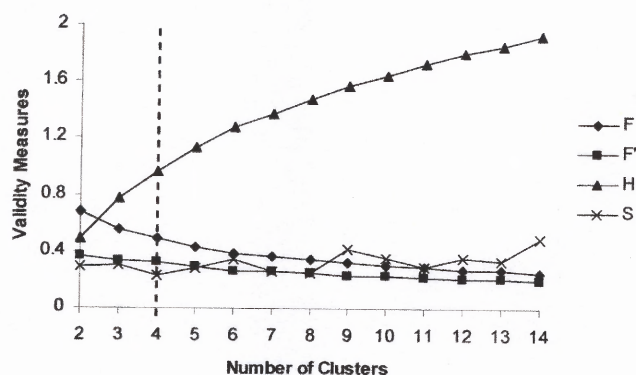


Figure A.6 Cluster validity plots for the $[N \times P1]_{T+R}$ proximity matrix.

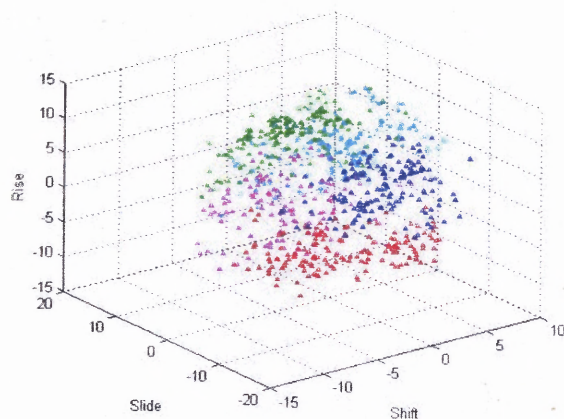


Figure A.7 (Slide, Shift, Rise) space for $[N \times P1]_{T+R}$, $c = 5$.

$[N \times P1]_T$:

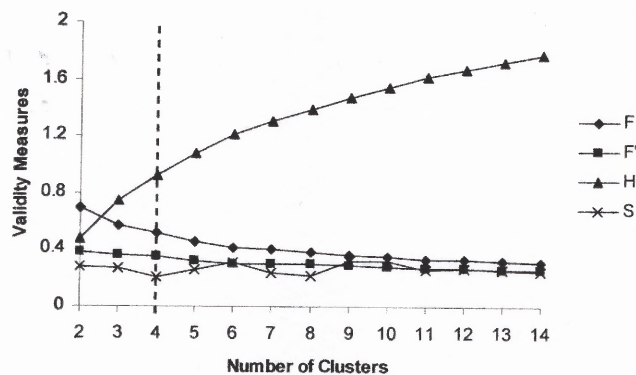


Figure A.8 Cluster validity plots for the $[N \times P1]_T$ proximity matrix.

[C_xP₂]_{T+R}:

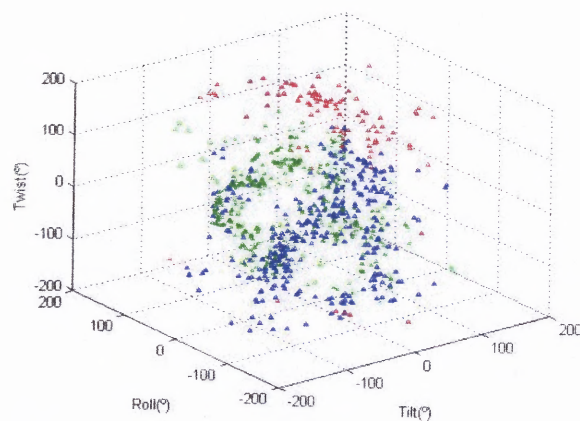


Figure A.9 (Roll, Tilt, Twist) space for [C_xP₂]_{T+R}, $c = 3$.

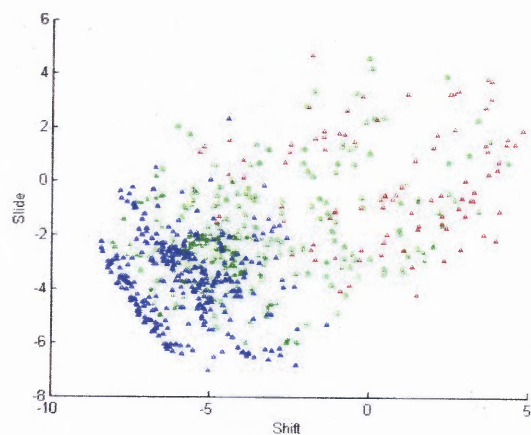


Figure A.10 (Shift, Slide) space for [C_xP₂]_{T+R}, $c = 3$.

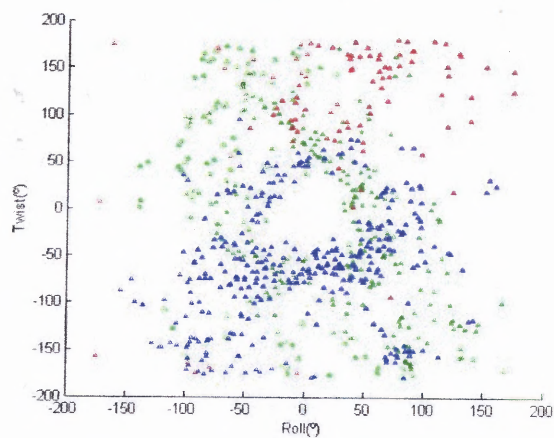


Figure A.11 (Roll, Twist) space for [C_xP₂]_{T+R}, $c = 3$.

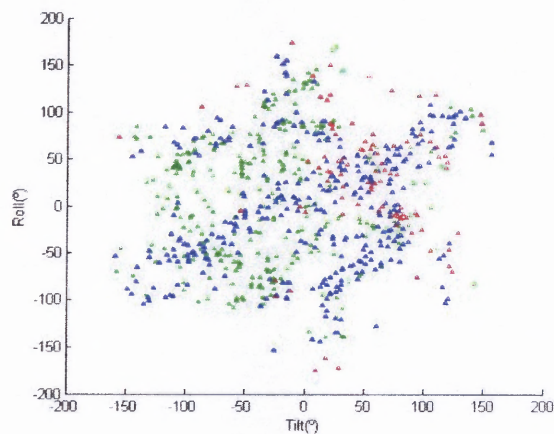


Figure A.12 (Tilt, Roll) space for $[CxP2]_{T+R}$, $c = 3$.

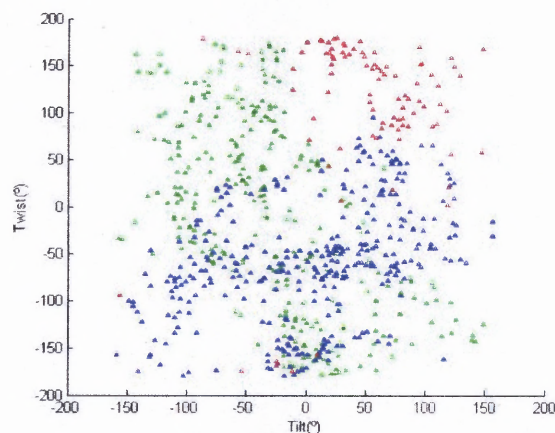


Figure A.13 (Tilt, Twist) space for $[CxP2]_{T+R}$, $c = 3$.

$[CxP1]_{T+R}$:

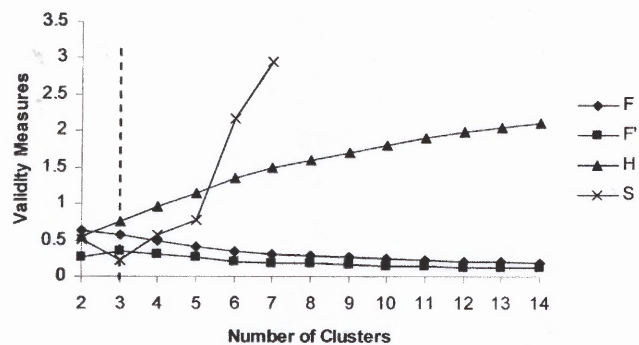


Figure A.14 Cluster validity plots for the $[CxP1]_{T+R}$ proximity matrix.

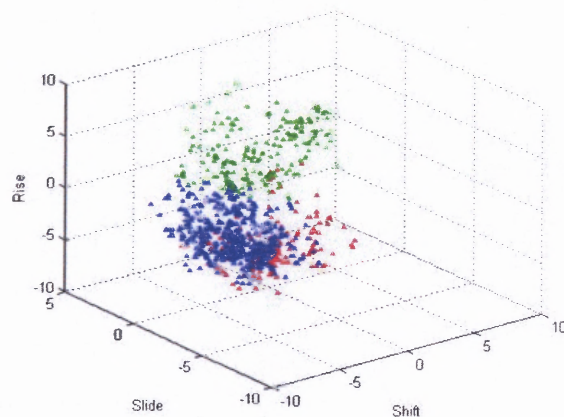


Figure A.15 (Slide, Shift, Rise) space for $[C_{xP1}]_{T+R}$, $c = 3$.

$[C_{xP1}]_T$:

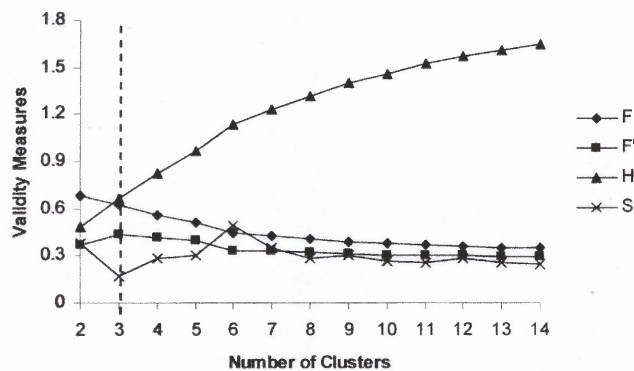


Figure A.16 Cluster validity plots for the $[C_{xP1}]_T$ proximity matrix.

$[N_{xC}]_T$:

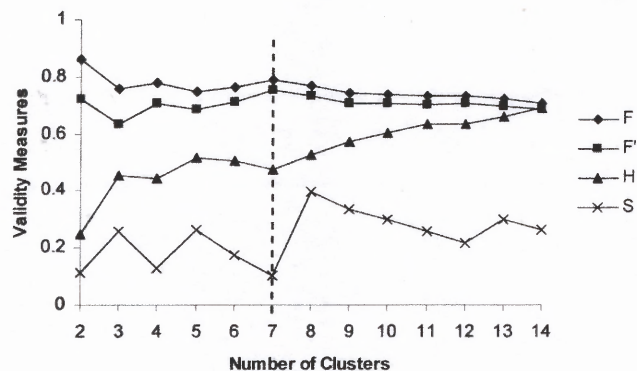


Figure A.17 Cluster validity plots for the $[N_{xC}]_T$ proximity matrix.

$[N \times C]_R$:

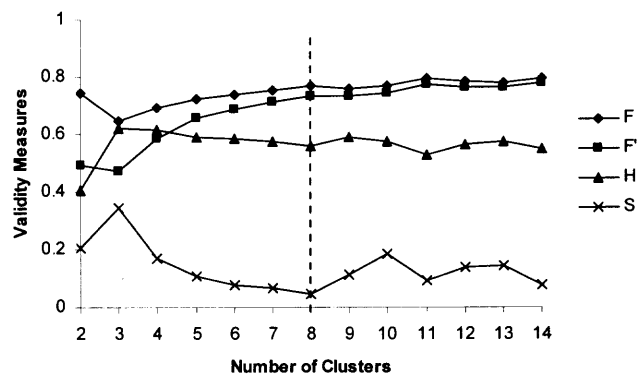


Figure A.18 Cluster validity plots for the $[N \times C]_R$ proximity matrix.

APPENDIX B

MP ANALOGUE DATA SET

The substitutions R1, R2, and R3 in Table B.1 correspond to those in the figure at the top of the table. The bold type letter “C” in the substituent listed in the R3 column corresponds to the 14th atom of the scaffold. The GIT (Georgia Institute of Technology) number is the identifier used by the Deutsch group. The DAT binding affinity is listed in nanomolar units while the pIC₅₀ column indicates the negative log of (molar) IC₅₀. Binding data was provided by Dr. Margaret Schweri, Mercer University Medical School. Methylphenidate is compound **39**. Compounds **42**, **43**, **45**, and **54** were provided by Dr. S. J. Gatley of Brookhaven National Laboratories. The binding affinity values are unpublished unless noted.

Table B.1 MP Analogues Data Set

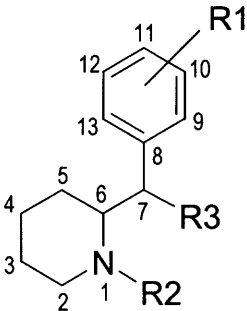
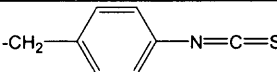
							
	<i>GIT Number</i>	<i>R1^a</i>	<i>R2</i>	<i>R3</i>	<i>IC₅₀ (nM)</i>	<i>pIC₅₀</i>	<i>Group</i>
1	AN-1-68.2	-H		-CO ₂ CH ₃	422.3	6.37	R2
2	BO-1-119.1	-H	-(CH ₂) ₃ Ph	-COH	193.5	6.71	R2+R3
3	BO-1-12.1	-H	-CH ₂ C≡C	-CO ₂ CH ₃	820.5	6.09	R2
4	BO-1-120.1	-H	-(CH ₂) ₄ Ph	-COH	622.5	6.21	R2+R3
5	BO-1-122.1	-H	-(CH ₂) ₂ Ph	-COH	1431	5.84	R2+R3

Table B.1 MP Analogues Data Set (Continued)

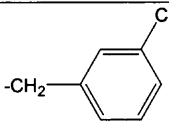
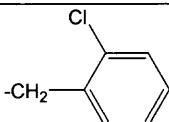
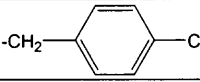
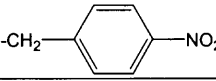
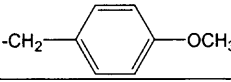
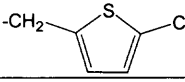
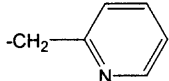
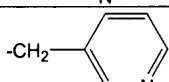
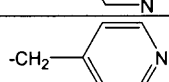
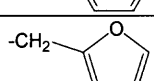
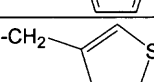
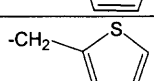
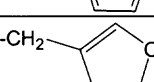
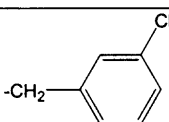
6	BO-1-128.1	-H	$-(\text{CH}_2)_3\text{Ph}$	$-\text{CO}_2\text{CH}_3$	267	6.57	R2
7	BO-1-13.1	-H		$-\text{CO}_2\text{CH}_3$	105.9	6.98	R2
8	BO-1-131.1	-H	$-(\text{CH}_2)_2\text{Ph}$	$-\text{CO}_2\text{CH}_3$	677.5	6.17	R2
9	BO-1-144.1	-H	$-(\text{CH}_2)_4\text{Ph}$	$-\text{CO}_2\text{CH}_3$	205.3	6.69	R2
10	BO-1-145.1	-H	$-(\text{CH}_2)_5\text{Ph}$	$-\text{CO}_2\text{CH}_3$	1572.5	5.80	R2
11	BO-1-146.1	-H	$-(\text{CH}_2)_6\text{Ph}$	$-\text{CO}_2\text{CH}_3$	656	6.18	R2
12	BO-1-15.1	-H		$-\text{CO}_2\text{CH}_3$	242.6	6.62	R2
13	BO-1-17.1	-H	$-\text{CH}_2\text{C}=\text{C}$	$-\text{CO}_2\text{CH}_3$	597.2	6.22	R2
14	BO-1-19.1	-H		$-\text{CO}_2\text{CH}_3$	31.2	7.51	R2
15	BO-1-21.1	-H		$-\text{CO}_2\text{CH}_3$	112.9	6.95	R2
16	BO-1-23.1	-H		$-\text{CO}_2\text{CH}_3$	79.1	7.10	R2
17	BO-1-30.1	-H		$-\text{CO}_2\text{CH}_3$	391.5	6.41	R2
18	BO-1-37.1	-H		$-\text{CO}_2\text{CH}_3$	368.5	6.43	R2
19	BO-1-43.1	-H		$-\text{CO}_2\text{CH}_3$	173.2	6.76	R2
20	BO-1-44.1	-H		$-\text{CO}_2\text{CH}_3$	127.9	6.89	R2
21	BO-1-45.1	-H		$-\text{CO}_2\text{CH}_3$	535.7	6.27	R2
22	BO-1-46.1	-H		$-\text{CO}_2\text{CH}_3$	142.8	6.85	R2
23	BO-1-47.1	-H		$-\text{CO}_2\text{CH}_3$	223.7	6.65	R2
24	BO-1-48.1	-H		$-\text{CO}_2\text{CH}_3$	459.3	6.34	R2
25	BO-1-96	-H	$-\text{CH}_2\text{CH}_3$	$-\text{COH}$	2338	5.63	R2+R3
26	BO-2-28.1	3,5-diCH ₃	-H	$-\text{CO}_2\text{CH}_3$	4685	5.33	R1
27	BO-2-40.1	3,5-diCl	-H	$-\text{CO}_2\text{CH}_3$	65.6	7.18	R1
28	BO-2-57.1	-H		$-\text{COH}$	25.8	7.59	R2+R3
29 ^b	BO3.2	-H	$-\text{CH}_2\text{Ph}$	$-\text{CO}_2\text{CH}_3$	52.9	7.28	R2

Table B.1 MP Analogues Data Set (Continued)

30	CE101.1	-H	-H	-COH	447.5	6.35	R3
31 ^c	EGK-266/1	4-OH	-H	-CO ₂ CH ₃	98	7.01	R1
32	EGK-276-A	3-CH ₂ OH, 4-OCH ₂ OH	-CH ₃	-CO ₂ CH ₃	620	6.21	R1+R2
33	EGK-276-B	4-OH	-CH ₃	-CO ₂ CH ₃	1215	5.92	R1+R2
34 ^c	LL81.2	4-NO ₂	-H	-CO ₂ CH ₃	493.8	6.31	R1
35 ^c	QS-1-114.1	3-NH ₂	-H	-CO ₂ CH ₃	265	6.58	R1
36 ^c	QS-1-128.1	4-NH ₂	-H	-CO ₂ CH ₃	34.5	7.46	R1
37 ^c	QS-1-138.1	4-OCH ₃	-H	-CO ₂ CH ₃	84.3	7.07	R1
38 ^c	QS-1-142.1	4-Cl	-H	-CO ₂ CH ₃	20.6	7.69	R1
39 ^c	QS-1-89.4	-H	-H	-CO ₂ CH ₃	83	7.08	-
40 ^c	QS-2-116.3	4- <i>t</i> -butyl	-H	-CO ₂ CH ₃	13450	4.87	R1
41 ^d	QS-2-124.2	3-Cl	-CH ₃	-CO ₂ CH ₃	160.5	6.79	R1+R2
42 ^c	QS-2-125.1	4-I	-H	-CO ₂ CH ₃	14	7.85	R1
43 ^c	QS-2-125.2	3-Br	-H	-CO ₂ CH ₃	4.2	8.38	R1
44 ^b	QS-2-133.1	4-CH ₃	-CH ₃	-CO ₂ CH ₃	139.8	6.85	R1+R2
45 ^c	QS-2-147.2	2-Br	-H	-CO ₂ CH ₃	1865	5.73	R1
46 ^c	QS-2-15.1	2-OCH ₃	-H	-CO ₂ CH ₃	100666.7	4.00	R1
47 ^c	QS-2-29.4	3-OCH ₃	-H	-CO ₂ CH ₃	287.5	6.54	R1
48 ^c	QS-2-40.1	2-OH	-H	-CO ₂ CH ₃	23050	4.64	R1
49 ^c	QS-2-41.2	3-OH	-H	-CO ₂ CH ₃	321	6.49	R1
50 ^c	QS-2-61.4	3-Cl	-H	-CO ₂ CH ₃	5.1	8.29	R1
51 ^c	QS-2-71.3	4-F	-H	-CO ₂ CH ₃	35	7.46	R1
52 ^c	QS-2-81.4	3,4-diCl	-H	-CO ₂ CH ₃	5.3	8.28	R1
53 ^c	QS-2-84.4	3,4-diOCH ₃	-H	-CO ₂ CH ₃	810	6.09	R1
54 ^c	QS-2-88.1	4-Br	-H	-CO ₂ CH ₃	6.9	8.16	R1
55 ^c	QS-2-99.3	2-Cl	-H	-CO ₂ CH ₃	1946.7	5.71	R1
56 ^c	WB47.4	2-F	-H	-CO ₂ CH ₃	1415	5.85	R1
57 ^c	WB48.4	3-F	-H	-CO ₂ CH ₃	40.5	7.39	R1
58 ^c	WB61.4	4-CH ₃	-H	-CO ₂ CH ₃	33	7.48	R1
59 ^c	WB71.5	3-CH ₃	-H	-CO ₂ CH ₃	21.4	7.67	R1
60	WB77.2	3-F	-H	-COH	281	6.55	R1+R3
61	XY-1-102.3	3,4-diCl	-H	-COH	4.2	8.38	R1+R3
62	XY-1-127.5	3,4-diCl	-H	-COCH ₃	1.7	8.77	R1+R3
63	XY-1-129.2	3-Cl	-CH ₂ Ph	-CO ₂ CH ₃	41.2	7.39	R1+R2
64	XY-1-144.4	-H	-CH ₂ Ph	-CON(CH ₃) ₂	1732.5	5.76	R2+R3
65	XY-1-147.4	-H	-CH ₂ Ph	-CONH ₂	384	6.42	R2+R3
66 ^d	XY-1-30.3	-H	-CH ₂ Ph	-COH	23.7	7.63	R2+R3
67 ^d	XY-1-44.5	-H	-CH ₂ Ph	-COCH ₃	17.8	7.75	R2+R3
68	XY-1-47.1	-H	-H	-COCH ₃	97.1	7.01	R3
69	XY-1-85.7	3,4-diCl	-CH ₂ Ph	CO ₂ CH ₃	76.3	7.12	R1+R2
70	XY-1-86.2	3,4-diCl	-CH ₂ Ph	-COH	2.7	8.57	R1+R2+R3
71	XY-1-89.5	3,4-diCl	-CH ₂ Ph	-COCH ₃	4.2	8.38	R1+R2+R3
72	XY-2-74.3	3,4-diCl	-H	-CONH ₂	16.4	7.79	R1+R3
73	ZL102.3	-H	-H	-CO ₂ CH ₂ Ph	1024.3	5.99	R3

Table B.1 MP Analogues Data Set (Continued)

74	ZL105.1	3-CH ₃	-CH ₃	-CO ₂ CH ₃	107.7	6.97	R1+R2
75	ZL21.1	-H	-H	-CO(CO)CH ₃	690	6.16	R3
76	ZL26.1	-H	-H	-CONH ₂	1728	5.76	R3
77 ^d	ZL32.1	-H	-CH ₃	-CO ₂ CH ₃	499	6.30	R2
78	ZL38.1	4-C ₂ H ₅	-H	-CO ₂ CH ₃	736.7	6.13	R1
79 ^e	ZL68.3	3,4-benzo	-H	-CO ₂ CH ₃	11	7.96	R1
80	ZL77.2	4-CF ₃	-H	-CO ₂ CH ₃	615	6.21	R1

^a Note that the naming convention for the R1 substituents is based upon the six positions of the phenyl ring. For example, analogue **48** (2-OH MP) is MP with an -OH substituent at the 2-position of the phenyl ring (i.e., at positions 9 or 13 in the figure that accompanies this table).

^b Reference 41.

^c Reference 37.

^d Reference 43.

^e Reference 42.

Table B.2 Data Set and Test Set Groups

<i>Set</i>	<i>R1* Analogues</i>	<i>R2* Analogues</i>	<i>R3* Analogues</i>
Data Set	26, 27, 32-38, 40-63, 68-76, 78-80	1-25, 28, 29, 32, 33, 41, 44, 63-67, 70, 71, 74, 77	2, 4, 5, 25, 28, 30, 60-62, 64-68, 70-73, 75, 76
TS1	26, 27, 32, 33, 34, 45, 53, 56, 57, 60, 71, 80	9, 32, 33, 67, 71	60, 67, 71
TS2	27, 32, 37, 56, 60, 71, 80	11, 32, 71	60, 68, 71
TS3	26, 27, 32, 33, 34, 41, 56, 57, 62, 71, 80	11, 32, 33, 41, 67, 71	30, 62, 67, 71

APPENDIX C

MINIMIZATION AND SEARCH PARAMETERS

Minimization Parameters

Method:	Powell
Initial Optimization:	Simplex
Termination:	Gradient
Gradient:	0.05 kcal/mol-Å
Force Field:	MMFF94
Charges:	MMFF94
Dielectric Function:	Constant
Dielectric Constant:	1
Non-bonded Cutoff:	8 Å

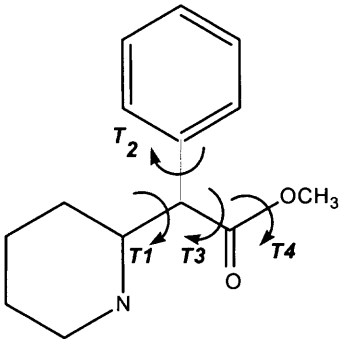
Systematic Search Parameters

Angle Increment:	30 degrees
Range:	0 – 359 degrees
Max. Energy Diff.:	20 kcal/mol
Use Electrostatics:	Yes
van der Waals Radius Scale Factors:	
General:	0.95
1-4:	0.87
H-bond:	0.65

APPENDIX D

TORSIONAL ANGLES FOR GEM CONFORMATIONS

Table D.1 Torsional Angles for MPN and MPP GEM Conformations

				
<i>GEM</i>	<i>T1</i>	<i>T2</i>	<i>T3</i>	<i>T4</i>
MPN	175.6	-105.7	-176.5	179.3
MPP	-169.2	-103.6	-166.4	179.3

APPENDIX E

COMFA PARAMETERS

Input parameters

CoMFA Field Class:	Tripos Standard
Field Types:	Both Steric and Electrostatic
Dielectric:	Distance
Smoothing:	None
Drop Electrostatics:	Within Steric Cutoff for Each Row
Steric Cutoff:	30 kcal/mol
ES Cutoff:	30 kcal/mol
Transition:	Smooth
Region:	Create Automatically

Parameters for CoMFA contour maps

a) mpn_c1_trn

Green/Yellow:	70/40
Blue/Red:	65/25

b) mpp_c1_trn

Green/Yellow:	70/45
Blue/Red:	70/20

APPENDIX F

MATLAB COMMANDS

The Matlab commands that were used to perform forward stepwise regression (see Section 6.3.1) are provided here. See section 7.1 for the nomenclature. In the following commands, if an X occurs in the name of the model it means that the commands listed were common to both the neutral and the protonated data sets.

Models MPX-RU

```
<< Stepwise regression for MPX-RU >>
```

```
stepwise(data, pIC50);
```

```
<< p_data matrix for MPN-RU >>
```

```
p_data(:,1) = data(:,10);
```

```
p_data(:,2) = data(:,9);
```

```
p_data(:,3) = data(:,11);
```

```
p_data(:,4) = data(:,12);
```

```
<< p_data matrix for MPP-RU >>
```

```
p_data(:,1) = data(:,9);
```

```
p_data(:,2) = data(:,10);
```

```
p_data(:,3) = data(:,11);
```

```
<< Statistics for MPX-RU >>
```

```
p_data = add_ones(p_data);
```

```
[b, bint, r, rint, stats] = regress(pIC50, p_data);
```

```
predicted = p_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-RU >>
```

```
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-RU >>
```

```
x = [3; 10];
```

```
y = [3; 10];
```

```
plot(x,y);
```

```
hold on;
```

```
plot(predicted, pIC50, 'o');
```

```
ylabel ('Predicted pIC50');
```

```
xlabel ('Experimental pIC50');
```

Models MPX-RR

```
<< Normalization for MPX-RR and generate minima and ranges >>
[mins, ranges, data_r] = normalize_ran(data);
```

```
<< Stepwise regression for MPX-RR >>
stepwise(data_r, pIC50);
```

```
<< p_data matrix for MPN-RR >>
p_data(:,1) = data_r(:,10);
p_data(:,2) = data_r(:,9);
p_data(:,3) = data_r(:,11);
p_data(:,4) = data_r(:,12);
```

```
<< p_data matrix for MPP-RR >>
p_data(:,1) = data_r(:,9);
p_data(:,2) = data_r(:,10);
p_data(:,3) = data_r(:,11);
```

```
<< Statistics for MPX-RR >>
p_data = add_ones(p_data);
[b, bint, r, rint, stats] = regress(pIC50, p_data);
predicted = p_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-RR >>
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-RR >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');
```

Models MPX-RS

```
<< Normalization for MPX-RS and generate means and standard deviations >>
[means, stds, data_s] = normalize_std(data);
```

```
<< Stepwise regression for MPX-RS >>
stepwise(data_s, pIC50);
```

```
<< p_data matrix for MPN-RS >>
p_data(:,1) = data_s(:,10);
p_data(:,2) = data_s(:,9);
p_data(:,3) = data_s(:,11);
p_data(:,4) = data_s(:,12);
```

```
<< p_data matrix for MPP-RS >>
p_data(:,1) = data_s(:,9);
p_data(:,2) = data_s(:,10);
p_data(:,3) = data_s(:,11);
```

```
<< Statistics for MPX-RS >>
p_data = add_ones(p_data);
[b, bint, r, rint, stats] = regress(pIC50, p_data);
predicted = p_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-RS >>
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-RS >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');
```

Models MPX-RM

```
<< Normalization for MPX-RM >>
```

```
[means, mads, data_m] = normalize_mad(data);
```

```
<< Stepwise regression for MPX-RM >>
```

```
stepwise(data_m, pIC50);
```

```
<< p_data matrix for MPN-RM >>
```

```
p_data(:,1) = data_m(:,10);
```

```
p_data(:,2) = data_m(:,9);
```

```
p_data(:,3) = data_m(:,11);
```

```
p_data(:,4) = data_m(:,12);
```

```
<< p_data matrix for MPP-RM >>
```

```
p_data(:,1) = data_m(:,9);
```

```
p_data(:,2) = data_m(:,10);
```

```
p_data(:,3) = data_m(:,11);
```

```
<< Statistics for MPX-RM >>
```

```
p_data = add_ones(p_data);
```

```
[b, bint, r, rint, stats] = regress(pIC50, p_data);
```

```
predicted = p_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-RM >>
```

```
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-RM >>
```

```
x = [3; 10];
```

```
y = [3; 10];
```

```
plot(x,y);
```

```
hold on;
```

```
plot(predicted, pIC50, 'o');
```

```
ylabel ('Predicted pIC50');
```

```
xlabel ('Experimental pIC50');
```

Models MPX-PU

```

<< PCA for MPX-PU >>
[PC, t_data, variances, t2] = princomp(data);

<< Stepwise regression for MPX-PU >>
stepwise(t_data, pIC50);

<< p_t_data matrix for MPN-PU >>
p_t_data(:,1) = t_data(:,3);
p_t_data(:,2) = t_data(:,4);
p_t_data(:,3) = t_data(:,12);
p_t_data(:,4) = t_data(:,1);
p_t_data(:,5) = t_data(:,13);

<< p_t_data matrix for MPP-PU >>
p_t_data(:,1) = t_data(:,4);
p_t_data(:,2) = t_data(:,12);
p_t_data(:,3) = t_data(:,2);
p_t_data(:,4) = t_data(:,13);

<< Statistics for MPX-PU >>
p_t_data = add_ones(p_t_data);
[b, bint, r, rint, stats] = regress(pIC50, p_t_data);
predicted = p_t_data * b;

<< Residuals plot with confidence intervals for MPX-PU >>
rcoplot(r, rint);

<< Predicted vs. Experimental plot for MPX-PU >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');

```

Models MPX-PR

```
<< Normalization for MPX-PR and generate mins and ranges >>
[mins, ranges, data_r] = normalize_ran(data);
```

```
<< PCA for MPX-PR >>
[PC, t_data, variances, t2] = princomp(data_r);
```

```
<< Stepwise regression for MPX-PR >>
stepwise(t_data, pIC50);
```

```
<< p_t_data matrix for MPN-PR >>
p_t_data(:,1) = t_data(:,3);
p_t_data(:,2) = t_data(:,12);
p_t_data(:,3) = t_data(:,5);
p_t_data(:,4) = t_data(:,2);
```

```
<< p_t_data matrix for MPP-PR >>
p_t_data(:,1) = t_data(:,3);
p_t_data(:,2) = t_data(:,5);
p_t_data(:,3) = t_data(:,12);
```

```
<< Statistics for MPX-PR >>
p_t_data = add_ones(p_t_data);
[b, bint, r, rint, stats] = regress(pIC50, p_t_data);
predicted = p_t_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-PR >>
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-PR >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');
```

Models MPX-PS

```
<< Normalization for MPX-PS and generate means and standard deviations >>
[means, stds, data_s] = normalize_std(data);
```

```
<< PCA for MPX-PS >>
[PC, t_data, variances, t2] = princomp(data_s);
```

```
<< Stepwise regression for MPX-PS >>
stepwise(t_data, pIC50);
```

```
<< p_t_data matrix for MPN-PS >>
p_t_data(:,1) = t_data(:,3);
p_t_data(:,2) = t_data(:,2);
p_t_data(:,3) = t_data(:,12);
p_t_data(:,4) = t_data(:,6);
p_t_data(:,5) = t_data(:,5);
```

```
<< p_t_data matrix for MPP-PS >>
p_t_data(:,1) = t_data(:,2);
p_t_data(:,2) = t_data(:,12);
p_t_data(:,3) = t_data(:,4);
p_t_data(:,4) = t_data(:,6);
```

```
<< Statistics for MPX-PS >>
p_t_data = add_ones(p_t_data);
[b, bint, r, rint, stats] = regress(pIC50, p_t_data);
predicted = p_t_data * b;
```

```
<< Residuals plot with confidence intervals for MPX-PS >>
rcoplot(r, rint);
```

```
<< Predicted vs. Experimental plot for MPX-PS >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');
```

Models MPX-PM

```
<< Normalization for MPX-PM >>
[means, mads, data_m] = normalize_mad(data);

<< PCA for MPX-PM >>
[PC, t_data, variances, t2] = princomp(data_m);

<< Stepwise regression for MPX-PM >>
stepwise(t_data, pIC50);

<< p_t_data matrix for MPN-PM >>
p_t_data(:,1) = t_data(:,3);
p_t_data(:,2) = t_data(:,12);
p_t_data(:,3) = t_data(:,6);
p_t_data(:,4) = t_data(:,2);
p_t_data(:,5) = t_data(:,4);

<< p_t_data matrix for MPP-PM >>
p_t_data(:,1) = t_data(:,2);
p_t_data(:,2) = t_data(:,12);
p_t_data(:,3) = t_data(:,6);
p_t_data(:,4) = t_data(:,3);
p_t_data(:,5) = t_data(:,5);
p_t_data(:,6) = t_data(:,4);
p_t_data(:,7) = t_data(:,11);

<< Statistics for MPX-PM >>
p_t_data = add_ones(p_t_data);
[b, bint, r, rint, stats] = regress(pIC50, p_t_data);
predicted = p_t_data * b;

<< Residuals plot with confidence intervals for MPX-PM >>
rcoplot(r, rint);

<< Predicted vs. Experimental plot for MPX-PM >>
x = [3; 10];
y = [3; 10];
plot(x,y);
hold on;
plot(predicted, pIC50, 'o');
ylabel ('Predicted pIC50');
xlabel ('Experimental pIC50');
```


APPENDIX G

D-SIM VERSION 1.0 PROGRAM CODE

The reference for this program is reference 145: Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modeling* 1997, **15**, 372-385.

```
/*
 * dsim.cpp
 *
 * D-SIM Version 1.0
 *
 * Author: Milind Misra
 *
 * Date Created:      August 29, 2005
 * Date Modified:    September 05, 2005
 *
 * This program calculates pairwise Soergel distances and then
 * identifies test and training sets for the input data set. It
 * implements a sphere exclusion algorithm in which the molecule
 * with the minimum maximum dissimilarity with the current test
 * set is selected for inclusion in the test set. The reference
 * for this program is 1997_jmgm_15_372 (present in RefBase). It
 * provides a description of the SE-MinMax algorithm that is used
 * here, along with descriptions of other dissimilarity-based
 * compound selection (DBCS) algorithms.
 *
 * Input:  A space-separated text file containing:
 *         1) Input matrix dimensions on the first line.
 *         2) Input descriptor matrix with values [0-1].
 *
 * Output: Test set molecule indexes.
 */

#include <vector>
#include <iostream>
#include <fstream>
#include <cassert>
#include <string>
#include <sstream>
#include <ctime>

using namespace std;

/*
 * toString()
 *
 * Define a toString function to convert basic types to string.
 */
template < class T >
inline std::string toString(const T & value)
{
    std::ostringstream strm;
    strm << value;
```

```

    return strm.str();
};

/*
 * molecule
 *
 * Define a struct that would uniquely identify a molecule and store
 * molecule specific data including its identity (index), the values
 * of its descriptors (a row from the input file), the values of its
 * dissimilarity with every other molecule (calculated by this
 * program), the sum of this molecule's dissimilarity values
 * (calculated by this program), and the status of the molecule
 * (0: original data set, -1: training set, and >0: test set; for a
 * test set molecule, its actual status value = the index for that
 * molecule). All molecules are initialized with status = 0.
 */
typedef struct {
    int index;
    vector<double> descriptors;
    vector<double> dissimilarity;
    double sumDissimilarity;
    int status;
} molecule;

// GLOBAL VARIABLES:

/*
 * v
 *
 * Declare a vector of (molecule) structs. Thus, a particular value
 * of dissimilarity between molecules i and j can be accessed by:
 *
 *   v.at(i).dissimilarity.at(j).
 *
 * Size of v = number of rows in the input data matrix.
 */
vector<molecule> v;

/*
 * testCount
 *
 * Stores the count of test set molecules.
 */
int testCount = 0;

/*
 * trainingCount
 *
 * Stores the count of training set molecules.
 */
int trainingCount = 0;

// FUNCTION DECLARATIONS:

    void calcDistances();
    void getData(string inputFileName);
    void calcTestSet(int algorithmType, double threshold, string inputFileName);
    int getMinSumIndex();
double getMaxDissimilarity();
    void markTrainingSet(int sumIndex, double threshold);
    void se_minMax(double threshold, string inputFileName);
    void se_minSum(double threshold, string inputFileName);
    void se_minMin(double threshold, string inputFileName);
    bool isMarked();
    void printLog(string inputFileName, string addText);
string logStatus();
    void printUsage();

```

```

    void printUsage(string errorMsg);
    void printOutput(string inputFileName, double threshold, int algorithmType);
string printHeader();
string printFooter();

// MAIN PROGRAM FUNCTION:

int main(int argc, char *argv[])
{
/*
 * threshold
 *
 * The value of threshold determines the radius of the hypersphere
 * around each of the test set molecules. All molecules that are
 * included in this hypersphere are excluded from the original data
 * set and are not considered for subsequent inclusion in the test
 * set. In theory, this value can be changed to yield different
 * test sets. Thus:
 *
 * threshold > 0.15: Theoretically smaller test set.
 * threshold < 0.15: Theoretically larger test set.
 *
 * Also:
 *
 * 0.0 < threshold < 1.0
 *
 * Once set, threshold will always find the SAME test set. In D-SIM
 * version 1.0, the user can specify the threshold as an argument
 * (with the flag "-t") on the command line otherwise it defaults to
 * 0.15. The program checks if the user-specified threshold is
 * between 0.0 and 1.0 and exits if it is not.
 */
    double threshold = 0.15;

/*
 * algorithmType
 *
 * The value of algorithmType determines the specific sphere-
 * exclusion algorithm used to calculate the requested test set.
 * Thus:
 *
 * algorithmType = 1: SE-MinMax (default)
 * algorithmType = 2: SE-MinSum
 * algorithmType = 3: SE-MinMin
 *
 * For a description of these algorithms, see 1997_jmgm_15_372
 * (present in RefBase). The program checks for user input and
 * exits if the value following the algorithm flag ("-a") is not
 * 1, 2, or 3.
 */
    int algorithmType = 1;    // SE-MinMax (Default)

/*
 * inputFileName
 *
 * The input file must be a text file with the format: input.txt.
 * The file name must be specified on the command line when running
 * D-SIM with the flag "-f". Additionally, the file name must be
 * entered without its extension (.txt). For example, if the input
 * data is contained in a file called sample.txt, then the name
 * entered on the command line must be just "sample". Thus, the
 * program should be invoked as:
 *
 * dsim -f sample
 * dsim -t 0.2 -f sample
 * dsim -a 3 -f sample
 * dsim -t 0.1 -f sample -a 2, etc.
 */

```

```

* The program checks for empty string and searches for the "."
* (dot) character. If either of these conditions is true, the
* program exits.
*/
string inputFileName = "";

int numFlags = 1;

// Parse command line arguments.
while ((numFlags<argc) && (argv[numFlags][0]!='-') && (argc>2)) {

    string inputFlag = argv[numFlags];

    if (inputFlag == "-t") {
        numFlags++;
        // Assign user-specified threshold.
        threshold = atof(argv[numFlags]);
    }
    else if (inputFlag == "-a") {
        numFlags++;
        // Assign user-specified algorithm.
        algorithmType = atoi(argv[numFlags]);
    }
    else if (inputFlag == "-f") {
        numFlags++;
        // Assign input file name root (used for output and log files).
        inputFileName = argv[numFlags];
    }
    else {
        string errorMsg = " ERROR: Unknown switch: ";
        errorMsg += toString(argv[numFlags]) + "\n";
        printUsage(errorMsg);
        exit(1);
    }

    numFlags++;
}

// Exit program if no file name was specified by the user.
if (inputFileName.empty()) {
    printUsage();
    exit(1);
}

// Exit program if file name contains an extension.
if (inputFileName.find(".", 0) != string::npos) {
    string errorMsg = " ERROR: Bad input file name.";
    errorMsg += "\n Provide text file name without extension:\n";
    printUsage(errorMsg);
    exit(1);
}

// Exit program if threshold specified is not between 0.0 and 1.0.
if (threshold <= 0.0 || threshold >= 1.0) {
    string errorMsg = " ERROR: Threshold must be between 0.0 and 1.0.";
    errorMsg += "\n Default threshold: 0.15\n";
    printUsage(errorMsg);
    exit(1);
}

// Exit program if algorithm specified is invalid.
if (algorithmType < 1 || algorithmType > 3) {
    string errorMsg = " ERROR: Three algorithms are currently available.";
    errorMsg += "\n Select:  1 (SE-MinMax) [default]";
    errorMsg += "\n          2 (SE-MinSum)";
    errorMsg += "\n          3 (SE-MinMin)\n";
    printUsage(errorMsg);
    exit(1);
}

```

```

// Declare starting time for finding total calculation time.
time_t startTime, endTime;

// Specify start time for calculations.
time(&startTime);

// Get current system data and time information.
time_t rawtime;
struct tm * timeinfo;
time (&rawtime);
timeinfo = localtime(&rawtime);

// Read input data and initialize vector of structs.
getData(inputFileName);

// Declare and assign program name for the log file.
string logFile = inputFileName + ".log";
ofstream logStream(logFile.data());
assert(logStream.is_open());
logStream << printHeader();
logStream.close();

string logText = "";
logText += "      " + toString(asctime(timeinfo));
logText += "      (LOGFILE)\n";
logText += "\nThis log file provides information about ";
logText += "successive test set\n";
logText += "identification steps.  The attribute \"status\" ";
logText += "is set to 0 for\n";
logText += "all molecules in the original data set.  When the ";
logText += "first test\n";
logText += "set molecule is identified, its status changes to ";
logText += "a positive\n";
logText += "number which identifies its index in the original ";
logText += "data set.\n";
logText += "The status of all molecules that are found to be ";
logText += "within the\n";
logText += "threshold specified for this analysis is set to ";
logText += "-1.  As more\n";
logText += "test set molecules are progressively identified by ";
logText += "the\n";
logText += "algorithm selected for this analysis, their status ";
logText += "is changed\n";
logText += "to reflect their index in the original data set.  ";
logText += "At each\n";
logText += "step, other molecules that satisfy the threshold ";
logText += "criterion are\n";
logText += "placed in the training set by changing their ";
logText += "status value from\n";
logText += "0 to -1.  At the end, there should be no 0 and as ";
logText += "many positive\n";
logText += "numbers as molecules identified for the test set.";
logText += "  The rest\n";
logText += "should have a value of -1.\n";
logText += "\nLog file:      " + inputFileName + ".log";
logText += "\nInput file:    " + inputFileName + ".txt";
logText += "\nOutput file:   " + inputFileName + ".out";
logText += "\n\nThreshold used: " + toString(threshold);
printLog(inputFileName, logText);

// Calculate Soergel distances.
calcDistances();

// Calculate test set for input data set.
calcTestSet(algorithmType, threshold, inputFileName);

// Print output to outputFile.
printOutput(inputFileName, threshold, algorithmType);

// Specify end time for calculations.
time(&endTime);

```

```

    logText = "\n\nTime taken: " + toString(difftime(endTime, startTime));
    logText += " second(s)\n";
    logText += printFooter();
    printLog(inputFileName, logText);

    return 0;
}

// FUNCTIONS:

void getData(string inputFileName) {

    // Declare and assign program name for the input file.
    string inputFile = inputFileName + ".txt";

    // Declare file stream for the input file.
    ifstream inputStream;

    cout << endl << " Opening " << inputFile << "...";

    // Activate file stream for the input file.
    inputStream.open(inputFile.data());

    // Exit program if input file not ready.
    assert(inputStream.is_open());

    cout << "DONE." << endl;
    cout << " Reading row and column size data...";

    // Declare variable to read double data from input file.
    double cursor = 0.0;

    // Read the first number from the input file.
    inputStream >> cursor;

    // Assign the first number to rowSize.
    int rowSize = int(cursor);

    // Read the second number from the input file.
    inputStream >> cursor;

    // Assign the second number to colSize.
    int colSize = int(cursor);

    cout << "DONE." << endl;
    cout << " Number of molecules in data set: " << rowSize << endl;
    cout << " Number of descriptors in data set: " << colSize << endl;
    cout << " Initializing vector of structs and reading ";
    cout << inputFile << "...";

    // Declare molecules and initialize v.
    for (int i=0; i<rowSize; i++) {
        molecule m;
        m.index = i;
        m.sumDissimilarity = 0.0;
        m.status = 0;

        for (int j=0; j<colSize; j++) {
            inputStream >> cursor;
            m.descriptors.push_back(cursor);
        }

        for (int k=0; k< rowSize; k++)
            m.dissimilarity.push_back(0.0);

        // Populate v with molecules.
        v.push_back(m);
    }

    cout << "DONE." << endl;
}

```

```

    cout << " Closing " << inputFile << "...";

    // Close file stream for the input file.
    inputStream.close();

    cout << "DONE." << endl;
}

void calcDistances() {

    for (int i=0; i<v.size(); i++) {
        for (int j=0; j<v.size(); j++) {

            // Initialize products to zero before each calculation.
            double sumProducts = 0.0;
            double sumSquares1 = 0.0;
            double sumSquares2 = 0.0;

            for (int k=0; k<v[i].descriptors.size(); k++) {
                sumProducts += v[i].descriptors[k] * v[j].descriptors[k];
                sumSquares1 += v[i].descriptors[k] * v[i].descriptors[k];
                sumSquares2 += v[j].descriptors[k] * v[j].descriptors[k];
            }

            // Calculate the Soergel distance for (i,j).
            v[i].dissimilarity[j] = sumSquares1+sumSquares2-sumProducts;
            v[i].dissimilarity[j] = sumProducts/v[i].dissimilarity[j];
            v[i].dissimilarity[j] = 1 - v[i].dissimilarity[j];

            // Calculate the sum of the dissimilarities for i.
            v[i].sumDissimilarity += v[i].dissimilarity[j];
        }
    }
}

void calcTestSet(int algorithmType, double threshold, string inputFileName) {

    // Declare and initialize variable for log file information.
    string logText = "";

    // Log initial values of each molecule's status (all zeroes).
    printLog(inputFileName, logStatus());

    // Log statement of zero test set count.
    logText = "Original data set status (test set count: " + toString(testCount) + ").\n";
    printLog(inputFileName, logText);

    logText = "\nSelecting first molecule based on ";
    logText += "minimum sum of dissimilarities...";
    printLog(inputFileName, logText);

    /*
    * Select the first molecule for the test set by setting the status
    * from 0 to the value of the index of the molecule that has the
    * minimum sum of dissimilarities.
    *
    * Declare sumIndex for the molecule with the above property and
    * determine the identity of the first test set molecule.
    */
    int sumIndex = getMinSumIndex();

    // Select the first test set molecule.
    v.at(sumIndex).status = v.at(sumIndex).index;

    // Increment count of test set molecules.
    testCount++;

    /*
    * Set the status from 0 to -1 for all molecules that have a Soergel
    * distance with the first test set molecule that is less than the

```

```

* threshold.
*
*/
markTrainingSet(sumIndex, threshold);

// Log status after the first molecule has been selected.
printLog(inputFileName, logStatus());

logText = "Test set count: " + toString(testCount) + "    ...DONE.\n";
printLog(inputFileName, logText);

switch(algorithmType) {
  case 2:
    logText = "\nSelecting the others by the SE-MinSum algorithm...";
    printLog(inputFileName, logText);
    se_minSum(threshold, inputFileName);
    break;
  case 3:
    logText = "\nSelecting the others by the SE-MinMin algorithm...";
    printLog(inputFileName, logText);
    se_minMin(threshold, inputFileName);
    break;
  default:
    logText = "\nSelecting the others by the SE-MinMax algorithm...";
    printLog(inputFileName, logText);
    se_minMax(threshold, inputFileName);
}
}

/*
* se_minMax()
*
* Calculates the remaining test set molecules by determining the
* molecule in the remaining set (molecules with status = 0) that
* is most dissimilar to AND least distant from the test set.
*
*/
void se_minMax(double threshold, string inputFileName) {

/*
* Declare variable for the molecule that is least distant
* from the test set. This variable will be set to 0 after
* every iteration of the while loop below.
*
*/
  int iMinIndex = 0;

/*
* Declare variable for dissimilarity value of the molecule
* corresponding to iMinIndex above. This variable must be
* set to 1.0 after every iteration of the while loop below.
*
*/
  double iMinValue = 1.0;

  // Declare variable for screen output of loop iterations.
  int pass = 1;

/*
* The following while loop block performs the selection of the
* rest of the test set (after the first molecule is selected
* above). The loop condition is a call to a boolean function
* that returns true only if there are molecules in the original
* data set that have not yet been marked as either belonging to
* the test set (status[] = 1) or belonging to the training set
* (status[] = -1).
*
* In this block, the indexes correspond as follows:
*
*   i --> ith current test set molecule
*   j --> jth current residual set molecule

```



```

*
* Thus, the loop does two main comparisons as follows:
*
*      max { (i1,j1), (i2,j1), (i3, j1), ... }
*      min { max { (i1,j2), (i2,j2), (i3, j2), ... } }
*      ^      max { (i1,j3), (i2,j3), (i3, j3), ... }
*      ^      ^
*      ^      ^
*      ^      jMaxValue and jMaxIndex
*      ^
*      iMinValue and iMinIndex
*
*/
while (!isMarked()) {

    int jMaxIndex = 0;
    double jMaxValue = 0.0;

    // For each test set molecule with each unmarked molecule.
    for (int j=0; j<v.size(); j++) {
        for (int i=0; i<v.size(); i++) {
            if (v.at(j).status==0 && v.at(i).status>0) {
                double jValue = v.at(j).dissimilarity.at(i);
                int jIndex = v.at(j).index;
                // Determine the molecule most dissimilar to the test set.
                if (jValue > jMaxValue) {
                    jMaxValue = jValue;
                    jMaxIndex = jIndex;
                }
            }
        }
    }

    // Determine the molecule least distant from the test set.
    if (jMaxValue < iMinValue) {
        iMinValue = jMaxValue;
        iMinIndex = jMaxIndex;
    }

    // Select another test set molecule.
    v.at(iMinIndex).status = v.at(iMinIndex).index;

    // Increment count of test set molecules.
    testCount++;

/*
* Set the status from 0 to -1 for all molecules that have
* a Soergel distance with the test set molecule just selected
* of less than the threshold.
*/
    markTrainingSet(iMinIndex, threshold);

    printLog(inputFileName, logStatus());

    string logText = "";
    logText = "Test set count: " + toString(++pass);
    printLog(inputFileName, logText);

    // Restore initialization for next iteration of the while loop.
    iMinValue = 1.0;
    iMinIndex = 0;

} // END of while loop.

string logText = "    ...DONE.";
printLog(inputFileName, logText);
}

/*
* se_minSum()

```

```

*
* Calculates the remaining test set molecules by determining the
* molecule in the remaining set (molecules with status = 0) that
* has the least sum of dissimilarities in every iteration.
*
*/
void se_minSum(double threshold, string inputFileName) {

/*
* Declare variable for the molecule that is least distant
* from the test set. This variable will be set to 0 after
* every iteration of the while loop below.
*
*/
    int iMinIndex = 0;

/*
* Declare variable for dissimilarity value of the molecule
* corresponding to iMinIndex above. This variable must be
* set to 1.0 after every iteration of the while loop below.
*
*/
    double iMinValue = 1.0;

    // Declare variable for screen output of loop iterations.
    int pass = 1;

/*
* The following while loop block performs the selection of the
* rest of the test set (after the first molecule is selected
* above). The loop condition is a call to a boolean function
* that returns true only if there are molecules in the original
* data set that have not yet been marked as either belonging to
* the test set (status[] = 1) or belonging to the training set
* (status[] = -1).
*
* In this block, the indexes correspond as follows:
*
*     i --> ith current test set molecule
*     j --> jth current training set molecule
*
* Thus, the loop does two main comparisons as follows:
*
*         sum { (i1,j1), (i2,j1), (i3, j1), ... }
* min { sum { (i1,j2), (i2,j2), (i3, j2), ... } }
*     ^   sum { (i1,j3), (i2,j3), (i3, j3), ... }
*     ^   ^
*     ^   ^
*     ^   jSumValue and jSumIndex
*     ^
*     iMinValue and iMinIndex
*
*/
    while (!isMarked()) {

        double jSumValue = getMaxDissimilarity();
        int jSumIndex = 0;

        // For each test set molecule with each unmarked molecule.
        for (int j=0; j<v.size(); j++) {
            for (int i=0; i<v.size(); i++) {
                if (v.at(j).status==0 && v.at(i).status>0) {
                    int jIndex = 0;
                    double jValue = 0.0;

                    for (int s=0; s<v.size(); s++) {
                        if (v.at(s).status>0) {
                            // Determine the sum of the dissimilarities.
                            jValue += v.at(j).dissimilarity.at(s);
                        }
                    }
                }
            }
        }
    }
}

```

```

        jIndex = v.at(j).index;

        // Determine the molecule with the least sum dissimilarities.
        if (jValue < jSumValue) {
            jSumValue = jValue;
            jSumIndex = jIndex;
        }
    }
}

iMinValue = jSumValue;
iMinIndex = jSumIndex;

// Select another test set molecule.
v.at(iMinIndex).status = v.at(iMinIndex).index;

// Increment count of test set molecules.
testCount++;

/*
 * Set the status from 0 to -1 for all molecules that have
 * a Soergel distance with the test set molecule just selected
 * of less than the threshold.
 */
markTrainingSet(iMinIndex, threshold);

printLog(inputFileName, logStatus());

string logText = "";
logText = "Test set count: " + toString(++pass);
printLog(inputFileName, logText);

// Restore initialization for next iteration of the while loop.
iMinValue = 1.0;
iMinIndex = 0;

} // END of while loop.

string logText = "    ...DONE.";
printLog(inputFileName, logText);
}

/*
 * se_minMin()
 *
 * Calculates the remaining test set molecules by determining the
 * molecule in the remaining set (molecules with status = 0) that
 * is most similar to AND most distant from the test set.
 */
void se_minMin(double threshold, string inputFileName) {

/*
 * Declare variable for the molecule that is least distant
 * from the test set. This variable will be set to 0 after
 * every iteration of the while loop below.
 */
    int iMinIndex = 0;

/*
 * Declare variable for dissimilarity value of the molecule
 * corresponding to iMinIndex above. This variable must be
 * set to 1.0 after every iteration of the while loop below.
 */
    double iMinValue = 1.0;

```

```

// Declare variable for screen output of loop iterations.
int pass = 1;

/*
 * The following while loop block performs the selection of the
 * rest of the test set (after the first molecule is selected
 * above). The loop condition is a call to a boolean function
 * that returns true only if there are molecules in the original
 * data set that have not yet been marked as either belonging to
 * the test set (status[] = 1) or belonging to the training set
 * (status[] = -1).
 *
 * In this block, the indexes correspond as follows:
 *
 *     i --> ith current test set molecule
 *     j --> jth current training set molecule
 *
 * Thus, the loop does two main comparisons as follows:
 *
 *         min { (i1,j1), (i2,j1), (i3, j1), ... }
 *     min { min { (i1,j2), (i2,j2), (i3, j2), ... } }
 *         ^      min { (i1,j3), (i2,j3), (i3, j3), ... }
 *         ^      ^
 *         ^      ^
 *         ^      jMinValue and jMinIndex
 *         ^
 *         iMinValue and iMinIndex
 */
while (!isMarked()) {

    int jMinIndex = 0;
    double jMinValue = 1.0;

    for (int j=0; j<v.size(); j++) {
        for (int i=0; i<v.size(); i++) {
            // For each test set molecule with each unmarked molecule.
            if (v.at(j).status==0 && v.at(i).status>0) {
                double jValue = v.at(j).dissimilarity.at(i);
                int jIndex = v.at(j).index;
                // Determine the molecule most similar to the test set.
                if (jValue < jMinValue) {
                    jMinValue = jValue;
                    jMinIndex = jIndex;
                }
            }
        }
    }

    // Determine the molecule most distant from the test set.
    if (jMinValue < iMinValue) {
        iMinValue = jMinValue;
        iMinIndex = jMinIndex;
    }

    // Select another test set molecule.
    v.at(iMinIndex).status = v.at(iMinIndex).index;

    // Increment count of test set molecules.
    testCount++;

/*
 * Set the status from 0 to -1 for all molecules that have
 * a Soergel distance with the test set molecule just selected
 * of less than the threshold.
 */
    markTrainingSet(iMinIndex, threshold);

    printLog(inputFileName, logStatus());
}

```

```

    string logText = "";
    logText = "Test set count: " + toString(++pass);
    printLog(inputFileName, logText);

    // Restore initialization for next iteration of the while loop.
    iMinValue = 1.0;
    iMinIndex = 0;

} // END of while loop.

string logText = "    ...DONE.";
printLog(inputFileName, logText);
}

/*
 * getMinSumIndex()
 *
 * The vector sumDissimilarity[] below is of size = rowSize and
 * contains the sum of the dissimilarities for each molecule. The
 * function getMinSumIndex() returns the location of the molecule
 * that has the minimum sum of dissimilarities. This molecule
 * becomes the first one to be included in the test set.
 */

int getMinSumIndex() {

    double minSumValue = v.at(0).sumDissimilarity;
    int minSumIndex = 0;

    for (int i=0; i<v.size(); i++) {
        if (v.at(i).sumDissimilarity < minSumValue) {
            minSumValue = v.at(i).sumDissimilarity;
            minSumIndex = v.at(i).index;
        }
    }

    return minSumIndex;
}

double getMaxDissimilarity() {

    double maxSumValue = v.at(0).sumDissimilarity;

    for (int i=0; i<v.size(); i++)
        if (v.at(i).sumDissimilarity > maxSumValue)
            maxSumValue = v.at(i).sumDissimilarity;

    return maxSumValue;
}

/*
 * markTrainingSet()
 *
 * This function: 1) Checks that only those molecule are affected
 * that remain in the original data set (i.e., status[] = 0) and have
 * not been included in either the test set (i.e., status[] = 1) or
 * the training set (i.e., status[] = -1). 2) Uses the value defined
 * for threshold to eliminate molecules from the test set. Thus, the
 * test set will not include those molecules whose Soergel distance
 * is less than the threshold value. 3) Increments the count for
 * the training set. This function does not return anything. The
 * variables affected are defined as global variables.
 */

void markTrainingSet(int index, double threshold) {

    for (int i=0; i<v.size(); i++) {
        if (v.at(i).status==0 && v.at(index).dissimilarity.at(i)<threshold) {

```

```

        // "Remove" from original data set molecules less dissimilar
        //      than threshold by setting status to -1.
        v.at(i).status = -1;

        // Increment number of training set molecules by one.
        trainingCount++;
    }
}

/*
 * isMarked()
 *
 * This boolean function returns true only if all molecules in the
 * original data set have been included in either the test set (marked
 * as 1) or the training set (marked as -1). If a single molecule
 * remains that has a value of status[] = 0, then this function
 * returns false indicating that all molecules in the original data
 * set have not yet been marked.
 */

bool isMarked() {
    bool marked = true;

    for (int i=0; i<v.size(); i++)
        if (v.at(i).status == 0) {
            marked = false;
            i = v.size();
        }

    return marked;
}

void printOutput(string inputFileName, double threshold, int algorithmType) {
    string algorithm = "";

    switch(algorithmType) {
        case 2:
            algorithm = "Sphere-Exclusion Minimum Sum (SE-MinSum).";
            break;
        case 3:
            algorithm = "Sphere-Exclusion Minimum Minimum (SE-MinMin).";
            break;
        default:
            algorithm = "Sphere-Exclusion Minimum Maximum (SE-MinMax).";
    }

    // Declare and assign program name for the output file.
    string outputFile = inputFileName + ".out";

    cout << " Opening " << outputFile << "...";

    // Declare and activate file stream for the output file.
    ofstream outputStream(outputFile.data());

    // Exit program if input file not ready.
    assert(outputStream.is_open());

    cout << "DONE." << endl;
    cout << " Writing to " << outputFile << "...";

    /*
    // Debug: Print dissimilarities to output file.
    for (int i=0; i<v.size(); i++) {
        for (int j=0; j<v.size(); j++) {
            outputStream << v[i].dissimilarity[j] << " ";
            if (j == v.size() - 1) {

```

```

        outputStream << endl;
    }
}
*/

// Get current system data and time information.
time_t rawtime;
struct tm * timeinfo;
time (&rawtime);
timeinfo = localtime(&rawtime);

outputStream << printHeader();
outputStream << "      " << asctime(timeinfo);
outputStream << "          (OUTPUTFILE)\n";
outputStream << endl;
outputStream << "D-SIM calculates the pairwise Soergel ";
outputStream << "distances and then" << endl;
outputStream << "identifies test and training sets for the ";
outputStream << "input data set." << endl;
outputStream << "The input descriptors must have values ";
outputStream << "between [0,1].\n";
outputStream << "It implements a sphere exclusion algorithm ";
outputStream << "in which the" << endl;
outputStream << "molecule with the minimum maximum dissimilarity ";
outputStream << "with the" << endl;
outputStream << "current test set is selected for inclusion in ";
outputStream << "the test set." << endl;
outputStream << "This is the SE-MinMax (default) algorithm. Two ";
outputStream << "other" << endl;
outputStream << "versions, SE-MinSum and SE-MinMin, are also ";
outputStream << "available." << endl << endl;
outputStream << "The algorithms used in this program have ";
outputStream << "been described" << endl;
outputStream << "in 1997_jmgm_15_372 that can be searched ";
outputStream << "for in RefBase." << endl;
outputStream << endl;
outputStream << " -----" << endl;
outputStream << endl;
outputStream << "The Tanimoto distance between two molecules, ";
outputStream << "i and j, is:" << endl;
outputStream << endl;
outputStream << "      S(i,j) = C / (A + B - C)" << endl;
outputStream << endl;
outputStream << "where, A: Sum of squared descriptor values ";
outputStream << "for i," << endl;
outputStream << "      B: Sum of squared descriptor values ";
outputStream << "for j, and" << endl;
outputStream << "      C: Sum of products of descriptor ";
outputStream << "values for i and j." << endl;
outputStream << endl;
outputStream << "The Soergel distance between i and j is then:";
outputStream << endl << endl;
outputStream << "      D(i,j) = 1 - S(i,j)" << endl;
outputStream << endl;
outputStream << " -----" << endl;
outputStream << endl;
outputStream << "Log file:      " << inputFileName << ".log" << endl;
outputStream << "Input file:   " << inputFileName << ".txt" << endl;
outputStream << "Output file:  " << inputFileName << ".out" << endl;
outputStream << endl << "Number of molecules: " << v.size() << endl;
outputStream << "Number of descriptors: " << v.at(0).descriptors.size();
outputStream << endl << endl;
outputStream << "Threshold dissimilarity: " << threshold << endl;
outputStream << "Algorithm:    " << algorithm << endl;
outputStream << endl;
outputStream << "Number of molecules in the test set: ";
outputStream << testCount << endl;
outputStream << "Number of molecules in the training set: ";
outputStream << trainingCount << endl;
outputStream << "Percentage of test set molecules: ";

```

```

outputStream << (float(testCount)/v.size() * 100) << "%" << endl;
outputStream << endl;
outputStream << " -----" << endl;
outputStream << endl;
outputStream << "Test set identified for this data set:";
outputStream << endl << endl;
outputStream << "      ";

// Output test set to output file if test set exists.
if (testCount != 0) {
    int formatCount = 0;

    for (int x=0; x<v.size(); x++) {
        if (v.at(x).status > 1) {
            outputStream << v.at(x).index + 1;
            formatCount++;
            if (formatCount%5 == 0 && x != v.size()-1)
                outputStream << endl << "      ";
            else if (x != v.size()-1)
                outputStream << " ";
        }
        if (x == v.size()-1)
            outputStream << endl;
    }
}
else
    outputStream << "Empty test set!" << endl;

outputStream << printFooter();

cout << "DONE." << endl;
cout << " Closing " << outputFile << "...";

// Close file stream for the output file.
outputStream.close();

cout << "DONE." << endl;
}

void printLog(string inputFileName, string addText) {

    string logFile = inputFileName + ".log";
    ofstream logStream;
    logStream.open(logFile.c_str(), ios::out | ios::app);
    assert(logStream.is_open());
    logStream << addText;
    logStream.close();
}

string logStatus() {

    string statusLog = "\n\n";

    for (int i=0; i<v.size(); i++) {
        v.at(i).status>0 ? (statusLog += toString(v.at(i).status+1)) : (statusLog +=
toString(v.at(i).status));
        statusLog += " ";
        if ((i+1) % 20 == 0 || i == v.size()-1)
            statusLog += "\n";
    }

    return statusLog;
}

void printUsage() {

    cout << printHeader();
    cout << endl;
    cout << " Usage (flag order independent): " << endl << endl;
    cout << "      dsim -t threshold -a algorithm -f inputfile";
    cout << endl << endl;
}

```



```

    cout << "  Examples for sample.txt:" << endl;
    cout << "    1.  dsim -f sample" << endl;
    cout << "    2.  dsim -f sample -t 0.2" << endl;
    cout << "    3.  dsim -a 3 -f sample" << endl;
    cout << "    4.  dsim -t 0.1 -f sample -a 2" << endl;
    cout << printFooter();
}

void printUsage(string errorMsg) {

    cout << printHeader();
    cout << endl << errorMsg << endl;
    cout << "  Usage (flag order independent): " << endl << endl;
    cout << "    dsim -t threshold -a algorithm -f inputfile";
    cout << endl << endl;
    cout << "  Examples for sample.txt:" << endl;
    cout << "    1.  dsim -f sample" << endl;
    cout << "    2.  dsim -f sample -t 0.2" << endl;
    cout << "    3.  dsim -a 3 -f sample" << endl;
    cout << "    4.  dsim -t 0.1 -f sample -a 2";
    cout << printFooter();
}

string printHeader() {

    string header = "\n          D-SIM version 1.0\n";
    header += "  ----- \n";

    return header;
}

string printFooter() {

    string footer = "\n -----";
    footer += "\n  Copyright (c) 2005 Milind Misra\n\n";

    return footer;
}

```

APPENDIX H

ALL POSSIBLE SUBSETS REGRESSION RESULTS

Note: "Number of Observations" refers to the number of training set analogues.

MPN data set

The SAS System			06:46 Friday, November 11, 2005 1		
The REG Procedure					
Model: MODEL1					
Dependent Variable: pIC50					
R-Square Selection Method					
Number of Observations Read			66		
Number of Observations Used			66		
Number in Model	R-Square	Variables in Model			
1	0.1698	ES11			
1	0.1447	ES09			
1	0.1381	ES10			
1	0.0586	ES12			
1	0.0164	ES13			
1	0.0077	ES01			
1	0.0043	ES02			
1	0.0021	ES07			
1	0.0021	ES03			
1	0.0003	ES08			
1	0.0002	ES04			
1	0.0001	ES05			
1	0.0000	ES06			
1	0.0000	ES14			

2	0.3640	ES09 ES10			
2	0.3634	ES09 ES11			
2	0.3391	ES09 ES12			
2	0.3265	ES09 ES13			
2	0.2373	ES08 ES09			
2	0.2247	ES08 ES11			
2	0.2209	ES02 ES09			
2	0.2133	ES10 ES11			
2	0.2066	ES11 ES13			
2	0.2012	ES08 ES10			
2	0.1999	ES01 ES09			
2	0.1922	ES07 ES11			
2	0.1872	ES03 ES09			
2	0.1871	ES05 ES11			

3	0.4501	ES09 ES10 ES11			
3	0.4050	ES09 ES10 ES13			
3	0.4035	ES09 ES10 ES12			
3	0.3846	ES09 ES11 ES12			
3	0.3830	ES02 ES09 ES10			
3	0.3801	ES09 ES11 ES13			
3	0.3756	ES01 ES09 ES10			
3	0.3745	ES03 ES09 ES10			
3	0.3734	ES02 ES09 ES11			
3	0.3707	ES03 ES09 ES11			

3	0.3704	ES05	ES09	ES10					
3	0.3703	ES08	ES09	ES11					
3	0.3694	ES08	ES09	ES10					
3	0.3694	ES06	ES09	ES10					

4	0.4637	ES05	ES08	ES10	ES11				
4	0.4618	ES06	ES08	ES10	ES11				
4	0.4527	ES02	ES09	ES10	ES11				
4	0.4523	ES03	ES09	ES10	ES11				
4	0.4518	ES04	ES08	ES10	ES11				
4	0.4511	ES08	ES09	ES10	ES11				
4	0.4504	ES09	ES10	ES11	ES13				
4	0.4503	ES06	ES09	ES10	ES11				
4	0.4503	ES04	ES09	ES10	ES11				
4	0.4503	ES05	ES09	ES10	ES11				
4	0.4502	ES09	ES10	ES11	ES14				
4	0.4501	ES09	ES10	ES11	ES12				
4	0.4501	ES01	ES09	ES10	ES11				
4	0.4501	ES07	ES09	ES10	ES11				

5	0.4861	ES06	ES08	ES10	ES11	ES13			
5	0.4825	ES03	ES08	ES10	ES11	ES14			
5	0.4784	ES05	ES08	ES10	ES11	ES13			
5	0.4771	ES03	ES08	ES09	ES10	ES11			
5	0.4767	ES05	ES08	ES09	ES10	ES11			
5	0.4754	ES06	ES08	ES09	ES10	ES11			
5	0.4739	ES05	ES08	ES10	ES11	ES14			
5	0.4717	ES02	ES08	ES09	ES10	ES11			
5	0.4712	ES06	ES08	ES10	ES11	ES12			
5	0.4712	ES03	ES04	ES09	ES10	ES11			
5	0.4710	ES04	ES08	ES09	ES10	ES11			
5	0.4699	ES06	ES08	ES10	ES11	ES14			
5	0.4698	ES04	ES08	ES10	ES11	ES14			
5	0.4681	ES05	ES08	ES10	ES11	ES12			

6	0.4996	ES06	ES08	ES10	ES11	ES12	ES13		
6	0.4992	ES03	ES08	ES10	ES11	ES13	ES14		
6	0.4990	ES04	ES08	ES10	ES11	ES12	ES13		
6	0.4981	ES06	ES08	ES10	ES11	ES13	ES14		
6	0.4953	ES06	ES08	ES09	ES10	ES11	ES13		
6	0.4945	ES05	ES08	ES10	ES11	ES13	ES14		
6	0.4919	ES05	ES08	ES10	ES11	ES12	ES13		
6	0.4904	ES05	ES08	ES09	ES10	ES11	ES13		
6	0.4899	ES03	ES08	ES09	ES10	ES11	ES14		
6	0.4890	ES04	ES08	ES10	ES11	ES13	ES14		
6	0.4889	ES03	ES07	ES08	ES10	ES11	ES13		
6	0.4886	ES03	ES08	ES10	ES11	ES12	ES14		
6	0.4874	ES01	ES06	ES08	ES10	ES11	ES13		
6	0.4872	ES06	ES07	ES08	ES10	ES11	ES13		

7	0.5191	ES01	ES06	ES08	ES10	ES11	ES12	ES13	
7	0.5140	ES04	ES08	ES09	ES10	ES11	ES12	ES13	
7	0.5117	ES06	ES08	ES09	ES10	ES11	ES12	ES13	
7	0.5103	ES03	ES08	ES10	ES11	ES12	ES13	ES14	
7	0.5100	ES03	ES08	ES09	ES10	ES11	ES12	ES13	
7	0.5100	ES04	ES08	ES10	ES11	ES12	ES13	ES14	
7	0.5097	ES02	ES06	ES08	ES10	ES11	ES12	ES13	
7	0.5089	ES02	ES04	ES08	ES10	ES11	ES12	ES13	
7	0.5073	ES04	ES05	ES06	ES08	ES10	ES12	ES13	
7	0.5071	ES01	ES05	ES08	ES10	ES11	ES12	ES13	
7	0.5071	ES05	ES08	ES09	ES10	ES11	ES12	ES13	
7	0.5070	ES06	ES08	ES10	ES11	ES12	ES13	ES14	
7	0.5067	ES03	ES07	ES08	ES10	ES11	ES12	ES13	
7	0.5064	ES01	ES04	ES08	ES10	ES11	ES12	ES13	

8	0.5595	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13
8	0.5399	ES02	ES06	ES07	ES08	ES10	ES11	ES12	ES13
8	0.5390	ES04	ES05	ES06	ES08	ES10	ES11	ES12	ES13
8	0.5373	ES04	ES05	ES06	ES08	ES09	ES10	ES12	ES13
8	0.5369	ES02	ES06	ES07	ES08	ES09	ES10	ES12	ES13
8	0.5311	ES03	ES05	ES07	ES08	ES09	ES10	ES12	ES13

8	0.5299	ES02	ES03	ES04	ES07	ES08	ES10	ES11	ES13					
8	0.5291	ES03	ES04	ES05	ES07	ES09	ES10	ES11	ES13					
8	0.5288	ES02	ES03	ES06	ES08	ES10	ES11	ES12	ES13					
8	0.5286	ES02	ES03	ES08	ES09	ES10	ES11	ES12	ES13					
8	0.5269	ES03	ES04	ES05	ES07	ES08	ES10	ES11	ES13					
8	0.5268	ES01	ES05	ES06	ES08	ES10	ES11	ES12	ES13					
8	0.5254	ES02	ES03	ES04	ES07	ES09	ES10	ES11	ES13					
8	0.5247	ES01	ES06	ES08	ES09	ES10	ES11	ES12	ES13					

9	0.5713	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
9	0.5701	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES12	ES13				
9	0.5700	ES02	ES03	ES04	ES06	ES08	ES10	ES11	ES12	ES13				
9	0.5670	ES03	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13				
9	0.5636	ES02	ES03	ES05	ES06	ES08	ES10	ES11	ES12	ES13				
9	0.5631	ES01	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13				
9	0.5629	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES12	ES13				
9	0.5623	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13	ES14				
9	0.5611	ES03	ES04	ES05	ES07	ES08	ES10	ES11	ES12	ES13				
9	0.5609	ES02	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13				
9	0.5586	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES12	ES13				
9	0.5574	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES12	ES13				
9	0.5568	ES02	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
9	0.5557	ES01	ES03	ES05	ES06	ES08	ES10	ES11	ES12	ES13				

10	0.5889	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
10	0.5858	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13			
10	0.5858	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
10	0.5806	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
10	0.5785	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES12	ES13	ES14			
10	0.5784	ES01	ES02	ES03	ES04	ES05	ES08	ES10	ES11	ES12	ES13			
10	0.5777	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES12	ES13	ES14			
10	0.5774	ES01	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13			
10	0.5771	ES01	ES02	ES03	ES04	ES06	ES08	ES10	ES11	ES12	ES13			
10	0.5764	ES03	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13	ES14			
10	0.5762	ES01	ES02	ES03	ES04	ES07	ES08	ES10	ES11	ES12	ES13			
10	0.5752	ES02	ES03	ES05	ES06	ES08	ES10	ES11	ES12	ES13	ES14			
10	0.5742	ES02	ES03	ES04	ES06	ES08	ES10	ES11	ES12	ES13	ES14			
10	0.5739	ES02	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13			

11	0.5980	ES01	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5974	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
11	0.5937	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
11	0.5935	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
11	0.5917	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5910	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5899	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
11	0.5889	ES01	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5884	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5883	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5883	ES01	ES02	ES03	ES04	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5874	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
11	0.5872	ES01	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
11	0.5871	ES01	ES02	ES03	ES04	ES05	ES08	ES09	ES10	ES11	ES12	ES13		

12	0.5996	ES01	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5988	ES01	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	
12	0.5984	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5982	ES01	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	
12	0.5981	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5975	ES01	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5970	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5962	ES01	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5960	ES01	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5921	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	
12	0.5920	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5920	ES01	ES02	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
12	0.5914	ES01	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	
12	0.5897	ES01	ES02	ES03	ES04	ES05	ES08	ES09	ES10	ES11	ES12	ES13	ES14	

13	0.5996	ES01	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14
13	0.5996	ES01	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14

13	0.5989	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	
13	0.5985	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5984	ES01	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5978	ES01	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5928	ES01	ES02	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5903	ES01	ES02	ES03	ES04	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5837	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13	ES14	
13	0.5806	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES12	ES13	ES14	
13	0.5626	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5514	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES13	ES14	
13	0.5352	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES14	
13	0.4462	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES11	ES12	ES13	ES14	

14	0.5996	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14

Full 14-descriptor model (MPN):

The SAS System 06:46 Friday, November 11, 2005 1

The REG Procedure
Model: MODEL1
Dependent Variable: pIC50

Number of Observations Read 66
Number of Observations Used 66

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	31.67883	2.26277	5.46	<.0001
Error	51	21.15270	0.41476		
Corrected Total	65	52.83152			

Root MSE 0.64402 R-Square 0.5996
Dependent Mean 6.76028 Adj R-Sq 0.4897
Coeff Var 9.52649

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-2.24897	31.74699	-0.07	0.9438
ES01	1	0.80056	2.16453	0.37	0.7130
ES02	1	59.50756	123.92791	0.48	0.6332
ES03	1	-138.98091	148.77248	-0.93	0.3546
ES04	1	123.61947	315.97755	0.39	0.6973
ES05	1	14.88885	325.97088	0.05	0.9637
ES06	1	-37.76386	34.60938	-1.09	0.2803
ES07	1	-0.34872	23.15511	-0.02	0.9880
ES08	1	16.51630	7.60475	2.17	0.0345
ES09	1	0.79803	0.55961	1.43	0.1599
ES10	1	-1.44211	0.32625	-4.42	<.0001
ES11	1	-0.44547	0.28607	-1.56	0.1256
ES12	1	16.84818	6.79830	2.48	0.0166
ES13	1	-44.34975	15.48467	-2.86	0.0061
ES14	1	-0.82715	2.64297	-0.31	0.7556

MPP data set

The SAS System

06:46 Friday, November 11, 2005 1

The REG Procedure

Model: MODEL1

Dependent Variable: pIC50

R-Square Selection Method

Number of Observations Read 66

Number of Observations Used 66

Number in Model	R-Square	Variables in Model
1	0.1686	ES11
1	0.1374	ES10
1	0.1361	ES09
1	0.0582	ES12
1	0.0173	ES13
1	0.0087	ES02
1	0.0082	ES03
1	0.0076	ES01
1	0.0027	ES05
1	0.0027	ES06
1	0.0025	ES04
1	0.0005	ES07
1	0.0001	ES14
1	0.0000	ES08

2	0.3621	ES09 ES11
2	0.3612	ES09 ES10
2	0.3434	ES09 ES12
2	0.3365	ES09 ES13
2	0.2548	ES08 ES09
2	0.2522	ES03 ES09
2	0.2480	ES02 ES09
2	0.2281	ES08 ES11
2	0.2276	ES05 ES09
2	0.2263	ES06 ES09
2	0.2198	ES04 ES09
2	0.2108	ES10 ES11
2	0.2097	ES11 ES13
2	0.2073	ES01 ES09

3	0.4497	ES09 ES10 ES11
3	0.4083	ES09 ES10 ES13
3	0.4043	ES09 ES10 ES12
3	0.3900	ES03 ES09 ES10
3	0.3886	ES02 ES09 ES10
3	0.3848	ES09 ES11 ES12
3	0.3816	ES09 ES11 ES13
3	0.3813	ES06 ES09 ES10
3	0.3801	ES05 ES09 ES10
3	0.3773	ES04 ES09 ES10
3	0.3767	ES01 ES09 ES10
3	0.3751	ES03 ES09 ES11
3	0.3724	ES02 ES09 ES11
3	0.3712	ES08 ES09 ES10

4	0.4525	ES03 ES09 ES10 ES11
4	0.4516	ES02 ES09 ES10 ES11
4	0.4505	ES08 ES09 ES10 ES11
4	0.4505	ES06 ES09 ES10 ES11
4	0.4503	ES05 ES09 ES10 ES11
4	0.4503	ES04 ES09 ES10 ES11
4	0.4503	ES09 ES10 ES11 ES13
4	0.4500	ES01 ES09 ES10 ES11
4	0.4499	ES09 ES10 ES11 ES14

4	0.4498	ES09 ES10 ES11 ES12
4	0.4497	ES07 ES09 ES10 ES11
4	0.4419	ES04 ES08 ES10 ES11
4	0.4341	ES07 ES08 ES10 ES11
4	0.4334	ES08 ES09 ES10 ES13

5	0.4723	ES04 ES08 ES09 ES10 ES11
5	0.4716	ES04 ES08 ES10 ES11 ES14
5	0.4673	ES01 ES06 ES08 ES10 ES11
5	0.4661	ES01 ES05 ES08 ES10 ES11
5	0.4649	ES05 ES08 ES09 ES10 ES11
5	0.4639	ES07 ES08 ES10 ES11 ES13
5	0.4631	ES03 ES08 ES09 ES10 ES11
5	0.4609	ES02 ES05 ES08 ES10 ES11
5	0.4607	ES03 ES06 ES09 ES10 ES11
5	0.4596	ES08 ES09 ES10 ES12 ES13
5	0.4593	ES06 ES08 ES09 ES10 ES11
5	0.4590	ES03 ES08 ES10 ES11 ES14
5	0.4585	ES03 ES05 ES09 ES10 ES11
5	0.4584	ES08 ES09 ES10 ES11 ES14

6	0.4907	ES01 ES06 ES08 ES10 ES11 ES13
6	0.4902	ES04 ES08 ES10 ES11 ES13 ES14
6	0.4896	ES03 ES05 ES07 ES09 ES10 ES11
6	0.4874	ES01 ES02 ES08 ES10 ES11 ES14
6	0.4848	ES01 ES03 ES08 ES10 ES11 ES14
6	0.4818	ES04 ES08 ES09 ES10 ES11 ES13
6	0.4817	ES02 ES03 ES08 ES10 ES11 ES14
6	0.4817	ES01 ES02 ES08 ES09 ES10 ES11
6	0.4811	ES04 ES08 ES10 ES11 ES12 ES14
6	0.4805	ES01 ES05 ES08 ES10 ES11 ES13
6	0.4804	ES04 ES08 ES09 ES10 ES11 ES14
6	0.4792	ES04 ES07 ES08 ES10 ES11 ES13
6	0.4791	ES02 ES05 ES08 ES10 ES11 ES13
6	0.4790	ES03 ES08 ES09 ES10 ES11 ES14

7	0.5248	ES01 ES06 ES08 ES10 ES11 ES12 ES13
7	0.5123	ES01 ES05 ES08 ES10 ES11 ES12 ES13
7	0.5119	ES03 ES05 ES07 ES08 ES10 ES11 ES13
7	0.5100	ES01 ES04 ES08 ES10 ES11 ES12 ES13
7	0.5097	ES02 ES04 ES08 ES10 ES11 ES12 ES13
7	0.5074	ES02 ES03 ES04 ES07 ES09 ES10 ES11
7	0.5069	ES03 ES05 ES07 ES09 ES10 ES11 ES14
7	0.5068	ES02 ES05 ES08 ES10 ES11 ES12 ES13
7	0.5045	ES03 ES05 ES07 ES09 ES10 ES11 ES13
7	0.5028	ES02 ES03 ES04 ES05 ES09 ES10 ES11
7	0.5027	ES01 ES02 ES08 ES10 ES11 ES13 ES14
7	0.5015	ES01 ES06 ES08 ES10 ES11 ES13 ES14
7	0.5012	ES02 ES03 ES08 ES10 ES11 ES12 ES13
7	0.5011	ES03 ES04 ES08 ES10 ES11 ES12 ES13

8	0.5608	ES03 ES05 ES07 ES08 ES10 ES11 ES12 ES13
8	0.5394	ES03 ES05 ES07 ES09 ES10 ES11 ES12 ES13
8	0.5389	ES03 ES05 ES07 ES08 ES09 ES10 ES12 ES13
8	0.5353	ES01 ES05 ES06 ES08 ES10 ES11 ES12 ES13
8	0.5352	ES02 ES06 ES07 ES08 ES09 ES10 ES12 ES13
8	0.5350	ES02 ES03 ES04 ES07 ES08 ES10 ES11 ES13
8	0.5346	ES04 ES05 ES06 ES08 ES09 ES10 ES12 ES13
8	0.5327	ES02 ES03 ES04 ES07 ES09 ES10 ES11 ES13
8	0.5312	ES03 ES05 ES07 ES08 ES10 ES11 ES13 ES14
8	0.5297	ES01 ES06 ES08 ES09 ES10 ES11 ES12 ES13
8	0.5294	ES02 ES03 ES06 ES08 ES10 ES11 ES12 ES13
8	0.5284	ES02 ES03 ES08 ES09 ES10 ES11 ES12 ES13
8	0.5274	ES01 ES04 ES06 ES08 ES10 ES11 ES12 ES13
8	0.5258	ES01 ES06 ES07 ES08 ES10 ES11 ES12 ES13

9	0.5792	ES02 ES03 ES04 ES06 ES08 ES10 ES11 ES12 ES13
9	0.5762	ES03 ES05 ES07 ES08 ES09 ES10 ES11 ES12 ES13
9	0.5705	ES02 ES03 ES05 ES06 ES08 ES09 ES10 ES12 ES13
9	0.5666	ES03 ES05 ES07 ES08 ES10 ES11 ES12 ES13 ES14
9	0.5660	ES01 ES03 ES05 ES06 ES08 ES10 ES11 ES12 ES13

9	0.5650	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES12	ES13					
9	0.5640	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES12	ES13					
9	0.5627	ES02	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13					
9	0.5627	ES01	ES03	ES05	ES07	ES08	ES10	ES11	ES12	ES13					
9	0.5621	ES03	ES04	ES05	ES07	ES08	ES10	ES11	ES12	ES13					
9	0.5617	ES02	ES03	ES05	ES06	ES08	ES10	ES11	ES12	ES13					
9	0.5612	ES03	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13					
9	0.5592	ES02	ES03	ES04	ES07	ES08	ES10	ES11	ES12	ES13					
9	0.5584	ES01	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13					
<hr/>															
10	0.5987	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5889	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5870	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5839	ES02	ES03	ES04	ES06	ES08	ES10	ES11	ES12	ES13	ES14				
10	0.5837	ES01	ES03	ES05	ES06	ES08	ES10	ES11	ES12	ES13	ES14				
10	0.5798	ES02	ES03	ES04	ES05	ES06	ES08	ES10	ES11	ES12	ES13				
10	0.5797	ES02	ES03	ES04	ES06	ES07	ES08	ES10	ES11	ES12	ES13				
10	0.5795	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14				
10	0.5793	ES01	ES02	ES03	ES04	ES06	ES08	ES10	ES11	ES12	ES13				
10	0.5791	ES01	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5781	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES12	ES13	ES14				
10	0.5766	ES02	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5765	ES01	ES03	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
10	0.5765	ES03	ES04	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13				
<hr/>															
11	0.6030	ES01	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14			
11	0.6005	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14			
11	0.5994	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5993	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5991	ES01	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5953	ES01	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5930	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14			
11	0.5930	ES01	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14			
11	0.5928	ES01	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5920	ES01	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5919	ES01	ES02	ES04	ES05	ES06	ES08	ES09	ES10	ES12	ES13	ES14			
11	0.5911	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13			
11	0.5886	ES01	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES12	ES13	ES14			
11	0.5867	ES01	ES02	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13			
<hr/>															
12	0.6063	ES01	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6060	ES01	ES02	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6059	ES01	ES02	ES03	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6038	ES01	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6022	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6020	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.6007	ES01	ES02	ES03	ES04	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.5997	ES01	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.5994	ES01	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
12	0.5994	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
12	0.5993	ES01	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13		
12	0.5963	ES01	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
12	0.5961	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14		
12	0.5947	ES01	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13		
<hr/>															
13	0.6075	ES01	ES02	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.6065	ES01	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.6064	ES01	ES02	ES03	ES04	ES05	ES06	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.6060	ES01	ES02	ES03	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.6023	ES01	ES02	ES03	ES04	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.6023	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5994	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	
13	0.5961	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES12	ES13	ES14	
13	0.5864	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES10	ES11	ES12	ES13	ES14	
13	0.5814	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5801	ES01	ES02	ES03	ES04	ES05	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14	
13	0.5576	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES13	ES14	
13	0.5380	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES14	
13	0.4733	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES11	ES12	ES13	ES14	
<hr/>															
14	0.6075	ES01	ES02	ES03	ES04	ES05	ES06	ES07	ES08	ES09	ES10	ES11	ES12	ES13	ES14

Full 14-descriptor model (MPP):

The SAS System 06:46 Friday, November 11, 2005 1

The REG Procedure
 Model: MODEL1
 Dependent Variable: pIC50

Number of Observations Read 66
 Number of Observations Used 66

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	32.09560	2.29254	5.64	<.0001
Error	51	20.73592	0.40659		
Corrected Total	65	52.83152			

Root MSE 0.63764 R-Square 0.6075
 Dependent Mean 6.76028 Adj R-Sq 0.4998
 Coeff Var 9.43217

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.62171	32.17029	0.02	0.9847
ES01	1	-2.24219	2.71493	-0.83	0.4127
ES02	1	-52.25195	146.75501	-0.36	0.7233
ES03	1	-0.87495	174.54579	-0.01	0.9960
ES04	1	-159.47558	360.94932	-0.44	0.6605
ES05	1	298.04119	361.54374	0.82	0.4136
ES06	1	-60.56944	32.09727	-1.89	0.0649
ES07	1	-9.80022	25.46842	-0.38	0.7020
ES08	1	13.97188	7.58913	1.84	0.0714
ES09	1	0.89358	0.54008	1.65	0.1042
ES10	1	-1.34982	0.32326	-4.18	0.0001
ES11	1	-0.33413	0.27442	-1.22	0.2290
ES12	1	17.08932	6.71213	2.55	0.0140
ES13	1	-45.77831	15.23681	-3.00	0.0041
ES14	1	-2.91421	2.83976	-1.03	0.3096

APPENDIX I

PATH INFORMATION FOR RESEARCH FILES

All important files used in the course of this research are located in Dr. Venanzi's research area. The directories where these files can be found are given below.

For computer programs:

/afs/cad/research/chem/venanzi/3/dissertation/programs/

For files related to superposition-dependent cluster analyses:

/afs/cad/research/chem/venanzi/3/dissertation/frc-1/

For files related to superposition-dependent cluster analyses:

/afs/cad/research/chem/venanzi/3/dissertation/frc-2/

For files related to QSAR of MP analogues:

/afs/cad/research/chem/venanzi/3/dissertation/qsar/

REFERENCES

1. *National Drug Threat Assessment 2001 - The Domestic Perspective*, National Drug Intelligence Center, U.S. Department of Justice, Report 2001-SQ0317-001, October, 2000.
2. Kuhar, M. J.; Ritz, M. C.; Boja, J. W. The dopamine hypothesis of the reinforcing properties of cocaine. *Trends in Neuroscience* **1991**, *14*, 299-302.
3. Giros, B.; Jaber, M.; Jones, S. R.; Wightman, R. M.; Caron, M. G. Hyperlocomotion and indifference to cocaine and amphetamine in mice lacking the dopamine transporter. *Nature* **1996**, *379*, 606-612.
4. Rothman, R. B.; Mele, A.; Reid, A. A.; Akunne, H.; Greig, N.; Thurkauf, A.; Rice, K. C.; Pert, A. Tight binding dopamine reuptake inhibitors as cocaine antagonists: A strategy for drug development. *FEBS Letters* **1989**, *257*, 341-344.
5. Rothman, R. B. High affinity dopamine reuptake inhibitors as potential cocaine antagonists: A strategy for drug development. *Life Sciences* **1990**, *46*, PL17-PL21.
6. Singh, S. Chemistry, design, and structure-activity relationship of cocaine antagonists. *Chemical Reviews* **2000**, *100*, 925-1024.
7. Carroll, F. I.; Gao, Y.; Rahman, M. A.; Abrams, P.; Parham, K.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Synthesis, ligand binding, QSAR, and CoMFA study of 3 β -(*p*-substituted phenyl)tropane-2 β -carboxylic acid methyl esters. *Journal of Medicinal Chemistry* **1991**, *34*, 2719-2725.
8. Carroll, F. I.; Mascarella, S. W.; Kuzemko, M. A.; Gao, Y.; Abraham, P.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Synthesis, ligand binding, and QSAR (CoMFA and classical) study of 3 β -(3'-substituted phenyl)-, 3 β -(4'-substituted phenyl)-, and 3 β -(3',4'-disubstituted phenyl)tropane-2 β -carboxylic acid methyl esters. *Journal of Medicinal Chemistry* **1994**, *37*, 2865-2873.
9. Yang, B.; Wright, J.; Eldefrawi, M. E.; Pou, S.; MacKerell Jr., A. D. Conformational, aqueous solvation, and pK_a contributions to the binding and activity of cocaine, WIN 32065-2, and the WIN vinyl analog. *Journal of the American Chemical Society* **1994**, *116*, 8722-8732.
10. Lieske, S. F.; Yang, B.; Eldefrawi, M. E.; MacKerell Jr., A. D.; Wright, J. (-)-3 β -Substituted ecgonine methyl esters as inhibitors for cocaine binding and dopamine uptake. *Journal of Medicinal Chemistry* **1998**, *41*, 864-876.

11. Zhu, N.; Harrison, A.; Trudell, M. L.; Klein-Stevens, C. L. QSAR and CoMFA study of cocaine analogs: Crystal and molecular structure of (-)-cocaine hydrochloride and N-methyl-3 β -(*p*-fluorophenyl)tropane-2 β -carboxylic acid methyl ester. *Structural Chemistry* **1999**, *10*, 91-103.
12. Muszynski, I. C.; Scapozza, L.; Kovar, K.-A.; Folkers, G. Quantitative structure-activity relationships of phenyltropanes as inhibitors of three monoamine transporters: Classical and CoMFA studies. *Quantitative Structure-Activity Relationships* **1999**, *18*, 342-353.
13. Davies, H. M. L.; Gilliatt, V.; Kuhn, L. A.; Saikali, E.; Ren, P.; Hammond, P. S.; Sexton, G. J.; Childers, S. R. Synthesis of 2 β -acyl-3 β -(substituted naphthyl)-8-azabicyclo[3.2.1] octanes and their binding affinities at dopamine and serotonin transport sites. *Journal of Medicinal Chemistry* **2001**, *44*, 1509-1515.
14. Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn-Z descriptors. *Journal of Medicinal Chemistry* **2000**, *43*, 4151-4159.
15. Paula, S.; Tabet, M. R.; Keenan, S. M.; Welsh, W. J.; Ball Jr., W. J. Three-dimensional structure-activity relationship modeling of cocaine binding to two monoclonal antibodies by Comparative Molecular Field Analysis. *Journal of Molecular Biology* **2003**, *325*, 515-530.
16. Paula, S.; Tabet, M. R.; Farr, C. D.; Norman, A. D.; Ball Jr., W. J. Three-dimensional quantitative structure-activity relationship modeling of cocaine binding by a novel human monoclonal antibody. *Journal of Medicinal Chemistry* **2004**, *47*, 133-142.
17. Newman, A. H.; Izenwasser, S.; Robarge, M. J.; Kline, R. H. CoMFA study of novel phenyl ring-substituted 3- α -(diphenylmethoxy)tropane analogues at the dopamine transporter. *Journal of Medicinal Chemistry* **1999**, *42*, 3502-3509.
18. Robarge, M. J.; Agoston, G. E.; Izenwasser, S.; Kopajtic, T.; George, C.; Katz, J. L.; Newman, A. H. Highly selective chiral N-substituted 3 α -[bis(4'-fluorophenyl)methoxy]tropane analogues for the dopamine transporter: Synthesis and Comparative Molecular Field Analysis. *Journal of Medicinal Chemistry* **2000**, *43*, 1085-1093.
19. Froimowitz, M.; Wu, K.-M.; Rodrigo, J.; George, C. Conformational preferences of the potent dopamine reuptake blocker BTCP and its analogs and their incorporation into a pharmacophore model. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 135-146.
20. Kulkarni, S. S.; Newman, A. H.; Houlihan, W. J. Three-dimensional quantitative structure-activity relationships of mazindol analogues at the dopamine transporter. *Journal of Medicinal Chemistry* **2002**, *45*, 4119-4127.

21. Froimowitz, M.; Patrick, K. S.; Cody, V. Conformational analysis of methylphenidate and its structural relationship to other dopamine reuptake blockers such as CFT. *Pharmaceutical Research* **1995**, *12*, 1430-1434.
22. Misra, M. Comparative Molecular Field Analysis (CoMFA) of phenyl ring substituted methylphenidates. Master's Thesis, New Jersey Institute of Technology, Newark, NJ, 1999.
23. Gilbert, K. M. Comparative Molecular Field Analysis (CoMFA) of protonated methylphenidate phenyl-substituted analogs. Master's Thesis, New Jersey Institute of Technology, Newark, NJ, 2002.
24. Gilbert, K. M.; Skawinski, W. J.; Misra, M.; Paris, K. A.; Naik, N. H.; Buono, R. A.; Deutsch, H. M.; Venanzi, C. A. Conformational analysis of methylphenidate: comparison of molecular orbital and molecular mechanics methods. *Journal of Computer-Aided Molecular Design* **2004**, *18*, 719-738.
25. Venanzi, C. A.; Misra, M.; Gilbert, K. M.; Buono, R. A.; Schweri, M. M.; Shi, Q.; Kim, D. I.; Deutsch, H. M. Comparative Molecular Field Analysis (CoMFA) of methylphenidate analogs with phenyl ring substituents. **Manuscript in preparation.**
26. Benedetti, P.; Mannhold, R.; Cruciani, G.; Pastor, M. GBR compounds and mepyraines as cocaine abuse therapeutics: Chemometric studies on selectivity using grid independent descriptors. *Journal of Medicinal Chemistry* **2002**, *45*, 1577-1584.
27. Misra, M.; Banerjee, A.; Davé, R. N.; Venanzi, C. A. Novel feature extraction technique for fuzzy relational clustering of a flexible dopamine reuptake inhibitor. *Journal of Chemical Information and Modeling* **2005**, *45*, 610-623.
28. Banerjee, A.; Misra, M.; Pai, D.; Woodley, R.; Davé, R. N.; Shih, L.-Y.; Lu, X.-J.; Srinivasan, A. R.; Olson, W. K.; Venanzi, C. A. Feature extraction using molecular planes parameters for fuzzy relational clustering of a flexible dopamine reuptake inhibitor. *Journal of Chemical Information and Modeling* **Submitted.**
29. Gilbert, K. M.; Venanzi, C. A. Hierarchical clustering analysis of flexible GBR 12909 dialkyl piperazine and piperidine analogs. *Journal of Computational Chemistry* **Submitted.**
30. Gilbert, K. M.; Boos, T. L.; Greiner, E.; Jacobson, A. E.; Lewis, D.; Matecka, D.; Prinszano, T.; Zhang, Y.; Rothman, R. B.; Rice, K. C. CoMFA and CoMSIA studies of the DAT/SERT selectivity of GBR 12909 analogs. *Journal of Medicinal Chemistry* **Submitted.**
31. Wang, S.; Sakamuri, S.; Enyedy, I. J.; Kozikowski, A. P.; Zaman, W. A.; Johnson, K. M. Molecular modeling, structure-activity relationships and functional antagonism studies of 4-hydroxy-1-methyl-4-(4-methylphenyl)-3-piperidyl 4-

- methylphenyl ketones as a novel class of dopamine transporter inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2001**, *9*, 1753-1764.
32. Volkow, N. D.; Ding, Y.-S.; Fowler, J. S.; Wang, G.-J.; Logan, J.; Gatley, S. J.; Dewey, S.; Ashby, C.; Liebermann, J.; Hintzemann, R.; Wolf, A. P. Is methylphenidate like cocaine? Studies on their pharmacokinetics and distribution in the human brain. *Archives of General Psychiatry* **1995**, *52*, 456-463.
 33. Ding, Y.-S.; Fowler, J. S.; Volkow, N. D.; Logan, J.; Gatley, S. J.; Sugano, Y. Carbon-11-*d-threo*-methylphenidate binding to dopamine transporter in baboon brain. *Journal of Nuclear Medicine* **1995**, *36*, 2298-2305.
 34. Schweri, M. M.; Skolnick, P.; Rafferty, M. F.; Rice, K. C.; Janowsky, A. J.; Paul, S. M. [³H]*Threo*-(+)-methylphenidate binding to 3,4-dihydroxyphenylethylamine uptake sites in corpus striatum: Correlation with the stimulant properties of ritalinic acid esters. *Journal of Neurochemistry* **1985**, *45*, 1062-1070.
 35. Schweri, M. M. N-ethylmaleimide irreversibly inhibits the binding of [³H]*threo*-(+)-methylphenidate to the stimulant recognition site. *Neuropharmacology* **1990**, *29*, 901-908.
 36. Schweri, M. M. Mercuric chloride and p-chloromercuriphenylsulfonate exert a biphasic effect on the binding of the stimulant [³H]methylphenidate to the dopamine transporter. *Synapse* **1994**, *16*, (188-194).
 37. Deutsch, H. M.; Shi, Q.; Gruszecka-Kowalik, E.; Schweri, M. M. Synthesis and pharmacology of potential cocaine antagonists. 2. Structure-activity relationship studies of aromatic ring-substituted methylphenidate analogs. *Journal of Medicinal Chemistry* **1996**, *39*, 1201-1209.
 38. Froimowitz, M.; Deutsch, H. M.; Shi, Q.; Wu, K.-M.; Glaser, R.; Adin, I.; George, C.; Schweri, M. M. Further evidence for a dopamine reuptake pharmacophore. The effect of N-methylation on *threo*-methylphenidate and its analogs. *Bioorganic & Medicinal Chemistry Letters* **1997**, *7*, 1213-1218.
 39. Deutsch, H. M. Structure-activity relationships for methylphenidate analogs and comparisons to cocaine and tropanes. *Medicinal Chemistry Research* **1998**, *8*, 91-99.
 40. Glaser, R.; Adin, I.; Shiftan, D.; Shi, Q.; Deutsch, H. M.; George, C.; Wu, K.-M.; Froimowitz, M. Solution and solid-state conformational and structural analysis of the N-methyl derivatives of (+)-*threo*-methylphenidate, (+)-*erythro*-methylphenidate and (+)-*threo*-p-methyl-methylphenidate HCl salts. *Journal of Organic Chemistry* **1998**, *63*, 1785-1794.
 41. Wayment, H. K.; Deutsch, H. M.; Schweri, M. M.; Schenk, J. O. Effects of methylphenidate analogues on phenethylamine substrates for the striatal

- dopamine transporter: Potential amphetamine antagonists? *Journal of Neurochemistry* **1999**, *72*, 1266-1274.
42. Deutsch, H. M.; Ye, X.; Shi, Q.; Liu, Z.; Schweri, M. M. Synthesis and pharmacology of site specific cocaine abuse treatment agents: A new synthetic methodology for methylphenidate analogs based on the Blaise Reaction. *European Journal of Medicinal Chemistry* **2001**, *36*, 303-311.
 43. Schweri, M. M.; Deutsch, H. M.; Massey, A. T.; Holtzman, S. G. Biochemical and behavioral characterization of novel methylphenidate analogs. *Journal of Pharmacology and Experimental Therapeutics* **2002**, *301*, 527-535.
 44. Deutsch, H. M.; Kim, D. I.; Holtzman, S. G.; Schweri, M. M.; Spealman, R. D. The synthesis and evaluation of new methylphenidates: Restricted rotation analogs, preliminary results. In *College on the Problems of Drug Dependence, 64th Annual Meeting*, Quebec City, Canada, 2002.
 45. Prisinzano, T.; Rice, K. C.; Baumann, M. H.; Rothman, R. B. Development of neurochemical normalization ("agonist substitution") therapeutics for stimulant abuse: Focus on the dopamine uptake inhibitor, GBR12909. *Current Medicinal Chemistry CNS Agents* **2004**, *4*, 47-59.
 46. Glowa, J. R.; Fantegrossi, W. E.; Lewis, D. B.; Matecka, D.; Rice, K. C.; Rothman, R. B. Sustained decrease in cocaine-maintained responding in rhesus monkeys with 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-hydroxy-3-phenylpropyl)piperazinyl decanoate, a long-acting ester derivative of GBR 12909. *Journal of Medicinal Chemistry* **1996**, *39*, 4689-4691.
 47. Lewis, D. B.; Matecka, D.; Zhang, Y.; Hsin, L. W.; Dersch, C. M.; Stafford, D.; Glowa, J. R.; Rothman, R. B.; Rice, K. C. Oxygenated analogues of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909) as potential extended-action cocaine-abuse therapeutic agents. *Journal of Medicinal Chemistry* **1999**, *42*, 5029-5042.
 48. Dutta, A. K.; Meltzer, P. C.; Madras, B. K. Positional importance of the nitrogen atom in novel piperidine analogs of GBR 12909: Affinity and selectivity for the dopamine transporter. *Medicinal Chemistry Research* **1993**, *3*, 209-222.
 49. Prisinzano, T.; Greiner, E.; Johnson II, E. M.; Dersch, C. M.; Marcus, J.; Partilla, J. S.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Piperidine analogues of GBR 12909: High affinity ligands for the dopamine transporter. *Journal of Medicinal Chemistry* **2002**, *45*, 4371-4374.
 50. Greiner, E.; Prisinzano, T.; Johnson II, E. M.; Dersch, C. M.; Marcus, J.; Partilla, J. S.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Structure-activity relationship studies of highly selective inhibitors of the dopamine transporter: N-

- benzylpiperidine analogues of [1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine. *Journal of Medicinal Chemistry* **2003**, *46*, 1465-1469.
51. Kolhatkar, R.; Cook, C. D.; Ghorai, S. K.; Deschamps, J.; Beardsley, P. M.; Reith, M. E. A.; Dutta, A. K. Further structurally constrained analogues of *cis*-(6-benzhydrylpiperidin-3-yl)benzylamine with elucidation of bioactive conformations: Discovery of 1,4-diazabicyclo[3.3.1]nonane derivatives and evaluation of their biological properties for the monoamine transporters. *Journal of Medicinal Chemistry* **2004**, *47*, 5101-5113.
 52. Dutta, A. K.; Xu, C.; Reith, M. E. A. Structure-activity relationship studies of novel 4-[2-[bis(4-fluorophenyl)methoxy]ethyl]-1-(3-phenylpropyl)piperidine analogs: Synthesis and biological evaluation at the dopamine and serotonin transporter sites. *Journal of Medicinal Chemistry* **1996**, *39*, 749-756.
 53. Dutta, A. K.; Reith, M. E. A.; Madras, B. K. Synthesis and preliminary characterization of a high-affinity novel radioligand for the dopamine transporter. *Synapse* **2001**, *39*, 175-181.
 54. Giros, B.; El Mestikawy, S.; Bertrand, L.; Caron, M. G. Cloning and functional characterization of a cocaine-sensitive dopamine transporter. *FEBS Letters* **1991**, *295*, 149-154.
 55. Chen, N. H.; Reith, M. E. A. Structure-function relationships for biogenic amine neurotransmitter transporters. In *Neurotransmitter Transporters: Structure, Function, and Regulation*; 2nd ed.; Reith, M. E. A., Ed.; Humana Press: Totowa, NJ, 2002; pp. 53-109.
 56. Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorganic & Medicinal Chemistry* **1995**, *3*, 411-428.
 57. Veith, M.; Hirst, J. D.; Brooks III, C. L. Do active site conformations of small ligands correspond to low free-energy solution structures? *Journal of Computer-Aided Molecular Design* **1998**, *12*, 563-572.
 58. Debnath, A. K. Three-dimensional quantitative structure-activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: Application of Comparative Molecular Field Analysis. *Journal of Medicinal Chemistry* **1999**, *42*, 249-259.
 59. Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry* **2004**, *47*, 2499-2510.
 60. Guarnieri, F.; Weinstein, H. Conformational memories and the exploration of biologically relevant peptide conformations: An illustration for the gonadotropin-

- releasing hormone. *Journal of the American Chemical Society* **1996**, *118*, 5580-5589.
61. Hopfinger, A. J.; Tokarski, J. S. Three-dimensional quantitative structure-activity relation analysis. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; pp. 105-164.
 62. Barnett-Norris, J.; Guarnieri, F.; Hurst, D. P.; Reggio, P. H. Exploration of biologically relevant conformations of anandamide, 2-arachidonylglycerol, and their analogs using conformational memories. *Journal of Medicinal Chemistry* **1998**, *41*, 4861-4872.
 63. Barnett-Norris, J.; Hurst, D. P.; Lynch, D. L.; Guarnieri, F.; Makriyanis, A.; Reggio, P. H. Conformational memories and the endocannabinoid binding site at the cannabinoid CB1 receptor. *Journal of Medicinal Chemistry* **2002**, *45*, 3649-3659.
 64. Greenidge, P. A.; Merette, S. A.; Beck, R.; Dodson, G.; Goodwin, C. A.; Scully, M. F.; Spencer, J.; Weiser, J.; Deadman, J. J. Generation of ligand conformations in continuum solvent consistent with protein active site topology: Application to thrombin. *Journal of Medicinal Chemistry* **2003**, *46*, 1293-1305.
 65. Bernard, D.; Coop, A.; MacKerell Jr., A. D. 2D conformationally sampled pharmacophore: A ligand based pharmacophore to differentiate delta opioid agonists from antagonists. *Journal of the American Chemical Society* **2003**, *125*, 3101-3107.
 66. Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988**, *110*, 5959-5967.
 67. Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 2002; Vol. 18, pp. 1-40.
 68. Davé, R. N.; Sen, S. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems* **2002**, *10*, 713-727.
 69. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach: Amsterdam, 1999.
 70. Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *Journal of Computational Chemistry* **1994**, *15*, 899-916.
 71. Murray-Rust, P.; Raftery, J. Computer analysis of molecular geometry, part IV: Classification of differences in conformation. *Journal of Molecular Graphics* **1985**, *3*, 50-59.
 72. Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*, Prentice Hall: NJ, 1988.

73. Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*, 4th ed.; Prentice-Hall: NJ, 1998.
74. Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; John Wiley: New York, 2000.
75. Chema, D.; Goldblum, A. The nearest neighbor method - Finding families of conformations within a sample. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 208-217.
76. Feher, M.; Schmidt, J. M. Metric and multidimensional scaling: Efficient tools for clustering molecular conformations. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 346-353.
77. Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 810-818.
78. Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press: New York, 1981.
79. Bezdek, J. C.; Keller, J.; Krishnapuram, R.; Pal, N. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers: Norwell, MA, 1999.
80. Davé, R. N.; Bhaswan, K. Adaptive fuzzy C-shells clustering and detection of ellipses. *IEEE Transactions on Neural Networks* **1992**, *3*, 643-662.
81. Davé, R. N.; Krishnapuram, R. Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems* **1997**, *5*, 270-293.
82. Xie, X. L.; Beni, G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1991**, *13*, 841-847.
83. Davé, R. N. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters* **1996**, *17*, 613-623.
84. Tucker, W. T. Counterexamples to the convergence theorem for the fuzzy c-means clustering algorithms. In *Analysis of Fuzzy Information*; Bezdek, J. C., Ed.; CRC Press: Boca Raton, FL, 1987; Vol. III, pp. 109-121.
85. Höppner, F.; Klawonn, F. Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression. *Proc. of the 4th Int. Conf. on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES)*: Brighton, U.K., 2000; pp. 162-165.
86. Höppner, F.; Klawonn, F.; Eklund, P. Learning indistinguishability from data. *Soft Computing Journal* **2002**, *6*, 6-13.

87. Hathaway, R. J.; Bezdek, J. C. NERF c-means: Non-Euclidean relational fuzzy clustering. *Pattern Recognition* **1994**, *24*, 429-437.
88. Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **1948**, *27*, 379-423.
89. Gath, I.; Geva, A. B. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1989**, *11*, 773-781.
90. Pai, D. Analysis of Molecular Conformations using Relative Planes. Master's Thesis, New Jersey Institute of Technology, Newark, NJ, 2004.
91. Lu, X.-J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research* **2003**, *31*, 5108-5121.
92. el Hassan, M. A.; Calladine, C. R. The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *Journal of Molecular Biology* **1995**, *251*, 648-664.
93. Lu, X.-J.; el Hassan, M. A.; Hunter, C. A. Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *Journal of Molecular Biology* **1997**, *273*, 668-680.
94. Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. B-DNA twisting correlates with base-pair morphology. *Journal of Molecular Biology* **1995**, *247*, 34-48.
95. Kosikov, K. M.; Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. DNA stretching and compression: large scale simulations of double helical structures. *Journal of Molecular Biology* **1999**, *289*, 1301-1326.
96. Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *Journal of Biomolecular Structure and Dynamics* **1988**, *6*, 63-91.
97. Lavery, R.; Sklenar, H. Defining the structure of irregular nucleic acids: conventions and principles. *Journal of Biomolecular Structure and Dynamics* **1989**, *6*, 655-667.
98. Dickerson, R. E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Research* **1998**, *26*, 1906-1926.
99. Soumpasis, D. M.; Tung, C.-S. A rigorous base-pair oriented description of DNA structures. *Journal of Biomolecular Structure and Dynamics* **1988**, *6*, 397-420.
100. Tung, C.-S.; Soumpasis, D. M.; Hummer, G. An extension of the rigorous base-unit oriented description of nucleic acid structures. *Journal of Biomolecular Structure and Dynamics* **1994**, *11*, 1327-1344.

101. Bansal, M.; Bhattacharyya, D.; Ravi, B. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Computer Applications in the Biosciences* **1995**, *11*, 281-287.
102. Pednault, E. P. D.; Babcock, M. S.; Olson, W. K. Nucleic acids structure analysis: a users guide to a collection of new analysis programs. *Journal of Biomolecular Structure and Dynamics* **1993**, *11*, 597-628.
103. Babcock, M. S.; Olson, W. K. The effect of mathematics and coordinate system on comparability and "dependencies" of nucleic acid structure parameters. *Journal of Molecular Biology* **1994**, *237*, 98-124.
104. Babcock, M. S.; Pednault, E. P. D.; Olson, W. K. Nucleic acid structure analysis: Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *Journal of Molecular Biology* **1994**, *237*, 125-156.
105. Lu, X.-J.; Babcock, M. S.; Olson, W. K. Overview of nucleic acid analysis programs. *Journal of Biomolecular Structure and Dynamics* **1999**, *16*, 833-843.
106. Lu, X.-J.; Olson, W. K. Resolving the discrepancies among nucleic acid conformational analyses. *Journal of Molecular Biology* **1999**, *285*, 1563-1575.
107. Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X.-J.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C.-S.; Westhof, E.; Wolberger, C.; Berman, H. M. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology* **2001**, *313*, 229-237.
108. SYBYL 6.9 Tripos Inc.; 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
109. Powell, M. J. D. Restart procedures for the conjugate gradient method. *Mathematical Programming* **1977**, *12*, 241-254.
110. Clark, M.; Cramer III, R. D.; van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *Journal of Computational Chemistry* **1989**, *10*, 982-1012.
111. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-3228.
112. Streitwieser, A. *Molecular Orbital Theory for Organic Chemists*, John Wiley & Sons Inc.: 1961.

113. Purcell, W. P.; Singer, J. A. A brief review and table of semiempirical parameters used in the Hückel molecular orbital method. *Journal of Chemical & Engineering Data* **1967**, *12*, 235-246.
114. Matecka, D.; Rothman, R. B.; Radesca, L.; de Costa, B. R.; Dersch, C. M.; Partilla, J. S.; Pert, A.; Glowa, J. R.; Wojnicki, F. H. E.; Rice, K. C. Development of novel, potent, and selective dopamine reuptake inhibitors through alteration of the piperazine ring of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909). *Journal of Medicinal Chemistry* **1996**, *39*, 4704-4716.
115. Matecka, D.; Lewis, D.; Rothman, R. B.; Dersch, C. M.; Wojnicki, F. H. E.; Glowa, J. R.; DeVries, A. C.; Pert, A.; Rice, K. C. Heteroatomic analogs of 1-[2-(diphenylmethoxy)ethyl]- and 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazines (GBR 12935 and GBR 12909) as high-affinity dopamine reuptake inhibitors. *Journal of Medicinal Chemistry* **1997**, *40*, 705-716.
116. Hsin, L. W.; Dersch, C. M.; Baumann, M. H.; Stafford, D.; Glowa, J. R.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Development of long-acting dopamine transporter ligands as potential cocaine-abuse therapeutic agents: Chiral hydroxyl-containing derivatives of 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(3-phenylpropyl)piperazine and 1-[2-(diphenylmethoxy)ethyl]-4-(3-phenylpropyl)piperazine. *Journal of Medicinal Chemistry* **2002**, *45*, 1321-1329.
117. Lewis, D.; Zhang, Y.; Prisinzano, T.; Dersch, C. M.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Further exploration of 1-{2-[bis-(4-fluorophenyl)methoxy]ethyl}piperazine (GBR 12909): Role of N-aromatic, N-heteroaromatic, and 3-oxygenated N-phenylpropyl substituents on affinity for the dopamine and serotonin transporter. *Bioorganic & Medicinal Chemistry Letters* **2003**, *13*, 1385-1389.
118. Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A* **1987**, *4*, 629-642.
119. Wishart, D. k-means clustering with outlier deletion, for data mining with mixed variables and missing values. In *Exploratory Data Analysis in Empirical Research*; Schwaiger, M., Opitz, O., Eds.; Springer: Berlin, 2002; pp. 216-226.
120. Saunders, M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum. *Journal of the American Chemical Society* **1987**, *109*, 3150-3152.
121. Fiorentino, A.; Pandit, D.; Gilbert, K. M.; Misra, M.; Dios, R.; Venanzi, C. A. Singular value decomposition of torsional angles of analogs of the dopamine reuptake inhibitor GBR 12909. *Journal of Computational Chemistry* **In press**.
122. Pandit, D.; Roosma, W. A.; Venanzi, C. A. **Manuscript in preparation.**

123. Rothman, R. B.; Lewis, B.; Dersch, C. M.; Xu, H.; Radesca, L.; de Costa, B. R.; Rice, K. C.; Kilburn, R. B.; Akunne, H. C.; Pert, A. Identification of a GBR12935 homolog, LR1111, which is over 4000-fold selective for the dopamine transporter, relative to serotonin and norepinephrine transporters. *Synapse* **1993**, *14*, 34-39.
124. Zhang, Y.; Rothman, R. B.; Dersch, C. M.; De Costa, B. R.; Jacobson, A. E.; Rice, K. C. Synthesis and transporter binding properties of bridged piperazine analogues of 1-{2-[bis(4-fluorophenyl)methoxy]ethyl}-4-(3-phenylpropyl)piperazine (GBR 12909). *Journal of Medicinal Chemistry* **2000**, *43*, 4840-4849.
125. Zhang, Y.; Joseph, D. B.; Bowen, W. D.; Flippen-Anderson, J. L.; Dersch, C. M.; Rothman, R. B.; Jacobson, A. E.; Rice, K. C. Synthesis and biological evaluation of tropane-like 1-{2-[bis(4-fluorophenyl)methoxy]ethyl}-4-(3-phenylpropyl)piperazine (GBR 12909) analogues. *Journal of Medicinal Chemistry* **2001**, *44*, 3937-3945.
126. Hsin, L. W.; Prisinzano, T.; Wilkerson, C. R.; Dersch, C. M.; Horel, R.; Jacobson, A. E.; Rothman, R. B.; Rice, K. C. Synthesis and dopamine transporter affinity of chiral 1-[2-[bis(4-fluorophenyl)methoxy]ethyl]-4-(2-hydroxypropyl)piperazines as potential cocaine abuse therapeutic agents. *Bioorganic & Medicinal Chemistry Letters* **2003**, *13*, 553-556.
127. MOE software available from Chemical Computing Group Inc.; 1010 Sherbrooke Street West, Suite 910, Montreal, Canada, H3A 2R7.
128. Randic, M. On characterization of molecular branching. *Journal of the American Chemical Society* **1975**, *97*, 6609-6615.
129. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press: New York, 1976.
130. Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*, Research Studies Press, John Wiley and Sons: Chichester, UK, 1986.
131. Kier, L. B. A shape index from molecular graphs. *Quantitative Structure-Activity Relationships* **1985**, *4*, 109-116.
132. Kier, L. B.; Hall, L. H. The kappa indices for modeling molecular shape and flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, UK, 1999.
133. Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*, Academic Press: San Diego, 1999.
134. Kier, L. B.; Hall, L. H. An atom-centered index for drug QSAR models. *Advances in Drug Research* **1992**, *22*, 1-38.

135. Kier, L. B.; Hall, L. H. An electrotopological state index for atoms in molecules. *Pharmaceutical Research* **1990**, *7*, 801-807.
136. Hall, L. H.; Mohnney, B. K.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Computer Sciences* **1991**, *31*, 76-82.
137. Hall, L. H.; Kier, L. B. An index of electrotopological state for atoms in molecules. *Journal of Mathematical Chemistry* **1991**, *7*, 229-241.
138. de Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with the electrotopological state indices: Corticosteroids. *Journal of Computer-Aided Molecular Design* **1998**, *12*, 557-561.
139. Halgren, T. A. Merck Molecular Force Field. I. Basis, form, scope, parameterization and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 490-519.
140. Neter, J.; Kutner, M. H.; Wasserman, W.; Nachtsheim, C. J. *Applied Linear Regression Models*, 3rd ed.; McGraw-Hill/Irwin: 1996.
141. Matlab Copyright 1994-2005 by The MathWorks Inc.; 3 Apple Hill Drive, Natick, MA 01760-2098 USA.
142. Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 357-369.
143. Golbraikh, A.; Shen, M.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 241-253.
144. Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspectives in Drug Discovery and Design* **1997**, *718*, 65-84.
145. Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modeling* **1997**, *15*, 372-385.
146. Willett, P.; Winterman, V. A comparison of some measures of inter-molecular structural similarity. *Quantitative Structure-Activity Relationships* **1986**, *5*, 18-25.
147. Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. Parameter based methods for compound selection from chemical databases. *Quantitative Structure-Activity Relationships* **1996**, *15*, 285-289.
148. Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.-D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of Ambergis Fragrance

- Compounds. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 582-595.
149. Johnson, M. A. A review and examination of the mathematical spaces underlying molecular similarity analysis. *Journal of Mathematical Chemistry* **1989**, *3*, 117-145.
150. Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbour searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences* **1986**, *26*, 36-41.
151. Dean, P. M. *Molecular Similarity in Drug Design*. Blackie Academic & Professional: Glasgow, 1995.
152. Willett, P. Using computational tools to analyse molecular diversity. In *Combinatorial Chemistry: A Short Course*; DeWitt, S. H., Czarnik, A. W., Eds.; American Chemical Society: Washington, 1997; pp. 17-48.
153. Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing chemical libraries for lead discovery. *Journal of Biomolecular Screening* **1996**, *1*, 65-73.
154. Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137-148.
155. Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry* **1997**, *40*, 1219-1229.
156. Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: a method for eliminating redundant variables. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 1160-1168.
157. SAS System for Windows version 9.1 Copyright 2005 SAS Institute Inc.; Cary, NC, USA.
158. Wold, H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*; Krishnaiah, P. R., Ed.; Academic Press: New York, 1966; pp. 391-420.
159. Wold, S.; Albano, C.; Dunn III, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. Multivariate Data Analysis in Chemistry. In *Chemometrics: Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; D. Reidel Publishing Company: Dordrecht, Holland, 1984.
160. Höskuldson, A. PLS regression models. *Journal of Chemometrics* **1988**, *2*, 211-228.

161. Frank, I. E.; Friedman, J. H. A statistical view of chemometrics regression tools. *Technometrics* **1993**, 35, 109-148.
162. Clark, R. D.; Sproun, D. G.; Leonard, J. M. Validating models based on large data sets. In *Rational Approaches to Drug Design: Proceedings of the 13th European Symposium on QSAR*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: 2001; pp. 475-485.