AUTOMATED DETECTION OF ANOMALOUS PATTERNS IN VALIDATION SCORES FOR PROTEIN X-RAY STRUCTURE MODELS

by

Eric David Williams

B.S. in Computer Engineering, University of Pittsburgh, 1999

M.S. in Electrical Engineering, University of Pittsburgh, 2001

M.S. in Intelligent Systems, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Intelligent Systems

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

Dietrich School of Arts and Sciences

This dissertation was presented

by

Eric Williams

It was defended on

June 21, 2013

and approved by

Gregory F. Cooper, Professor, Biomedical Informatics/Intelligent Systems

Shyam Visweswaran, Assistant Professor, Biomedical Informatics/Intelligent Systems

Xingha Lu, Associate Professor, Biomedical Informatics

Advisor: John M. Rosenberg, Professor, Biological Sciences/Intelligent Systems

AUTOMATED DETECTION OF ANOMALOUS PATTERNS IN VALIDATION SCORES FOR PROTEIN X-RAY STRUCTURE MODELS

Eric Williams, PhD

University of Pittsburgh, 2013

Structural bioinformatics is a subdomain of data mining focused on identifying structural patterns relevant to functional attributes in repositories of biological macromolecular structure models. This research focused on structures determined via x-ray crystallography and deposited in the Protein Data Bank (PDB).

Protein structures deposited in the PDB are products of experimental processes, and only approximately model physical reality. Structural biologists address accuracy and precision concerns via community-enforced consensus standards of accepted practice for proper building, refinement, and validation of models. Validation scores are quantitative partial indicators of the likelihood that a model contains serious systematic errors.

The PDB recently convened a panel of experts, which placed renewed emphasis on troubling anomalies among deposited structure models. This study set out to detect such anomalies. I hypothesized that community consensus standards would be evident in patterns of validation scores, and deviations from those standards would appear as unusual combinations of validation scores.

Validation attributes were extracted from PDB entry headers and multiple software tools (e.g., WhatCheck, SFCheck, and MolProbity). Independent component analysis (ICA) was used for attribute transformation to increase contrast between inliers and outliers. Unusual patterns were sought in regions of locally low density in the space of validation score profiles, using a novel standardization of Local Outlier Factor (LOF) scores.

Validation score profiles associated with the most extreme outlier scores were demonstrably anomalous according to domain theory. Among these were documented fabrications, possible annotation errors, and complications in the underlying experimental data. Analysis of deep inliers revealed promising support for the hypothesized link between consensus standard practices and common validation score values.

Unfortunately, with numerical anomaly detection methods that operate simultaneously on numerous continuous-valued attributes, it is often quite difficult to know *why* a case gets a particular outlier score. Therefore, I hypothesized that IF-THEN rules could be used to post-process outlier scores to make them comprehensible and explainable. Inductive rule extraction was performed using RIPPER. Results were mixed, but they represent a promising proof of concept.

The methods explored are general and applicable beyond this problem. Indeed, they could be used to detect structural anomalies using physical attributes.

TABLE OF CONTENTS

LIST OF EQUATIONS XX

LIST OF HYPOTHESES XXIII

ACKNOWLEDGEMENTS XXIV

1.0 INTRODUCTION 1

2.0 BACKGROUND AND SIGNFICANCE 3

2.1 STRUCTURAL BIOLOGY AND BIOINFORMATICS 3

2.1.1 Structural Bioinformatics 3

2.1.2 Structural Biology 4

2.1.3 Due Diligence in Structural Biology 5

2.1.4 Due Diligence in Structural Bioinformatics 6

2.1.5 Motivation for This Research 7

2.2 PROTEIN STRUCTURE DETERMINATION 8

2.2.1 Protein Crystallization 8

2.2.2 X-Ray Diffraction 9

2.2.3 Model Building and Refinement 11

2.3 STRUCTURE MODEL VALIDATION 12

2.3.1 What is validation? 12

2.3.2 What are common validation methods? 12

2.3.3 Diffraction Data Quality 13

2.3.4 Crystal Parameter Checks 14

2.3.5 Model Parameter Checks 15

2.3.6 Covalent Geometry Checks 17

2.3.7 Model-Data Agreement 17

2.3.8 Stereochemistry Checks 20

2.4 ANOMALIES 21

2.4.1 What are anomalies? Why are they important? 21

2.4.2 What are anomalies in structural bioinformatics? 22

2.4.3 Why is it important to identify anomalies in validation data? 23

2.4.4 Why are anomalies hard to detect? 23

2.4.5 Why is it hard to automate anomaly detection? 24

2.4.5.1 Known Knowns 25

2.4.5.2 Known Unknowns 25

2.4.5.3 Unknown Unknowns 25

2.4.5.4 Unknown Knowns 26

2.4.5.5 Algorithmic Difficulties 27

2.4.5.6 Data Difficulties 29

2.4.5.7 Curse(s) of Dimensionality 30

2.4.6 Has anyone attempted to find anomalies in the PDB before? 33

2.4.7 How did I get here? 36

2.5 A BRIEF HISTORY OF ANOMALOUS PROTEIN STRUCTURE MODELS 38

2.5.1 Early Difficulties 39

2.5.2 Concerns About Automation 40

2.5.3 Spectacular Retractions 41

2.5.4 Documented Frauds 42

3.0 THESIS 46

3.1 INTRODUCTION 46

3.2 CONSENSUS PRACTICES, COMMON VALIDATION SCORE PATTERNS, AND ANOMALIES 48

3.3 DETECTING DEVIATIONS FROM COMMON PATTERNS 49

4.0 DATA COLLECTION 51

4.1 BENCHMARK DATA 51

4.1.1 UCI Machine Learning Repository 51

4.1.2 Wine Quality 51

4.1.3 SECOM 52

4.1.4 Communities and Crime 52

4.1.5 Cardiotocography 53

4.2 PDB DATA 53

4.2.1 Protein Structure Models 53

4.2.2 Deposited Validation Scores and Descriptive Data 54

4.2.3 Verify3D 55

4.2.4 Errat 55

4.2.5 ProSA 56

4.2.6 TAP 57

4.2.7 MolProbity 58

4.2.8 ProCheck 59

4.2.9 SFCheck 60

4.2.10 WhatCheck 61

4.2.11 Constructed Validation Scores 62

4.2.11.1 Optical Resolution Estimation 63

4.2.11.2 B-factor (Atomic Displacement Parameter) Statistics 63

4.2.12 Missing Values 65

4.2.13 Missing Value Summaries and Complete Attribute List 66

5.0 OUTLIER SCORING 67

5.1 LOCAL OUTLIER FACTOR 67

5.2 OUTLIER SCORE STANDARDIZATION 69

5.2.1 Introduction 69

5.2.2 Probability Distribution Fitting 70

5.2.2.1 Published Fits to LOF 70

5.2.2.2 MLE Fits to LOF and LOF-1 73

5.2.2.3 MLE Fits to 1-1/LOF 79

5.3 OUTLIER SCORING EVALUATION CRITERIA 85

5.3.1 Loss Function 85

5.3.2 Expected Cost 86

5.3.3 Randomization Test 87

5.4 NEIGHBORHOOD SIZE EXPERIMENTS 88

5.4.1 Hypotheses 89

5.4.2 Methods 89

5.4.3 Results 89

5.5 ATTRIBUTE TRANSFORMATION AND DIMENSIONALITY REDUCTION EXPERIMENTS 90

5.5.1 Introduction 90

5.5.2 Hypotheses 93

5.5.3 Methods 93

5.5.3.1 Scaling 93

5.5.3.2 PCA 94

5.5.3.3 ICA 96

5.5.4 Results 101

5.6 SEMANTIC DIMENSIONALITY REDUCTION EXPERIMENT 102

5.6.1 Introduction 102

5.6.2 Hypothesis 103

5.6.3 Methods 104

5.6.4 Results 104

5.7 CONTEXTUALIZATION 105

5.7.1 Introduction 105

5.7.2 Contextual Attribute Preparation 106

5.7.3 Year of Deposition 106

5.7.4 Optical Resolution 108

5.7.5 Merging Contextual Scores 109

5.8 MAXIMAL OUTLIER ANALYSIS 112

5.8.1 Introduction 112

5.8.2 Conjecture 113

5.8.3 Methods ... 113

5.8.4 Results ... 114

5.8.4.1 Documented Fraud ... 116

5.8.4.2 Viruses With Misreported NCS ... 117

5.8.4.3 Possibly Problematic ... 120

5.8.4.4 Obvious Glitch ... 122

5.8.4.5 Strong Anisotropy ... 122

5.8.4.6 Unclear or Ambiguous ... 125

5.8.4.7 Exceptionally High Quality ... 129

5.9 DEEP INLIER ANALYSIS ... 129

5.9.1 Introduction ... 129

5.9.2 Hypothesis ... 131

5.9.3 Methods ... 131

5.9.4 Results ... 132

6.0 RULE EXTRACTION ... 134

6.1 INTRODUCTION ... 134

6.2 HYPOTHESIS ... 135

6.3 GENERAL METHODS ... 136

6.3.1 RIPPER ... 136

6.3.2 Rule Extraction Evaluation Criteria ... 137

6.4 OUTLIER SCORE EXPLANATION AND PREDICTION ... 138

6.4.1 Benchmark Experiment Methods ... 138

6.4.2 Benchmark Experiment Results ... 139

6.4.3 PDB Experiments Methods.. 143

6.4.3.1 Experiment 1 143

6.4.3.2 Experiment 2 144

6.4.3.3 Experiment 3 145

6.4.4 PDB Experiments Results 145

6.5 EXPLANATION OF DEEP INLIERS AND EXTREME OUTLIERS...... 151

6.5.1 Introduction 151

6.5.2 Methods 151

6.5.3 Results... 152

6.5.3.1 Deep Inliers... 152

6.5.3.2 Extreme Outliers 156

7.0 FUTURE WORK 164

7.1 LOCAL VALIDATION 164

7.2 TRANSFORMATION AND REDUCTION 165

7.3 K NEAREST NEIGHBORS 165

7.4 LOF 166

7.5 CONTEXTS 167

7.6 HUBS 168

7.7 ULTRAMETRICITY.. 168

8.0 SUMMARY AND CONCLUSIONS 170

8.1 PRIMARY RESULTS... 170

8.2 SECONDARY RESULTS... 171

8.3 CONTRIBUTION... 172

8.3.1 Novelty ... 173

8.3.2 Interestingness ... 173

8.3.3 Plausibility ... 174

8.3.4 Intelligibility ... 174

8.3.5 Broader Context and Applicability ... 174

APPENDIX A ... 175

BIBLIOGRAPHY ... 184

LIST OF TABLES

Table 1 Amoroso distribution parameter values for a sample of members of the family 75

Table 2 Goodness-of-fit scores for LOF using maximum likelihood estimation to fit parameters 77

Table 3 Goodness-of-fit scores for LOF-1 using maximum likelihood estimation to fit parameters 78

Table 4 Goodness-of-fit scores for 1-1/LOF using maximum likelihood estimation to fit parameters 80

Table 5 Expected costs for the unreduced ICA run for each benchmark set. The complete set of tables can be found in Appendix AError! Reference source not found.. 90

Table 6 Example of benchmark experiment results, showing MSE and expected costs for nonviolent crime at k=10 for each attribute transformation experiment. The complete set of tables can be found in Appendix A. 101

Table 7 Example of benchmark experiment results, showing MSE and expected costs for wine quality data in the unreduced ICA condition 102

Table 8 Counts of entries with outlier scores meeting or exceeding a series of threshold values 114

Table 9 Comparison of score distributions for potential hubs and all scored entries. 133

Table 10 Benchmark performance statistics for (0.9, 1.0) bin of 10-bin discretization 140

Table 11 Wine Quality performance statistics for 10-bin discretization, unreduced ICA condition, MinPts = 20... 140

Table 12 Simplified example rule predicting 10-bin outlier score for the wine quality data set 141

Table 13 Benchmark performance statistics for (0.67, 1.00) bin of 3-bin discretization 141

Table 14 Wine Quality performance statistics for 3-bin discretization, unreduced ICA condition, MinPts = 20 142

Table 15 Simplified example rule predicting 3-bin outlier score for the wine quality data set.. 142

Table 16 F-measure scores for cross-validated rule learning for PDB data. Asterisks indicate misleading values due to RIPPER's use of a default class. 146

Table 17 Performance statistics for rules learned from outlier scores based on the unreduced ICA condition (without dummy class).. 146

Table 18 Performance statistics for rules learned from outlier scores based on the unreduced ICA condition (with dummy class)... 147

Table 19 Number of attributes (in the improper subset) used for outlier scoring, number of rules extracted, and performance statistics for predictions. Overall accuracy refers to all predictions. Other statistics given are related specifically to predictions for the (0.8, 1.0) bin. 147

Table 20 Example of a (simplified) rule predicting outlier scores from the unreduced condition 149

Table 21 Predicted outlier score ranges for cases that were not scored by LOF 150

Table 22 Example of a (simplified) rule predicting outlier scores in the range [0.0, 0.2].......... 153

Table 23 Example of a (simplified) rule predicting outlier scores in the range [0.0, 0.1].......... 154

Table 24 Example of a (simplified) rule predicting outlier scores in the range [0.00, 0.05]...... 155

Table 25 Example of a (simplified) rule predicting outlier scores in the range [0.000, 0.025].. 156

Table 27 Example of a (simplified) rule predicting outlier scores in the range (0.8, 1.0].......... 158

Table 28 Example of a (simplified) rule predicting outlier scores in the range (0.9, 1.0].......... 159

Table 29 Example of a (simplified) rule predicting outlier scores in the range (0.90, 0.95]...... 160

Table 30 Example of a (simplified) rule predicting outlier scores in the range (0.95, 1.00]...... 161

Table 31 Example of a (simplified) rule predicting outlier scores in the range (0.950, 0.975].. 162

Table 32 Example of a (simplified) rule predicting outlier scores in the range (0.975, 1.000].. 163

Table 33 MSE and expected costs for wine quality data, control (scaled) condition................ 175

Table 34 MSE and expected costs for wine quality data, unreduced ICA condition 175

Table 35 MSE and expected costs for wine quality data, reduced PCA condition 176

Table 36 MSE and expected costs for wine quality data, reduced ICA condition 176

Table 37 MSE and expected costs for SECOM data, control (scaled) condition...................... 176

Table 38 MSE and expected costs for SECOM data, unreduced ICA condition 177

Table 39 MSE and expected costs for SECOM data, reduced PCA condition.......................... 177

Table 40 MSE and expected costs for SECOM data, reduced ICA condition 177

Table 41 MSE and expected costs for violent crime data, control (scaled) condition............... 178

Table 42 MSE and expected costs for violent crime data, unreduced ICA condition............... 178

Table 43 MSE and expected costs for violent crime data, reduced PCA condition.................. 178

Table 44 MSE and expected costs for violent crime data, reduced ICA condition................... 179

Table 45 MSE and expected costs for nonviolent crime data, control (scaled) data................. 179

Table 46 MSE and expected costs for nonviolent crime data, undreduced ICA condiotion...... 179

Table 47 MSE and expected costs for nonviolent crime data, reduced PCA condition............ 180

Table 48 MSE and expected costs for nonviolent crime data, reduced ICA condition............. 180

Table 49 MSE and expected costs for total crime data, control (scaled) condition................... 180

Table 50 MSE and expected costs for total crime data, unreduced ICA condition................... 181

Table 51 MSE and expected costs for total crime, reduced PCA condition.............................. 181

Table 52 MSE and expected costs for total crime, reduced ICA condition................................ 181

Table 53 MSE and expected costs for cardiotocography data, control (scaled) condition......... 182

Table 54 MSE and expected costs for cardiotocography data, unreduced ICA condition 182

Table 55 MSE and expected costs for cardiotocography data, reduced PCA condition 182

Table 56 MSE and expected costs for cardiotocography data, reduced ICA condition 183

LIST OF FIGURES

Figure 1 Distribution of LOF scores (*MinPts* = 100) for 1000 randomly generated Gaussian data points, with dimensionalities equal to powers of 2 from 1 to 7 70

Figure 2 Calculation of probabilistic LOF scores according to a Gaussian distribution 71

Figure 3 Calculation of probabilistic LOF scores according to a gamma distribution 71

Figure 4 Probability and cumulative density functions for a Cauchy distribution 72

Figure 5 Probability and cumulative density functions for an F-distribution. B is a beta function, and I is a regularized incomplete beta function. 72

Figure 6 Cullen and Frey plot for LOF in the improper subset of the PDB data......................... 74

Figure 7 Weibull distribution fit plots for LOF-1 in the improper subset of the PDB data.......... 78

Figure 8 Cullen and Frey plot for 1-1/LOF in the improper subset of the PDB data 81

Figure 9 Kumaraswamy distribution fit plots for 1-1/LOF in the improper subset of the PDB data 82

Figure 10 Combined scatter and contour plot of MolProbity score and *pLOF* using the 1-1/*LOF* fit to a Kumaraswamy distribution 83

Figure 11 Combined scatter and contour plot of optial resolution and pLOF using the 1-1/LOF fit to a Kumaraswamy distribution 84

Figure 12 Combined scatter and contour plot of R_{free} and pLOF using the 1-1/LOF fit to a Kumaraswamy distribution 85

Figure 13 Failed idea for an overall outlier score as components connected in series 110

Figure 14 Plot of MolProbity Score and final *pLOF* score on the x and y axes, respectively.... 111

Figure 15 Plot of reported Rfree and final *pLOF* score on the x and y axes, respectively......... 112

LIST OF EQUATIONS

Equation 1 Formula for the calculation of the Matthews coefficient 14

Equation 2 Formula for the expected ratio of water molecules to total atoms at a given resolution, in a model of a crystal diffracted at room temperature 15

Equation 3 Formula for the ratio of water molecules to total atoms at a given resolution, in a model of a crystal diffracted at cryo-cooled temperature 16

Equation 4 Formula for R-factors 18

Equation 5 Formula for the volume of a crystal's unit cell, based on its dimensions (a, b, and c) and its vertex angles (α, β, and γ) 54

Equation 6 Formula for the volume of a crystal's asymmetric unit, based on its unit cell volume and the multiplicity of the asymmetric unit 54

Equation 7 Expected value of the Verify3D score, given the number of residues in the model, based on linear regression using 2011 data.. 55

Equation 8 Fractional difference between actual and estimated Verify3D scores 55

Equation 9 Expected value of combined ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data.. 56

Equation 10 Expected value of surface ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data.. 56

Equation 11 Expected value of pairwise ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data.. 57

Equation 12 Expected value of combined ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data.. 57

Equation 13 Expected value of surface ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data.. 57

Equation 14 Expected value of pairwise ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data.. 57

Equation 15 Fractional difference between actual and estimated ProSA energy or z-scores....... 57

Equation 16 Formula for overall MolProbity Score 59

Equation 17 Formula for the estimation of optical resolution from nominal resolution 63

Equation 18 Definition of atomic displacement parameters calculated from atomic B-factors ... 64

Equation 19 Definition of mean model atomic displacement.. 64

Equation 20 Regression equation for estimating mean atomic displacement from optical resolution.. 64

Equation 21 Formula for fractional delta u, based on a model's mean u and an estimate of u calculated from optical resolution.. 65

Equation 22 The ratio of main-chain to side-chain atomic displacement parameter means 65

Equation 23 The ratio of main-chain to side-chain atomic displacement parameter standard deviations 65

Equation 24 Definition of reachability-distance 67

Equation 25 Definition of local reachability density. (Angle brackets indicate average.) 68

Equation 26 Definition of local outlier factor (LOF).. 68

Equation 27 Probability and cumulative distributions of the Amoroso distribution 75

Equation 28 Probability and cumulative distribution functions for the Kumaraswamy distribution, and the relationship between Kumaraswamy and beta distributions........................ 80

Equation 29 Square error loss function and mean square error 86

Equation 30 Unbiased and biased estimated costs based on means squared errors...................... 87

Equation 31 Classical (top) and modern (bottom) definitions of kurtosis. The classical definition is in terms of central moments. The modern definition is in terms of cumulants. The latter is actually excess kurtosis, with the kurtosis of a Gaussian distribution subtracted. 92

Equation 32 Expected value of each eigenvalue under the broken stick null hypothesis............. 95

Equation 33 Signals as linear combinations of independent components 97

Equation 34 Signal mixing in vector and matrix notations.. 97

Equation 35 Unmixing matrix, W, as the inverted mixing matrix, A.. 97

Equation 36 Definition of covariance 97

Equation 37 Differential entropy of a probability distribution, p_x... 99

Equation 38 Negative entropy (aka negentropy) of a probability distribution, p_x, defined in relation to a Gaussian distribution of the same mean and variance, ϕ_x.. 99

Equation 39 General form of a negentropy estimator.. 99

Equation 40 Functions for use in negentropy estimation... 99

LIST OF HYPOTHESES

Hypothesis 1 Consensus Practices and Common Patterns 48

Hypothesis 2 Anomalous Validation Score Patterns 48

Hypothesis 3 Optimal Neighborhood Size 89

Hypothesis 4 ICA Transformation... 93

Hypothesis 5 Simple PCA Dimensionality Reduction 93

Hypothesis 6 ICA From Simple PCA Dimensionality Reduction... 93

Hypothesis 7 Semantic Dimensionality Reduction (Outlier Scoring) 103

Hypothesis 8 Semantic Dimensionality Reduction (Rule extraction) 104

Hypothesis 9 Rule Extraction 135

ACKNOWLEDGEMENTS

In the words of the Grateful Dead, “What a long, strange trip it’s been!” This degree is long overdue. For a long time, I wondered if the light I saw at the end of the tunnel was the exit or an oncoming train. Getting to this point has caused a lot of stress for me, my family, and my advisor. I am therefore greatly indebted to them for their patience and forbearance. Words cannot adequately express my gratitude to them, and the contrition I feel for the difficulties I have caused. Thank you.

I also owe a debt of gratitude to my committee. They gave me a second chance when even I wasn’t sure I deserved one. Thank you.

I am also grateful for the support and advice I have received from my friends and colleagues. They are too numerous to mention them all, but special recognition must go to Kevin Hejna, Erin Patrizio, Fr. David Abernethy (and the whole Pittsburgh Oratory community), Philip Ganchev, Jonathan Lustgarten, and Katrina Romagnoli. They know why. Thank you.

1.0 INTRODUCTION

“The most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ (I found it!), but ‘That's funny...’” – attributed to Isaac Asimov

Structural bioinformatics studies seek evolutionarily or functionally relevant structural features of proteins, nucleic acids, and other biological macromolecules. Computational techniques and tools assist researchers in searching vast repositories of very complex data, containing tens of thousands of structures described by dozens or hundreds of variables, for patterns that may be weak or noisy. This research focused on structures determined via x-ray crystallography and deposited in the Protein Data Bank (PDB). X-ray protein crystallography is a sub-discipline of structural biology in which structures are determined by crystallizing soluble proteins, diffracting x-rays off those crystals, measuring reflection intensities in the resulting diffraction pattern, and attempting to produce a macromolecular structure capable of producing an equivalent pattern. The results of these experiments are typically shared in PDB for use by fellow crystallographers and other various consumers.

One important aspect of informatics studies in a wide variety of domains is anomaly detection. Anomalies are cases that, independently or collectively, do not conform to domain theoretical expectations or deviate markedly from other members of the same data set or sample. Anomalies in structural bioinformatics correspond to novel or unusual structural features, such as rare protein folds or structural differences in evolutionally related proteins. However, before such novelties can be reliably identified and characterized, a different kind of anomaly must be identified in the data used to represent macromolecule structures.

Though the products of structure determination experiments are commonly referred to as structures, they only approximate physical reality and are more accurately described as models. Structure model anomalies may be explainable as real, and potentially interesting, structural deviations from ideality, but unless the underlying experimental data are of very high quality, some skepticism is prudent.

PDB data present a particularly challenging and interesting domain for the outlier detection. Practically none of the ~70,000 PDB models is labeled as problematic, so a detection algorithm must infer outlier status from validation scores produced by multiple tools. Deposited over ~40 years, the data form a highly heterogeneous set of models built from experimental data of varying quality and according to varying techniques. Due to this heterogeneity, a large proportion of the entries are missing at least one field (validation score), with many missing several.

This research has demonstrated the feasibility of applying unsupervised anomaly scoring and detection methods to the identification of anomalous protein structure models from multiple validation measures representing noisy, partial, and overlapping indications of model quality. However, the opacity and unintelligibility of common anomaly scoring methods are barriers to their acceptance by structural biologists. Therefore, the use of IF-THEN rules was explored as post-processing technique for making them comprehensible and explainable.

2.0 BACKGROUND AND SIGNFICANCE

"When selecting data sets for deriving general principles about, say, protein structures it is important to filter out those that might give misleading results simply because they are unlikely to be sufficiently accurate or precise to contribute meaningful or correct data to the analysis…[If] you put unsound data into your analysis, you will get unsound conclusions out."[1]

2.1 STRUCTURAL BIOLOGY AND BIOINFORMATICS

2.1.1 Structural Bioinformatics

A variety of techniques for discovering patterns and knowledge from data sets in an algorithmic fashion are generally referred to as data mining. The broader processes of data preparation, selection, and cleaning, combined with incorporation of relevant prior knowledge and interpretation of data mining results constitute knowledge discovery in databases. The study of these processes is part of the field of informatics. These computational techniques and tools assist researchers in searching vast repositories of very complex data, testing a large number of potential statistical models to find robust representations for signals that may be weak and/or obscured by a high degree of noise. When the data are generated by scientific observation or measurement, and patterns are translated into testable hypotheses, mining efforts are sometimes referred to as automated scientific discovery or computational discovery[2-5].

The analysis of databases containing biochemical structures has given rise to the emerging field of structural bioinformatics. Structural bioinformatics seeks to analyze, describe, and in some applications predict, the three-dimensional structure of proteins, nucleic acids, and other biological macromolecules. The structure of biological macromolecules is important because structure determines function.

2.1.2 Structural Biology

Structural biology leverages the combined knowledge and techniques of molecular biology, biochemistry, and biophysics to determine the structures of macromolecules (proteins and nucleic acids in particular). Biologically relevant structural regularities are generalized from these structures. Then, structural regularities are in turn related to functions. X-ray protein crystallography is a sub-discipline in which structures are determined by crystallizing soluble proteins, diffracting x-rays off those crystals, measuring reflection intensities in the resulting diffraction pattern, and attempting to produce a macromolecular structure capable of producing equivalent pattern. The results of these experiments are typically deposited into the Worldwide Protein Data Bank (wwPDB or simply PDB)[6] for use by fellow crystallographers and other various consumers.

A common misconception, held primarily by non-structural biologists, is that once a structure model has been accepted by the PDB it can be regarded as essentially free of error[7]. This is a mistaken notion with potentially disastrous consequences, from wasted scarce research resources to the design of hazardous pharmaceuticals.

Though the products of structure determination experiments are commonly referred to as structures, they are more accurately described as models. To some degree, every protein structure model is "wrong". All approximate reality, lacking detail at some level[8], and may be inaccurate and/or imprecise[9]. The number and degree of errors varies greatly from model to model, as explained in section 2.3).

Unfortunately, the most biologically interesting structures tend to also be the most difficult to determine, and models of them are often of low detail and reliability. Nevertheless, if those structures are to be studied, bioinformaticians must rely on those deeply flawed models. Consequently, there is a need for means to judiciously “make do” with the models available while avoiding seriously problematic outliers.

2.1.3 Due Diligence in Structural Biology

Structural biologists deal with the approximate nature of protein models through community-enforced standardization of methodologies. As technologies and techniques evolve, so do common practices.

One set of standard methodologies governs the building and refinement of structure models, expressed in terms of what makes a structure model “good”. Good structure models should make sense in light of current accepted knowledge and practices of physics, chemistry, crystallography, protein structures, statistics, and biology[10,11]. That is, a reliable model should have no interpenetrating non-bonded atoms, its covalent geometry should be properly restrained/constrained, its structure must adequately explain the experimental diffraction data, its structure must “look” like a protein, its stereochemistry must be reasonable, over-fitting should be minimal, and the structure it posits should explain previous observations and enable one to make predictions that can be tested experimentally.

Another set of standard methodologies governs the validation of structure models, both before and after to publication/deposition. In order for a model depositor, a journal referee, a structural bioinformatician, or any end user of a model to be confident that it is unlikely to be problematic, confounding, or misleading, he applies several software tools to obtain validation scores. A validation score is a quantitative measure of a particular aspect of a model's validity. Thus, it gives an incomplete and imprecise indication of the likelihood that it contains serious systematic errors. When a crystallographer examines these scores for a particular model, he is looking for possible anomalies or outliers. In these contexts, "anomalous" means that a model may be unsuitable for structural visualization, a confounder in large-scale structural studies, and/or likely to lead biological investigations astray. Models that score consistently poorly across multiple validators should be regarded with suspicion[12]. Also, those with highly inconsistent scores are suspect[13]. This process helps depositors to avoid contributing noise to the PDB, referees to stop suspicious structural details and/or theories from being published, and PDB users from including confounding data in large-scale structural studies or other studies for which reliable models are required in order to avoid reaching wrong conclusions.

2.1.4 Due Diligence in Structural Bioinformatics

Structural biologists expect that anyone making use of models in the PDB should be able to independently apply domain theory to assess their validity. *Ergo, caveat emptor*. A detailed and conscientious analysis of each model considered for informatics use would be ideal. This would involve reading the relevant paper(s), applying appropriate validation tools, inspecting the results, and using domain expertise to determine whether authors' conclusions are supported by their deposited data. If only a small handful of potential models are under consideration, manual evaluation and selection of problem-free exemplars (or alternately anomalous exemplars for detailed study) may be feasible. However, this is rarely the case, and more often, an investigator must choose from hundreds, thousands, or even tens of thousands of models. Consequently, an informatician must rely on large-scale automation. This automation must necessarily rely on a greatly simplified version of the evaluation that a responsible structural biologist would perform.

2.1.5 Motivation for This Research

This research is timely. Responding to twenty years of calls from the crystallography community for more frequent and more thorough validation of structure models[8,14-21], and for appropriate validation tools to be provided to depositors, referees, and consumers, a few years ago the wwPDB convened a Validation Task Force (VTF). Such calls had become more fervent in recent years, thanks in large part to some high-profile retractions[22,23]. The VTF has recently finished its deliberations and published recommendations for updating validation criteria, stating which types/sources should be used, how they should be used, and how the results should be presented to various types of users.

VTF's recommendations state, in part, that consumers of protein structure models, particularly non-specialists, need to be able to critically assess the quality of models before using them in studies. Multiple indicators of model quality should presented for users' benefit because a "good model will typically score well on most if not all validation criteria, whereas a poor one will score poorly on most criteria"[11]. The information provided should be "easily understandable by scientists", "not require a deep understanding of crystallographic or validation methodology", "clearly measure quality", and "not depend strongly on arbitrary cutoffs"[12]. This study represents a demonstration of the feasibility of using outlier scores to achieve these goals.

2.2 PROTEIN STRUCTURE DETERMINATION

In coarse terms, there are three experimental stages in protein x-ray crystallography: crystallization, diffraction, and model building/refinement. Error propagated from each stage affects the accuracy and precision of protein structure models. Tests of accuracy ask, "How close are the results on average to the truth (regardless of their precision)?", and tests of precision ask, "If you were to repeat the experiment, how much would you expect the results to vary (regardless of their accuracy)?" Accuracy is a property of the model and observed error; precision is a property of crystal quality, experimental data quality, and expected error/uncertainty[24].

2.2.1 Protein Crystallization

Globular proteins (the primary focus of this study) are more or less water-soluble. As such, they are highly mobile and disordered in their natural state, making them difficult to visualize to study. To make proteins more amenable to structural analysis, they are crystallized into three-dimensional lattices that are grown to 20 – 100 µm. However, lattices in protein crystals are not uniformly well ordered. Chemical heterogeneity, multiple conformations, intrinsically disordered regions, local flexibility, and small thermal atomic motions all cause atoms in the protein to not be in exactly the same position in every unit cell (repeating macromolecular structure) in a crystal lattice. Furthermore, the unit cells themselves are subject to the problems of mosaicity (long-range disorder) and weak lattice contacts (where molecules in different unit cells meet). That is, most crystals are not a single lattice but blocks of nearly identical lattices oriented at different angles with respect to each other. Mosaicity measures the average angle of divergence between blocks. Within each block, unit cells may not be consistently aligned, due to weak bonds between them.

Crystalized proteins are also surrounded by a large number of highly mobile (i.e., disordered) solvent molecules. Bulk solvent is both a blessing and a curse, because while it beneficially allows ligands to flow through channels to binding sites, but it also facilitates the transportation of free radicals, which are caused by ionizing X-ray radiation and cause damage to protein structure, and contributes to background noise in diffraction patterns[25,26].

These inconsistencies become problematic at the diffraction stage, because x-ray diffraction patterns are spatially and temporally averaged images. Identical and perfectly repeating unit cells would produce high-contrast diffraction patterns with noiseless reflections measurable out to device limits. Indeed, this is very nearly the case in small molecule crystallography.

2.2.2 X-Ray Diffraction

The diffraction experiment is the last physical experiment in the sequence of steps of structure determination. The quality/information content of a structure model is limited by the quality/information content of these underlying diffraction data[27].

Proteins are too small to be seen using light or even electron microscopes. Instead, x-ray diffraction must be employed. Unfortunately, no lens with sufficient index of refraction to refocus x-rays exists, which greatly complicates the structure determination process.

When a focused beam of x-rays is diffracted off a crystal, it produces a pattern of reflections. Systematic errors in this diffraction pattern may arise from variations in rotation during sampling of reciprocal space, rapid fluctuations in beam intensity, or errors in shutter synchronization. These errors are difficult or impossible to detect or model. The most difficult systematic error to account for is radiation damage, which contributes to a crystal's disorder by causing structural changes over time during x-ray exposure, and different reflections change at different rates[28].

In addition to a non-trivial amount of noise intrinsic to the process of generating, collecting, and measuring reflections, disorder from crystallization strongly attenuates reflections, reducing their intensities. The diffraction pattern, represented mathematically as a set of structure factors, is the Fourier transform of the structure's three-dimensional atomic coordinates (or equivalently, its electron density function). Inverting that transform to recover the electron density (and therefore the structure) would require measuring the amplitude, frequency, and phase of each reflection, but phases cannot be recorded and measured directly. Consequently, they must be acquired by indirect means, a process that may produce errors affecting the interpretation of the electron density.

A typical protein has a molecular weight of ~50kDa, ~3,500 non-hydrogen atoms, and 14,000 isotropic atomic displacement parameters or 31,500 anisotropic atomic displacement parameters. Unfortunately, if the diffraction data are resolved to 3Å there will be only ~8,500 reflections; at 2Å there would be ~28,000 reflections. Because the number of observations is close to the number of parameters, the mathematical problem of determining protein structures is often only weakly over-determined and sometimes under-determined, making it only barely solvable[25].

2.2.3 Model Building and Refinement

Model building refers to the process by which a protein's polypeptide chains are threaded through a rough early version of the electron density map build from the diffraction data. Unless the electron density map has very high resolution, there are many choices to be made in that process and those choices are not always made appropriately. Errors at this stage can include misthreading due to backbone connectivity errors, misalignments/misregistrations (residue sequence shifted by one or more positions from their matched electron densities), and misplacement or incorrect chirality of side-chains (which tend to be more disordered)[29-31]

The refinement stage attempts to correct the accumulated mistakes of the previous stages. However, it may also introduce new mistakes as well. The process of model refinement involves a number of interrelated decisions, some of which are implicit. They may be based on statistics, personal preference, experience, instinct, or dogmatic principles. For instance, simply choosing to accept a program's default parameter settings can have profound effects on refinement outcomes[32].

In refinement, the atomic coordinates are adjusted to improve agreement between observed structure factors and those calculated from the model. It is a weakly over-determined, and sometimes under-determined, process, which cannot be entirely rescued by merely reducing the number of the parameters being adjusted. Rather, the space of acceptable models to be searched must be constrained. Additional physical information must be brought to bear, most prominently from known protein stereochemistry. Since atomic positions are not independent of each other, and are related in physically limited ways, the search for an optimal structure in model space can be constrained to make convergence computationally tractable. Convergence to the *global* optimum, however, is far from guaranteed.

2.3 STRUCTURE MODEL VALIDATION

2.3.1 What is validation?

Validation is the process of evaluating the how successfully the previous stages were performed. Structural biologists generally refer to a valid model as being one that is the most consistent with data – both diffraction data and prior knowledge. This is primarily a test of accuracy, in which the most accurate model is the one that globally maximizes the total likelihood and minimally over-fits to errors in the data. Non-specialist end users more typically understand validation to be a test of precision that asks the question, "How useful is the model in terms of the reliability of conclusions drawn from it (e.g., regarding structure-function relationships), assuming the accuracy has been verified?" In selecting subsets of models for use in structural informatics studies, we are concerned with both types of validation.

2.3.2 What are common validation methods?

A thorough and exhaustive description all of the available and commonly used validation tools and techniques is beyond the scope of this setting. For the purpose of this dissertation, the types of validation scores to detect outliers will be classified as pertaining to diffraction data quality, crystal parameter checks, model parameter checks, model-data agreement, covalent geometry checks, and stereochemistry checks. At least one representative example of each is given. Specific details about these and additional validation scores are provided in section 4.2, as warranted by their relevance to the stated goals of this study.

2.3.3 Diffraction Data Quality

Typical indicators of data quality are completeness (the ratio of reflections observed to theoretical reflection count), signal-to-noise ratios (reflection intensities vs. reflection spreads), Wilson B-factor (an estimate of structural disorder), nominal resolution (the smallest measurable spacing between Bragg planes), and optical resolution (the theoretical distance between the two closest resolvable objects in the electron density). These scores are to varying degrees related and interdependent. For instance, the number of measurable intensities is known to be inversely proportional to the cube of the nominal resolution, and the degree of structural disorder is the primary limiting factor on the nominal resolution.

While nominal resolution should not be equated with model quality[33] (or, strictly speaking, data quality), it does provide an intuitive guide to how diffraction data quality affects model quality. Resolution is essentially the aperture of the diffraction imaging experiment. Data at different resolutions support different levels of structure model detail[25,32,34]. Unless a structure is extremely novel and important, 4 Å is the limit of publishable structural studies[35].

- At >3Å, only the main chain can be traced. Local electron density peaks merge and maxima appear at locations not corresponding to atomic positions.
- At 3Å, residues of amino acids alanine and isoleucine should be distinguishable.

- At 2.7Å, Hydrogen-bonded water molecules can be distinguished, and solvent modeling may be cautiously attempted.
- At 2.0Å, leucine and isoleucine are distinguishable and models can be called “high resolution”.
- At 1.5Å, there may be enough data to attempt anisotropic refinement.
- At 1.0Å, hydrogen atoms begin to be discernible and individual non-hydrogen atoms can be seen as discrete balls in the electron density plot.
- .48Å, is the current record for best macromolecular resolution achieved[36].

1.2Å is a criterion for atomic resolution, because it is the length of a carbon-oxygen double bond, the shortest covalent bond in proteins not involving hydrogen atoms[25]. Also, with > 50% completeness in the 1.1 – 1.2Å range (i.e., $F > 4\sigma(F)$ for > 50% of the theoretical reflections), direct methods of structure solution may be applied[37,38].

2.3.4 Crystal Parameter Checks

Crystal parameter checks evaluate the appropriateness of such things as unit cell dimensions, space group, non-crystallographic symmetry, solvent content.

The Matthews coefficient[39,40] is an expression of the specific volume of a crystallized protein, as shown in Equation 1. Here V_{unit} is the volume of the unit cell, n_{asym} is the number of asymmetric units per unit cell, and W is the molecular weight of the asymmetric unit.

$$V_M = V_{unit} \Big/ \left(n_{asym} * W \right) = V_{asym} / W$$

Equation 1 Formula for the calculation of the Matthews coefficient

Lower values of V_M have been associated with crystals diffracting to higher resolutions, suggesting that tightly packed proteins tend to diffract better than those that are loosely packed[40]. A quasi-linear correlation was found between the natural log of V_M and resolution[39], with structures off the diagonal being "obvious outliers"[24].

2.3.5 Model Parameter Checks

Model parameter checks explicitly modeled water molecules, and atomic displacement parameters (overall, TLS, isotropic, or anisotropic).

In order for water molecules to produce measurable reflections, sites must maintain sufficiently high occupancy during diffraction to contribute to the resulting pattern[41]. In addition, as the number of protein atoms in a model rises, the number of modeled water molecules falls. This is due to the presence of more and larger buried hydrophobic regions[41].

The ratio of modeled water molecules to the number of modeled protein atoms has been found to correlate with resolution as indicated in Equation 2 and Equation 3 for room temperature and cryo-cooled conditions, respectively[41]. Consequently, there are expected to be slightly less than one modeled water molecule per residue at 2Å. At 1Å, 1.6 – 1.7 water molecules per residue are expected.

$$N_{HOH} / N_{atoms} = .301 - 0.095 * resolution$$

Equation 2 Formula for the expected ratio of water molecules to total atoms at a given resolution, in a model of a crystal diffracted at room temperature

$$N_{HOH} / N_{atoms} = .334 - 0.11 * resolution$$

Equation 3 Formula for the ratio of water molecules to total atoms at a given resolution, in a model of a crystal diffracted at cryo-cooled temperature

The decision process for selecting a B-factor model in PDB-REDO[32], an automated re-refinement tool, presents an interesting case study of both guiding principles in model parameter selection and potential validation criteria.

- If the number of reflections per atom (RPA) is greater than 18, there is sufficient over-determination to use individual atomic anisotropic B-factors.
- If 18 > RPA > 13.5, atomic B-factors are initialized to the Wilson B-factor, automatic refinements are performed with anisotropic and isotropic B-factors, and Hamilton R-factors[42,43] are used to determine whether the additional parameters in the anisotropic model are statistically justified.
- If 13.5 > RPA > 3, isotropic atomic B-factors are used.
- If RPA < 3, TLS refinement (essentially anisotropic group – rather than individual atomic – B-factors) is applied and optimized first. Then, refinements with a single overall isotropic B-factor and isotropic atomic B-factors are tried and compared as above.
- If TLS cannot be used, e.g., due to unstable refinement, isotropic atomic B-factors are used.

Another important factor in assessing the appropriateness of atomic displacement parameters (B-factors) is how they vary throughout a structure. Protein side chains are known to be more flexible and more disordered than main chains, so side chain atoms are therefore expected to have larger B-factors than main-chain atoms[44-52]. Furthermore, the surface-molten character of proteins and correlation between B-factors and solvent accessibility generate the expectation that B-factors should generally increase with distance from a protein's core[53].

2.3.6 Covalent Geometry Checks

During the refinement process, the enormous space of possible structures is reduced drastically by restraining certain properties based on rules of covalent geometry. These properties are derived from experiments in small molecule, atomic resolution crystallography[54] and theoretical quantum mechanical calculations. Validation tools typically report the agreement of main chain bond lengths and main chain bond angles with these expectations. They may be presented as either rms deviations or outlier percentages.

2.3.7 Model-Data Agreement

The most prominent measures of model-data agreement are the R-factors, R_{work} and R_{free}[a], which calculate the sum-squared error between observed and calculated diffraction patterns from working set (i.e., training set) reflections and free set (i.e., validation set) reflections, respectively. That is, the more accurately and precisely a structure model reflects the atomic structure that produced the observed diffraction data, the smaller difference between observed

[a] The 'R' stands for either "residual" or "reliability", depending on whom you ask.

and calculated (hypothetical) structure factors should be. Consequently, refinement can be guided by seeking to minimize the R-factors.

$$R = \frac{\sum \left\| F_{obs} \right| - \left| F_{calc} \right\|}{\sum \left| F_{obs} \right|}$$

Equation 4 Formula for R-factors

Prior to 1992, there was only one R-factor. Then Brünger, Jones, and many others observed that R could be arbitrarily reduced by increasing the number of adjustable parameters used to describe the structure or the number of restraints imposed during refinement of the structure, or by reducing either the number of experimental observations, as described above[11]. As a form of single-fold cross-validation, Brünger advised splitting the R-factor into a training set, R_{work}, and a validating set, R_{free}[55]. In current refinement practices, structures are typically refined until a sensible R_{free} value is reached. This can be likened to guiding machine learning using a validation set while learning from a training set. Since R_{free} should be less prone to over-fitting, and low R_{work} values are no longer a primary goal in refinement, R_{work} should be proportional to R_{free}. In turn, both should be proportional to resolution, as described above. It is important to remember, however, that "the goal of a protein refinement is usually to answer a biological question, rather than minimizing any given statistic, per se."[56]

One can also measure the difference between R_{free} and R_{work} as a form of validation. When R_{work} is low and R_{free} is high, some degree of over-fitting is likely to be present in the model. It has been suggested that the difference between R_{work} and R_{free} should be as small as possible[11,19,21], generally less than 5%, or the ratio close to unity[11,21,57].

There are several categories of errors that cause observed and calculated diffraction patterns to differ[24,56]:

- Random experimental (i.e. measurement) errors in observed amplitudes
- Missing higher resolution data due to weak diffraction
- Errors in the algebraic form of the structure factor model (i.e., how anisotropy, anharmonicity, disorder, and the like are expressed)
- Parameter errors in the structure factor model (e.g., errors in scaling, bulk solvent corrections, atomic parameters, misplaced/missing atoms, and inadequately modeled disorder)
- Insufficiently accurate atomic model convergence to an incorrect minimum
- Under-refinement (insufficient iterations for convergence).

Random errors require better crystals and/or better data collection techniques/technology to correct. In small molecule structure determinations random errors dominate. Structure factor model errors produce errors in both calculated amplitudes and phases. Due to the limited resolution of macromolecular diffraction patterns, structure factor model errors dominate[56]. Parameter value errors can be caused by overly tight positional constraints

For macromolecules with typical nominal resolutions, the model error component in calculated structures dominates with about four times the effect of data errors. Thus, the precision of the data may be less than 5% while the final R-factors are 15-20%, even with optimal model parameterization and all parameter value errors corrected[24].

2.3.8 Stereochemistry Checks

Allowed and disallowed regions for main-chain torsion angle combinations in alpha helices, beta sheets, and loops are mapped in Ramachandran plots. Along the backbone of a protein chain, there are three torsion angles, φ, ψ, and, ω. Combinations of φ, which is along the N-Cα bond, and ψ, which is along the Cα-C bond, are strongly restricted by steric repulsion[b]. Ramachandran plots of φ and ψ combinations in a model are generally displayed as points against a contour map of regions for which angle combinations do not result in steric clashes. The third torsion angle, ω, is along the C-N bond and is typically constrained to a tight distribution around 180°, so it is often treated like a covalent geometry measure.

It is generally agreed that a realistic model should have most of its amino acid residues in the most favored regions of the φ-ψ main chain torsion angle (Ramachandran) plot because the disfavored regions represent steric clashes[58]. It is reasonable to expect that this would be easier to achieve at high resolution[18,59,60]. Similar to the percentage of outlying φ-ψ combinations, the percentage of side chain torsion angle pairs (rotamers), χ1 and χ2, that are within tolerated regions is commonly reported. As with torsion angles, certain combinations of side chain angles are more likely due to steric clashes resulting from others. For both of these measures, one would expect poor scores at low resolutions. As resolution improves, so should these scores. Furthermore, we expect strong correlations between these scores and R-factors[18,59].

[b] Steric repulsion is caused by electrostatic repulsion of electron shells. Steric clashes occur when non-bonded atoms have interpenetrating electron shells.

Some validation tools also report a measure of violations of van der Waals radii in terms of how many non-bonded nuclei are too close to each other (producing steric clashes). During refinement, van der Waals radii are enforced via the 6-12 potential. That is, the attraction between two atoms changes in proportion the sixth power of the distance between their radii and the repulsion changes in proportion to the twelfth power of that distance, resulting in contact distances that are more or less energetically favorable. Though minimized during refinement, the number of outliers among non-bonded interactions has greater value as a validator than directly restrained measures, though, because minimizing a many-body energy equation is not the same as imposing particular distances between atoms. If the number of bad contacts reflects the amount of free energy well, more bad contacts should be flagged at poorer resolutions than at better resolutions.

2.4 ANOMALIES

2.4.1 What are anomalies? Why are they important?

One important aspect of informatics studies in a wide variety of domains is anomaly detection. Anomalies are cases that, independently or collectively, do not conform to expected behavior[61] and are inconsistent with or deviate markedly from other members of the set within which they occur[62]. Anomalies are important because they “translate to significant (and often critical) information”[63]. Sources of anomalies include instrumentation/measurement error, human error, rare circumstances, novel phenomena, and malicious activity[63]. Investigators may wish to treat them as noise to be filtered out, previously unknown signals to be accommodated, problems to be investigated, or novel targets for research; these ends are not entirely mutually exclusive. A common trait to anomalies is that they are interesting to domain experts.

Anomaly detection has been studied in the statistical community since at least the last 19th century[64], and it is still an active area of research in data mining and automated scientific discovery[61]. Some recent examples are the detection of potential clinical errors, computer intrusion, network intrusion, fraud (bank, credit card, insurance, identity, etc.), mechanical faults, structural defects, and census abnormalities, to name but a handful.

2.4.2 What are anomalies in structural bioinformatics?

Anomalies in structural bioinformatics arise from many sources and are interesting for different reasons. Since the aim of structural bioinformatics is to identify structural patterns in biological macromolecules relevant to functional attributes, anomalies correspond to novel or unusual structural features. For instance, certain protein folds may be rare, or evolutionarily related proteins may have important structural differences.

However, before such novelties can be reliably identified and characterized, a different kind of anomaly must be identified in the data used to represent macromolecule structures. For reasons elaborated in section 2.2, structure models are necessarily approximate. Due to their approximate nature, structure models are of varying reliability (also referred to as validity or quality). Reliability of models is assessed in part by using validation tools that assess their accuracy and precision with respect to what is known about protein structure. Anomalies among these models, identifiable by unusual validation score combinations (consistently good, consistently poor, or conspicuously inconsistent) are the focus of my research.

Structure model anomalies scores may be explainable as real, and potentially interesting, structural deviations from ideality, but unless the underlying experimental data are of very high quality, some skepticism is prudent. The aim of my research was to discover patterns in protein structure model validation data that would allow structural biologists to triage tens of thousands of models and quickly identify those sufficiently unusual to warrant closer inspection.

2.4.3 Why is it important to identify anomalies in validation data?

Avoiding the inclusion of anomalies of this type is particularly important because "if you put unsound data into your analysis, you will get unsound conclusions out." Anomalies, which are generated from different distributions, can confound statistical analyses. That is, if they are too great in number or too severe in magnitude they are likely to disrupt efforts to characterize inlying data. Mischaracterization of functionally relevant structural patterns may lead to incorrect scientific conclusions or predictions. At best, further research based on unsound conclusions will waste resources such as time, money, and effort. At worst, they may result in dangerous situations or products, such as might be the case in structure-based pharmaceutical drug design.

2.4.4 Why are anomalies hard to detect?

Unfortunately, the detection of outliers is not trivial in large datasets. Indeed, the allusion to a needle in a haystack is quite apt here. If the stack is very small, one might be reasonably expected to complete the task, though it would almost certainly be tiresome drudgery to do so. If the stack is very large, the task becomes practically impossible. The time commitment and cognitive load of finding a small number of highly varied and disparate outliers among hundreds to tens of thousands of database entries are far too great.

For outlier detection to be feasible on a large scale, automation is required. Since automated methods cannot be as thorough, as flexible, or as nuanced as experts would be, a great deal of expertise must be condensed into an algorithm that is feasible to implement. While the resulting detection algorithms may be comparatively simple, producing them is often quite challenging.

In addition to the challenges related to the nature of anomalies, there is the significant problem that database entries are rarely labeled as being a “normal” case or an “outlier”. The prospects of getting domain experts to carefully review and label more than a few dozen cases (if even that many) are exceptionally poor, let alone tens of thousands. Even if labels could be obtained for a subset of the data, there is no guarantee that a representative sample of outliers would be present in that subset.

For protein structure models, an additional difficulty arises because the validation scores associated with them were developed for human use on a case-by-case basis, and in concert with extra-computational knowledge. They provide partial and overlapping evaluations of model quality, and therefore complicate attempts to find independent, complementary, and comprehensive attributes.

2.4.5 Why is it hard to automate anomaly detection?

While the anomaly detection algorithms may be relatively simple, producing them is often quite challenging. The challenges may be understood in light of surprisingly astute observation made by the leader of a political bureaucracy.

"[T]here are known knowns; there are things we know that we know. There are known unknowns; that is to say there are things that, we now know we don't know. But there are also unknown unknowns – there are things we do not know, we don't know." — *former United States Secretary of Defense, Donald Rumsfeld*[65]

2.4.5.1 Known Knowns

"Known knowns" are observed and well-characterized anomalies. Such cases may be produced in a particular time period, having been generated and measured using particular instruments or particular experimental techniques, for instance. They should be relatively easy for automated methods to identify in the data using well-understood error propagation techniques. Unfortunately, this type of anomaly is rarely the most common or most severe type in a data set.

2.4.5.2 Known Unknowns

"Known unknowns" are unobserved particular instances of known anomaly types. While the phenomena responsible for these outliers may be well understood, the distributions of values they produce for variables in the data may not be particularly complete. Furthermore, the boundaries between normal and outlying cases may be poorly resolved, if at all.

2.4.5.3 Unknown Unknowns

Unknown unknowns are unobserved anomalies of unexpected or ill-characterized anomaly types. Fraud is the most serious anomaly of this type. Examples of fraud are thankfully rare. However, this rarity provides a paucity of data with which to train computational detection methods to find them. They are also inherently irregular, since we can presume that perpetrators will attempt to avoid the mistakes that got their predecessors caught. Furthermore, fraudulent data are likely to be obfuscated and disguised as legitimate, thereby making their “signatures” distinct from ordinary errors. Honest mistakes can be expected to follow predictable patterns in obedience to situational conditions, such as physical laws; if parameter A’s value is “too good to be true” due to an over-fitting model, we can expect parameter B’s values to be noticeably suboptimal. However, when dishonesty is operant, known causal relationships cannot be counted on, and one is faced with the difficult proposition of identifying distinct unintended consequences of manipulating fabricated data to appear real.

Another type of unknown unknown is an error in the methodology used to generate or collect case data that is not widely known until advances are made in the scientific understanding of a field. Programming a tool to find those mistakes well after they were made is necessary for many applications, but insufficient for detecting mistakes being made presently. It may seem to be a catch-22 to know how to find mistakes you do not know you are making, but it is vitally important to do so in order to avoid including misleading information in analyses. Consequently, algorithms must be developed that are capable of finding peculiar cases in the data that seem to defy current domain theory.

2.4.5.4 Unknown Knowns

Interestingly, Rumsfeld omitted “unknown knowns”. These are things we think we know, but actually do not (because we are mistaken). In this context, unknown knowns are likely to be found in the prior knowledge used to supplement experimental data and guide model building and refinement. Kleywegt crystallographer expressed the need to critically examine prior knowledge with a quote from Mark Twain[15]. “The trouble with most of us is that we know too much that ain’t so.” Ironically, this quote was not critically examined. It seems to be a mutated form of a paraphrase Twain made of an aphorism of Josh Billings (real name Henry Wheeler Shaw). It has been attributed to a wide variety of people, but Mark Twain is most often credited. Countless people “know” Twain said it, but that “ain’t so”[66]. (Then again, perhaps that was Kleywegt’s point.) An example of an unknown known in prior protein knowledge relates to the statistical distribution of the peptide bond angle. Historically, it was constrained to be planar. However, later research indicated that it can deviate from planarity by several degrees in well-determined structures[67,68].

2.4.5.5 Algorithmic Difficulties

A naïve approach to detection outliers might be to define a normal region in attribute-value space and declare anything outside that region anomalous. This is the approach typically taken in univariate statistical methods. Unfortunately, this simple approach in complicated by a number of challenges[61].

It is often difficult to define a complete and unbroken boundary around normal cases. When regions of attribute-value space are sparse, it is hard to drawn a complete boundary without over-fitting to noise. Boundaries between normal and anomalous cases are often imprecise and may even overlap. This is likely to be the case when the set of available attributes does not include all of the dimensions necessary to cleanly separate classes. A more complicated situation arises when attributes combine in non-linear ways, leading to non-linear decision boundaries, such as in the case of the logical operator XOR.

In cases involving malicious actors, such as in frauds, anomalous cases may be disguised as normal cases. Worse, the profiles of anomalous cases are likely to change over time as old camouflage tactics are neutralized and actors adapt. So far, this does not seem to be the case in structural biology, as the known cases of fraud were not particularly sophisticated or clever. However, as the existence of undiscovered sophisticated frauds in the PDB is an "unknown unknowns"[65], we cannot confidently state that they do not exist.

In addition to evolving anomalies, there may be evolving normal profiles. Indeed, this is the case in data extracted from the PDB. Over the course of time, technologies and techniques for structure determination and validation have improved, resulting in older models appearing sub-optimal by modern standards.

Different outlier detection methods are better suited to different kinds of data (e.g., categorical, discrete ordinal, or continuous). In the presence of mixed sets of attribute types, some methods require extensive adjustment to work, while others are completely inapplicable.

Lastly, labeled training data may be either mislabeled or missing. Mislabeled cases add noise to the data, causing machine learning techniques to over-fit or fail to converge close to the global optimum. Labeled data may not be available in large quantities or at all, though, so many methods of outlier detection may not be applicable. Such is the case for structure models in the PDB, which lack "ground truth" labels of normality or otherwise.

2.4.5.6 Data Difficulties

In addition to the challenges related to the nature of anomalies, the datasets within which they must be detected introduce their own difficulties. The greatest of these is that database entries are rarely labeled as “normal” or “outlier”. This automatically rules out the use of supervised machine learning techniques, which attempt to infer functional relationships between input values (database entry fields) and output values (classifications). Furthermore, the prospects of getting domain experts to carefully review and label more than a few dozen cases (if even that many) are exceptionally poor, let alone tens of thousands. Consequently, unsupervised methods, which attempt to infer the probability distributions of input variables irrespective of any target classification, must be relied upon to discover “emergent” properties of the data that set outliers apart from normal data.

Machine learning techniques are only as reliable and informative as the data on which they operate. Scientific data, especially in very large databases, are highly heterogeneous and irregular. This is because experimental data may be acquired using different tools and techniques, or according to different assumptions, theoretical foundations, and experimental conditions. Heterogeneity may be augmented by temporal changes in databases. That is, the distributions of variables characterizing each entry may change over time. If they do, a case that might be an outlier compared to contemporary data may be ordinary in comparison to data from other time periods, and vice versa.

Additional complications arise from related problems concerning the availability of discriminating variables describing database entries. These complications include incomplete data fields and redundant/overlapping fields. Both problems reduce the number of independent indicators of possible anomalousness. Cases with missing data (due either to omission or to changes in preferred descriptive or diagnostic criteria) are difficult to directly compare to other cases. They also reduce a detector’s ability to check variable values against each other for consistency. Redundancy is good from the perspective of ensuring robust observation of phenomena, especially in light of the preceding problem. However, redundant measurements for the same or overlapping phenomena may make cases artifactually appear more dissimilar than they really are by magnifying small differences. This can be especially problematic if independent aspects of observations are not all equally redundant.

Finally, even perfectly collected and homogeneous data may be tricky to consistently and unambiguously divide into “normal” and “anomalous” sets. Distinctions between the two distributions are typically subtle and non-linear, with neighboring clusters having overlapping boundaries. A truism of statistics states that rare events happen, and indeed, it is not uncommon for cases in the tails of populations to be normal and not really anomalies. Conversely, true anomalies may have attribute values well within the range occupied by normal cases. Where distributions overlap like this, one cannot be certain which group a case belongs to without additional information, such as contextual metadata.

2.4.5.7 Curse(s) of Dimensionality

The curse of dimensionality is an umbrella term that covers several related problems. The following review is adapted from “A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data”[69]. The reader is advised to consult that article and references therein for more detailed treatments of these concerns. The problems of high dimensionality can be divided into concentration of distances, dimensions that are highly redundant, noisy, and/or irrelevant, combinatorial explosion, subspace selection paradox, data snooping bias, hubness, and difficulty classifying scores.

Concentration of distances refers to the phenomenon of the convergence of the minimum and maximum between-point distances as dimensionality approaches infinity in the limit. The effect of concentration is to reduce the difference between intra- and inter-cluster point distances.

When the number of irrelevant attributes (little signal, mostly noise) and/or the amount of additive noise in relevant attributes is high, false positive and false negatives rates can be high. False positives may occur when normal cases occupy low density regions of feature space because distances are inflated by noise. False negatives may occur when anomalous cases share feature space with normal cases, due to noisy distance calculations.

Combinatorial explosion refers to the fact that the number of possible attribute-value pairs in a feature space increases exponentially with the number of dimensions. Even if each dimension is discretized to a fixed number of values, *V*, the number of combinations will be V^A, where *A* is the number of dimensions. At *V=10* and *A=100*, the number of combinations is a googol – more than the number of atoms in the known universe. Without adequate sampling of feature space, it is likely that rare, but normal, cases will be misidentified as anomalies, even when all of the dimensions are relevant and informative.

Attempts to avoid other curses by selecting highly discriminative subspaces (subsets of attributes) via dimensionality reduction (either case-by-case or global) create another problem. A case may be anomalous in certain subspaces but not others. A paradox arises in nearest neighbor methods because relevant subspaces are needed to find appropriate neighbors, but determining the relevancy of subspaces assumes appropriate neighborhoods have already been established. Additionally, the number of subspaces to explore is prohibitively large (i.e., another facet of combinatorial explosion). There are 2^A-1 possible subspaces to examine, if attribute dimensionality is considered globally. That total is multiplied by the number of cases if relevant subsets are examined on a case-by-case basis. Both numbers are computationally prohibitive in high dimensionality situations. The number of possible subspaces explodes if additive combinations of dimensions (such as affine linear transformations) are permitted. The systematic exploration of numerous candidate subspaces also leads to the following problem.

Data snooping bias (also known as the problem of multiple comparisons or hypothesis tests) is a byproduct of combinatorial explosion, wherein the likelihood that a point will appear to be an outlier in at least one dimension increases exponentially as the number of dimensions increases, thereby increasing the false positive rate. That is, given enough dimensions, at least one subspace can always be found such that each point appears to be an outlier in it. Irrelevant dimensions only exacerbate the problem.

Difficulty in interpreting outlier scores arises from the concentration of distances and resulting concentration of outlier scores. While a meaningful ranking of outlier scores may still be generated, establishing cut-offs separating inliers from outliers becomes increasingly difficult at higher dimensionalities.

In contrast to common concerns in high-dimensional Euclidean spaces, the authors of "A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data"[69] argue that these problems are often misunderstood and overstated. In particular, highly correlated components greatly reduce the severity of distance contraction/concentration, noise effects, and attribute irrelevancy problems. As long as a large proportion of the attributes are relevant and highly correlated, the effects of noisy and irrelevant attributes should be negligible. Indeed, highly correlated attributes should create beneficial point concentration, such that inliers are concentrated together and outliers are very distant from them. We know that at least for modern PDB depositions (say the last 10-15 years) validation scores are highly correlated, which suggests that dimensionality reduction may not be necessary. However, since ICA tends to produce heavy-tailed distributions (see section 5.5), performing it on an unreduced set of attributes could exacerbate the production of spurious outliers resulting from combinatorial explosion.

2.4.6 Has anyone attempted to find anomalies in the PDB before?

There have been broad descriptions of statistical patterns of some of these quality indicators, but outlier detection has not been applied. Attempts to separate high and low-quality models, such as PDB SELECT, have used simple cut-offs, such as resolution better than 2Å and R-factor better than 20%, or used simple formulations of resolution, R-factors, and some stereochemical scores, such as ASTRAL[70,71]. Their chief disadvantage is in their simplicity; they include only a small number of validation scores and ignore contextual clues such as year of deposition. More recently, Read and Kleywegt demonstrated preliminary work in case-controlled (i.e., contextual) validation[14]. However, they neither combined validation scores into a single metascore, nor reported the discovery of any notable outliers. Even so, their results did inform my selection of contextual attributes (section 5.7).

The closest comparison to the work described here is that by Brown and Ramaswamy[72]. They used multiple linear regression (a form of semi-supervised outlier detection) to predict "structure quality metrics" from contextual information such as resolution, x-ray source, and deposition date. Based on these regression equations, the residuals for each metric were calculated across the whole PDB and converted to z-scores (having zero mean and unitary variance). This made metric values from any given models comparable, regardless of context. PCA was then applied to the z-scores (a form of outlier detection in subspaces), and the eigenvectors accounting for 50% of the variation were kept. Lastly, the Euclidean distance from the origin to the transformed points served as a single scalar score of model "quality" (a form of distance-based outlier detection).

I posit that the Brown and Ramaswamy method is inadequate and prone to error in at least three ways. First, they included only nine validation scores. Second, there is considerable non-linearity in several of the validation score relationships (a flaw also mentioned by Read and Kleywegt, who were also inspired by this work). Third, they only kept the components responsible for 50% of the variation, whereas 80-90% is more typical. I suspect that only three components were chosen to make visualization simple. However, much information may have been lost as a result. Had the authors kept the first six components, 79.4% of the variance would have been accounted for - 88.7% if they would have kept seven. Note that while there is a somewhat steep drop in the scree plot from the seventh to the eighth component, seven dimensions is not a significant reduction from nine. It is not clear what additional benefit reducing dimensionality via PCA provided over simply calculating 9D Euclidean distances.

In contrast, though my methods involve the use of attribute transformation and the selection of component subspaces for dimensionality reduction, I my choice of ICA over PCA was guided by properties of the former I suspected would be conducive to outlier detection, namely dimensions that are approximately statistically independent and have distributions with long tails. Furthermore, by performing the projection separately for each context I can potentially find small-scale near-linear relationships between attributes, even if the global data are non-linearly related. It also allows for the underlying latent dimensions of potential model defects to be determined with different emphases on attributes in each context. I have not attempted linear combination of contextual attribute values to predict validation score values, and while overall scores for degree of "outlierness" are obtained from my method, no attempt is made to define or measure quality, *per se*, since the term is highly context-dependent, rather loaded, and deceptively subjective[c].

I considered including Brown and Ramaswamy's quality metric among my chosen validation scores, but the concerns noted above and those addressed by Read and Kleywegt's work cast doubt on the validity of their methods and results. I generally preferred to avoid composite scores, unless their methods had been demonstrated to be based on up-to-date knowledge, and their results were demonstrably valid and useful.

[c] Nevertheless, the use of "quality" in the literature of structural biology is pervasive and nearly ubiquitous, making it difficult to avoid entirely.

2.4.7 How did I get here?

Starting with my master's project, completed in November 2007, and continuing until late 2011, I worked on a project called "Omega". Driven by many of the same motivations as the work described here, the goal was to create a single score by which protein structure models in the Protein Data Bank could be ranked for the purpose of keeping models with scores worse than a certain threshold out of structural bioinformatics studies.

Omega was presented as a new, composite measure of model validity, based on the assumption that a model having relatively good scores on multiple individual measures is less likely to be problematic than one with relatively poor scores. I included validators that are well represented in the PDB and would (by general agreement of structural biologists) certainly be included in a more thorough (manual) evaluation. They represent a necessary, but neither complete nor exhaustive, set of validation criteria.

Though this composite validator bore some resemblance to an outlier detection score, I did not frame the problem that way. Rather, I trapped myself with the mistaken intention of combining multiple validation scores into a single estimate of model quality. Unfortunately, "quality" is a contentious concept, and one that is defined somewhat subjectively and according to intended model uses[12,14,72-75]. Framing the problem in terms of outlier detection allowed me to identify potentially problematic structure models based on a variety validation scores without trying to both define and measure quality. I also avoided subjective weights for validation scores, which would require a pool of experts greater than one (my advisor) to properly evaluate.

Omega's other fatal flaw was its mathematical formulation. It was designed as a weighted sum of rank and indicator flag scores, which in turn was ranked. Ranking expressed how a score compared to others in the database (or subset), and flags expressed whether a score surpassed a level beyond which a structural biologist would begin to regard a model's validity with suspicion. The architecture of a weighted linear sum was based on a mistaken assumption that validation scores are strongly correlated and should always or nearly always increase or decrease in values together. That assumption was shown to be inappropriate when a decade-long fraud, involving twelve PDB entries and hundreds of citations, was revealed in 2009[76]. A defining characteristic of these fraudulent models is physically nonsensical inconsistencies in validation scores. Under the Omega formulation, the resulting mixture of very good and very poor scores averages out to overall scores that are "middle of the road", rather than obvious outliers.

In order to fix Omega so it could handle inconsistencies, I attempted to develop a consistency component for the formula. This led to various experiments with variances, root-mean-square differences, maximum deviations, and factor analysis (such as principal components analysis). The PCA produced some tantalizing possibilities, but the fraudulent models had too many close neighbors even in the transformed feature spaces to be consistently discriminative as outlier detectors. The false positive rate would have been unacceptably high.

In an attempt to find more distance between fraudulent models and others likely to be either normal or honest mistakes, I accelerated a step that was planned anyway. This was to incorporate more (and more varied) validation scores, as well as to reduce the ambiguity of all scores. For instance, one score improved by fitting a regression line between it and another score. A new score was then formed from the residuals, which were found to be correlated with well-established validators.

Despite some partial successes, I unfortunately had to accept that the Omega project could no longer proceed, having realized that its flaws were too deep and serious. Fortunately, however, I realized that my efforts, as what I had been doing in calculating flag scores, fitting regression lines, and performing factor analysis was implicit outlier detection.

2.5 A BRIEF HISTORY OF ANOMALOUS PROTEIN STRUCTURE MODELS

> *"In the first 50 years of biomacromolecular crystallography, there has probably not been a single error-free structure (in the sense that it cannot be improved upon, now or in the future). This means first of all that crystallographers must do their utmost to find and (if possible) fix the major errors in their model prior to deposition and publication. However, since there is no way of keeping even highly suspicious models out of the public database (or of evicting them), users of structural information should also find out about potentially problematic aspects of any model they intend to use as a molecular-replacement probe, to design mutants or ligands, to produce homology models, to compare with related structures or to simulate. In other words, validation is crucial for both the producers and the consumers of biomacromolecular structures and validation tools should be used both to assess the overall quality of a model and to assess the reliability of particularly interesting aspects (active-site residues, interface residues, ligands, inhibitors, cofactors etc.)."*[15]

2.5.1 Early Difficulties

In the late 1980s and early 1990s, several deposited structures were found to be incorrect. The errors were serious, significantly affecting conclusions drawn from the structures, but they were honest mistakes. Blame was placed on limitations in the experimental data, lack of adequate validation tools, and pressure to be first to publish[77,78]. They precipitated a series of papers vigorously calling for routine and thorough validation of published models, reporting of informative validation scores in publications, and mandatory deposition of structure factor data (i.e., observed diffraction intensities/amplitudes)[8,18-21,79]. In response to this minor crisis in crystallography, a number of new quality indicators and validation tools were developed[29,31,42,54,55,68,80-95]. Additional errors have been found over time, but few have had significant effects on structural biology or related fields.

As of 1998, the crystallographic community still felt some hostility toward routine validation of structures at deposition and the mandatory submission of coordinates prior to journal publication[92]. They were reluctant to release their coordinates before publication for fear of being scooped by competitors and out of desire for exclusive rights to examine details of their structure before anyone else. They objected to routine validation because structures are solved, published, and deposited for a variety of reasons and uses. Structures are not always deposited in a fully refined state, and they argued that it is not always possible or desirable to refine all structures with extreme care. Rather, it is often enough to affirmatively answer the question, “Do the data presented justify the conclusions drawn?” Lastly, the quality of a structure model is limited by the quantity and quality of the underlying data. However, it was decided that it was important to provide at least a minimal guarantee of the validity/accuracy of models to the broader scientific community, because not all end users of structure models have the training necessary to reliably distinguish acceptable models from unacceptable ones.

Routine validation for PDB depositions has since become an accepted and unremarkable practice. More recently, debates arose over the deposition of structure factor files. However, that too is now mandatory and accepted as well. The current debate is over whether to require the archival and availability of raw diffraction image files. At the time of this writing it is not clear whether that, too, will become routine, since the space requirement per frame of data is very high, making this a rather expensive proposition.

Though routine validation prior to deposition has been performed for more than a decade, it is still important to revalidate older models using modern tools. Older structure models lack the accuracy and precision afforded by modern software packages (from data collection to validation), so they are suboptimal by modern criteria and generally receive poorer scores from validation tools than models based on similar macromolecules and diffraction of similar quality[32,34,96-98]. Likewise, today's optimal models will be suboptimal by future standards. Improvement of macromolecular structure determination methods over time stands in stark contrast to chemical crystallography, which has essentially been using the same algorithms for structure determination since the mid-1980s[99].

2.5.2 Concerns About Automation

At the beginning of the new millennium, concern was expressed over increasing use of automated structure determination methods. Some feared that, "Although increased automation might result in a reduction of human errors during model building, it may equally well lead to an increase of errors if too much faith is put in results obtained with magical black boxes."[17] A decade later, these fears seem to be somewhat ill-founded, as structures produced by the various structural genomics pipelines around the world have been shown to be as good as or better than those produced by hypothesis-driven structural biology[14,72].

2.5.3 Spectacular Retractions

However, concerns that vigilance in validation should be renewed were not entirely unfounded. The year 2007 brought the "great pentaretraction"[23,100-123], in which five high-profile journal articles and their associate PDB entries were retracted after the authors discovered serious flaws in their structure models. Most of the retracted structures were incomplete (having only C_α atoms in the backbone and no side chains) and reported resolutions too poor for use in this study. The paucity of detail in models presented as only C^α coordinates "substantially undermines" efforts to independently assess their validity[118]. Nevertheless, they illustrate well the need for more thorough validation and outlier detection. Indeed, not only was the work and career of the lead investigator set back, so was the work of other labs working on the same family of proteins[108].

The mistake made by the authors was seemingly simple. A program developed in-house for converting intensities to amplitudes inverted the signs on Friedel pairs of reflections (equivalent reflections generated by crystallographic symmetry). That is, intensities $I\left(hkl\right)$ and $I\left(\overline{h}\overline{k}\overline{l}\right)$ became amplitudes $F\left(\overline{h}\overline{k}\overline{l}\right)$ and $\mathrm{F}\left(hkl\right)$. This led to the building of structure models into inverted maps with incorrect topologies, such that all atom locations were essentially wrong[118]. The resulting relatively high R-factors were erroneously attributed to intrinsic crystal disorder. A novel technique, called multi-copy refinement, was used to get the R-factors into a publishable range. Multi-copy refinement is a method that uses an ensemble of non-interacting models to account for certain forms of disorder[124]. Unfortunately, the observation-to-parameter ratio is reduced in proportion to the number of copies, which is potentially problematic even at high resolutions[125]. The single-copy R-factors had been alarmingly high and should have been cause for concern, especially when combined with rather poor resolutions. Furthermore, the random selection of the free set of reflections was inappropriate for the non-crystallographic symmetry (NCS) present[126].

Two years later, an unrelated structure was retracted because of gross inaccuracies. Its replacement was then also retracted due to lack of precision in functionally important regions[7,24,127,128]. Far worse than either of the preceding embarrassments, however, was the case of fraud made public at the end of 2009.

2.5.4 Documented Frauds

After suspicions were raised in 2007 by researchers working on some of the same structures[129], structure models deposited by H.M. Krishna Murthy were subjected to integrity investigations by the University of Alabama Birmingham. In an admirably thorough report, the investigators recommended that several papers be retracted and that the PDB expunge eleven models deposited over approximately a decade (a twelfth had already been retracted by the author)[13] in late 2009. Numerous physically improbable and impossible features were highlighted, including poor covalent geometry, poor core packing, chemically impossible close-contacts between non-bonded atoms, improbably high solvent contents (and bizarre solvent characteristics), B-factor profiles not reflective of distance from the core or solvent accessibility, anomalously poor

Ramachandran plots at given resolutions, anomalously good agreement between models and diffraction data with given poor stereochemistry, and physically improbable gaps in the lattice.

The PDB responded to the controversy by saying that models would only be expunged when the affected journals formally requested a retraction[130], and by forming a Validation Task Force to make recommendations for the simplification and wider use of validation tools[12]. It should be noted that at the time of this writing eight of the offending structures had not been retracted.

With the exposure and ramifications of UAB fraud still fresh in minds of crystallographers and those who consume their structure models, another fraud has been revealed in 2012[131-134]. This case was found via "validation by re-refinement"[131,134]. Specifically, a routine search of the PDB-REDO database[32,96-98,135] for a particular birch pollen protein revealed a deposition with a significant discrepancy between reported R-factors and unexpectedly low (for the resolution) R-factors calculated for a conservative re-refinement. This discovery led to a more detailed investigation of the structural details.

Numerous side chains did not fit the experimental electron density. Furthermore, several atom occupancies are set to zero in unreasonable ways. Ordinarily, zeroed occupancies are used to account for poorly defined side chain atoms in the electron density, due to a high degree of disorder or multiple conformations, instead of allowing atomic B-factors to be large. Even used correctly this is practice may not appropriate, because validation tools frequently omit "unoccupied" atoms from geometry checks. There is little support for physically highly improbable application to main-chain atoms, such as some C^{β} and backbone O atoms in the suspect model[134]. Suspicion was increased when it became apparent that the inclusion of unoccupied atoms did not result in unaccounted-for density in a $2mF_{obs} - DF_{calc}$ difference map.

Like Krishna-Murthy's models, the effects of bulk solvent, which adds noise to the diffraction pattern, seem to be non-existent; this is a red flag. Signal-to-noise ratios were also an order of magnitude higher than expected for real observed diffraction data (F_{obs}). Assuming that the reflections reported as observed were actually calculated (F_{calc}), further re-refinement was performed, which produced near-impossible R_{free} and R_{work} values of 0.040 and 0.019, respectively. These values are within the range of values associated with small molecule structures[136]. It was concluded that this was another case of fraud, in which no actual diffraction experiment took place and calculated structure factors were used in place of observed data.

The investigator points out that nearly all of the recent significant controversies resulted from the desire to confirm existing hypotheses (i.e., confirmation bias), and should have been avoided through critical investigation of the underlying macromolecular structures. Indeed, they could have been avoided through use of Bayesian methods, which combine prior knowledge with current data to produce posterior probabilities for posited structural models[7].

One prominent structural biologist responded to the recent controversies saying, "It is plain that improved techniques are needed, so that these problems can be uncovered before the structures get into the databases."[99] I agree in substance, but not in particulars. I agree that such problems can and should be discovered prior to deposition, or at least prior to journal publication. I disagree that new *validation* methods must be developed to uncover fraud. Rather, the combination of multiple existing methods into new *metavalidation* methods is likely to be quite adequate. Indeed, most of Krishna Murthy's structures are worrisome outliers according to several criteria and seemingly suspiciously good according to others; together they paint a picture of physical implausibility. Furthermore, most of the validation tools used to investigate the suspected fraudulent models were readily available at the time of their deposition and publication, and other similar tools, some rather venerable, would have also flagged them[92].

3.0 THESIS

3.1 INTRODUCTION

Protein crystallography's experimental processes have "satisficing"[137] _ENREF_1 quality to them, due to the costly nature of optimization. Conscientious structural biologists follow community-enforced standards of best practice for proper structure model building, refinement, and validation, as expressed in peer-reviewed publications and shared domain lore. However, "best" is a moving target that changes as tools and techniques improve. Consequently, it is not uncommon for researchers to refine their models only as well as their immediate uses demands and aim for validity no worse than their contemporaries would achieve (on average) with the same experimental data. This behavior should be evident in common patterns of model validation score values.

In a 2009 article, "On vital aid: The why, what and how of validation", Kleywegt laments the "creative" ways that crystallographers "manage to produce new types of errors with regularity" and that "there is no way of keeping even highly suspicious models out of the public database (or of evicting them)"[15]. Nevertheless, the kind of "vital aid" he rightly believes validation provides in producing better models and detecting errors is implicitly dependent on the vast majority of models being acceptable and having a great deal of consistency and regularity with respect to their validation scores.

Published in the same issue as “On vital aid”, Read and Kleywegt presented “Case-controlled structure validation” as a proof of concept for validating structure models with respect to those of similar age, diffraction data quality, and size[14]. Expressing an opinion that goes back at least xx years, they present ideal model building and refinement as doing “the best that could possibly be achieved with the amount of information available or at least the best that can be achieved by the most talented crystallographers using current methods”. However, they report that crystallographers typically only aim to do as well as “the average seen so far in comparable structures in the PDB”. This “schooling” behavior hints at common validation scores associated with informal standards. Additionally, their discussion of a lack of consensus regarding the reporting of disordered regions weakly implies that there *is* community consensus with respect to other aspects of structure refinement and validation. Lastly, the existence of comparable controls, against which models with unusual properties can be compared, implies strong commonalities among inliers.

The high-profile Validation Task Force, in its 2011 report on best practices for performing validation and presenting results, highlighted “isolated instances of high-profile structures that are entirely incorrect, incorrect in essential features, or likely fabricated”[12]. If instances of high-profile errors and fabrications are isolated, this implies that the vast majority – nearly all – of the models in the PDB is not substantially incorrect or fabricated. This is an indirect way of saying that competent crystallographers are following accepted standards of practice. The task force, which included Read and Kleywegt, advocated resolution-relative validation as a means to help a depositor “to judge how well the model approaches the best that could be achieved with the experimental data using current refinement methods and to catch slip-ups”, and an end user to “choose widely among similar deposited structures”. This technique relies on the fact that there are known current standards of “best” refinement. Additionally, the task force recommended presenting each validation criterion as a point on a distribution, in addition to reporting the raw scores. On the one hand, the value of presenting points in comparison to probability distributions is dependent on scores being tightly distributed around ideal values. On the other, shifts in those distributions over time mean model refinement and validation standards are changing with time. Thus, “commonness” is not a static concept, with respect to patterns of validation scores resulting from consensus standards.

3.2 CONSENSUS PRACTICES, COMMON VALIDATION SCORE PATTERNS, AND ANOMALIES

Hypothesis 1 Consensus Practices and Common Patterns

If community-accepted consensus standards regarding validation practice are evident in patterns of validation scores, then violations of those standards will appear as unusual combinations of validation scores.

Hypothesis 2 Anomalous Validation Score Patterns

Unsupervised anomaly detection methods can reliably identify unusual patterns of validation scores of protein structure models in the PDB.

3.3 DETECTING DEVIATIONS FROM COMMON PATTERNS

The aim of my research was to discover patterns in protein structure model validation data that would allow structural biologists to triage tens of thousands of models and quickly identify those sufficiently unusual to warrant closer inspection. To the best of my knowledge, explicit outlier scoring on PDB data never been attempted before.

I first tuned and tested the reliability of my anomaly detection methods by finding predetermined outliers in benchmark data (sections 4.1, 5.4, 5.5, and 6.0). I then applied those methods to finding unusual patterns of validation score values for protein structure models in the PDB (sections 4.2, 5.6, 5.7, 5.8, 5.9, and 6.0). "Unusual" was qualitatively defined as consistently poor or highly inconsistent score values. These unusual patterns are often associated with physical absurdities in models. However, not all unusual patterns are problematic. Nonetheless, they all require closer expert inspection, because rare, but real, structural features that produce anomalous validation scores are diagnosed by exclusion of less benign anomalies. Method reliability was further verified by confirming correlation between extremely poor "gold standard" validation scores and very high outlier scores, by ensuring that the models with the most extremely high outlier scores are demonstrably anomalous according to domain theory sections (5.8 and 6.5).

Unfortunately, outlier scoring is sufficiently opaque, having no explanatory power or value, that it would engender healthy skepticism in any domain, including structural biology. Therefore, I hypothesized that disjunctive sets of conjunctive IF-THEN rules could be used to post-process outlier scores, associating them with values from small groups of validators, and thereby making them comprehensible and explainable.

Support for this hypothesis was established in three ways. The first was to demonstrate better-than-random accuracy for rules predicting outlier scores from benchmark data (sections 6.4.1 and 6.4.2). The second was to do the same for protein structure model validation and outlier scores (sections 6.4.3 and 6.4.4). The third was to subjectively assess whether rules generated from validation scores and PDB models were consistent with domain theory (section 6.5).

4.0 DATA COLLECTION

4.1 BENCHMARK DATA

4.1.1 UCI Machine Learning Repository

Benchmark data were obtain from the UCI (University of California Irvine) Machine Learning Repository, “a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms”[138]. Selection guidelines for data sets were comparable dimensionality, comparable cardinality, few or no missing values, and real-valued attributes, though none of the sets meet all of the criteria.

4.1.2 Wine Quality

The wine quality data set is the union of two sets representing samples of red and white Portuguese wine samples. The original purpose of the data was to model and predict wine quality (as defined by experts’ subjective assessment) from physicochemical test measurements[139]. There are eleven input attributes and one target attribute. The inputs are real-values, and the target is an integer in the range 1 to 10. No definition of “anomaly” was deposited for the set, so I labeled the lowest and scores (3 and 9, respectively) as outliers.

For this, and all subsequent data sets, inliers were coded as 0.0, and outliers were coded as 1.0. This was to facilitate MSE calculations from probabilistic outlier scores.

4.1.3 SECOM

The SECOM set contains data from a semiconductor manufacturing process. Signals were collected from sensors and process measurement points, containing a combination of useful information, irrelevant information, and noise. No citation request was provided on the set's repository page.

The original purpose of the data set was to test feature selection methods and their effects on the accurate prediction of yield excursions downstream in the process. The 590 input attributes are real-valued, and the target attribute is a binary indication of pass/fail for each product on the line (case). Failures were treated as outliers.

4.1.4 Communities and Crime

The communities and crime (hereafter "crime") data set combines unnormalized socio-economic data from the 1990 Federal Census and law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey[140-142]. Data from the 1995 FBI Uniform Crime Report were also provided, but they were not used due to numerous incomplete cases. The original problem was to use regression to predict crime rates from demographic data.

2215 cases were used, represented by 147 attributes. Of the 147 attributes, 28 were rejected due to too many missing values. Another 18 were potential targets, representing raw and per 100,000 population counts of various violent and nonviolent crimes. Of these, I only retained the per

100K population violent, nonviolent, and total counts. The final set had 101 input attributes and 3 target attributes. Separate outlier scoring runs were performed for each target. I labeled the top and bottom 1% of each count as outliers.

4.1.5 Cardiotocography

The cardiotocography data set contains 2126 processed fetal heart rate and uterine contraction recordings made during childbirths. The original purpose was to develop automated analysis tools for cardiotocograms to be used in clinical settings[143]. The 22 input attributes consist of integer or real-valued diagnostic measurements. The target attribute corresponds to three fetal distress statuses: normal, suspect, and pathologic. Pathologic cases were coded at outliers, and the rest were inliers.

4.2 PDB DATA

Protein structure models built from x-ray diffraction data were obtained from the Worldwide Protein Data Bank[144,145] (wwPDB). See section 2.2.1 for a description of how these models are produced.

4.2.1 Protein Structure Models

Protein structure models and diffraction data were obtained using an rsync[146] script acquired from the RCSB PDB[147]. The advanced web query interface was used to acquire a list of PDB IDs associated with entries containing protein, produced via x-ray crystallography, refined to 4Å or better, and deposited on or before December 31, 2011. This query resulted in 68,359 entries. An additional 1686 entries were acquired from the RCSB's repository of obsolete entries (as of April 24, 2012), bringing the total number of entries to 70,055.

4.2.2 Deposited Validation Scores and Descriptive Data

Using the list described above, validation scores and descriptive data were downloaded from the RCSB PDB in the form of several reports. Some of the fields refer to the structure as a whole, and others refer to particular chains or entities. Only those fields addressing whole structure models were utilized in this study.

The unit cell dimension report provides angles α, β, and γ; lengths a, b, and c; space group; and Z number (space group multiplicity). Asymmetric unit volume is calculated from these values as follows.

$$V_{unit} = abc\sqrt{1+2\cos(\alpha)\cos(\beta)\cos(\gamma)-\cos^2(\alpha)-\cos^2(\beta)-\cos^2(\gamma)}$$

Equation 5 Formula for the volume of a crystal's unit cell, based on its dimensions (a, b, and c) and its vertex angles (α, β, and γ)

$$V_{asym} = V_{unit}/Z$$

Equation 6 Formula for the volume of a crystal's asymmetric unit, based on its unit cell volume and the multiplicity of the asymmetric unit

The refinement details and parameters reports provide the resolution, R_{all} (the residual difference between observed and calculated structure factors in the absence of a validation set), R_{work}

("training" set residual), R_{free} ("validation" set residual), structure determination method, and the number of reflections used for refinement. R_{all} was used to populate empty R_{work} fields.

4.2.3 Verify3D

Verify3D[81,148,149] is a validation tool that evaluates the compatibility of a protein's amino acid residues with their associated environments in the crystal. Residues are determined to be exposed, polar, or buried, and the reported score is the sum of log-transformed joint probabilities for residue types and environments. Of the 70,055 entries, Verify3D was successfully run on 62,474 (89%) of them.

$$\mathrm{E}\left[Verify3DScore \mid length\right] = .455 * \mathrm{length} - 2.047$$

Equation 7 Expected value of the Verify3D score, given the number of residues in the model, based on linear regression using 2011 data

$$\Delta V3D_{frac} = \frac{\left(Verify3DScore - E\left[Verify3DScore \mid length\right]\right)}{E\left[Verify3DScore \mid length\right]}$$

Equation 8 Fractional difference between actual and estimated Verify3D scores

4.2.4 Errat

Errat[29] identifies possible model errors by highlighting improbable relative frequencies of non-bonded atomic interactions. The score for a model is the percentage of residues within the 95%

contour of a multi-dimensional Gaussian based on the relative frequencies of non-bonded carbon-carbon, carbon-oxygen, carbon-nitrogen, oxygen-oxygen, oxygen-nitrogen, and nitrogen-nitrogen interactions (using only five frequencies, because the sixth is redundant). Before use in this study, scores were subtracted from 100 so that higher quantities are qualitatively worse. Errat was successfully run on 69, 271 entries (99%).

4.2.5 ProSA

ProSA (Protein Structural Analysis)[31,150] evaluates the correctness of a model's fold by calculating pseudo-energies for the given fold and decoys, based on potentials of mean force for C_β-C_β atomic interactions. Three energies (and their associated z-scores based on a database of known protein structures) are reported: using a distance-based pair potential, using a potential that captures the solvent exposure of protein residues, and using a combination of the two. ProSA2003 was successfully run on 65,334 entries (93%).

$$E\left[ProSA_{energy}^{combined}\,|\,length\right] = -.725 * length\ -\ 23.926$$

Equation 9 Expected value of combined ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data

$$E\left[ProSA_{energy}^{surface}\,|\,length\right] = -.071 * length\ -\ 2.601$$

Equation 10 Expected value of surface ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data

$$E\left[ProSA_{energy}^{pair}\,|\,length\right] = -.371 * length\ -\ 10.920$$

Equation 11 Expected value of pairwise ProSA energy score, given the number of residues in a model, based on linear regression using 2011 data

$$E\left[ProSA_z^{combined} \mid length\right] = -.450 * \sqrt{length} \ - \ 1.978$$

Equation 12 Expected value of combined ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data

$$E\left[ProSA_z^{surface} \mid length\right] = -.333 * \sqrt{length} \ - \ 1.242$$

Equation 13 Expected value of surface ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data

$$E\left[ProSA_z^{pair} \mid length\right] = -.402 * \sqrt{length} \ - \ .765$$

Equation 14 Expected value of pairwise ProSA z-score, given the number of residues in a model, based on linear regression using 2011 data

$$\Delta ProSA_{frac} = \frac{\left(ProSA - E\left[ProSA \mid length\right]\right)}{E\left[ProSA \mid length\right]}$$

Equation 15 Fractional difference between actual and estimated ProSA energy or z-scores

4.2.6 TAP

TAP (Torsion Angle Propensity)[151] evaluates the likelihood of a model's stereochemistry being correct. This is achieved by calculating the conditional probability of seeing a particular combination of (discretized) torsion angles given a particular residue type, converting those probabilities to pseudo-energies, and reporting the normalized cumulative pseudo-energy for an entire structure model. Before use in this study, scores were subtracted from 100 so that higher quantities are qualitatively worse. TAP was successfully run on 67,956 entries (97%).

4.2.7 MolProbity

MolProbity[152] (implemented as the ramalyze, rotalyze, cbetadev, and clashscore routines of Phenix[153]) evaluates aspects of a model's stereochemistry using all atomic contacts, including likely hydrogen positions, whether they are deposited with the model or calculated by MolProbity. To avoid any attempt to alter or improve upon a deposited model, the reduce routine (which adds explicit hydrogen atoms), was run with the "NOADJust" option. This means that no attempt is made to optimize any atomic coordinates by changing bond lengths, or bond angles, or flipping side-chains.

MolProbity performs checks on main-chain torsion angles (i.e. Ramachandran analysis), reporting the percentage of φ-ψ angle pairs that fall within regions that are favored, disfavored (but allowed), and disallowed (due to steric clashes). Similarly, the rotamer analysis reports the percentage of side-chain torsion angle pair (χ_1-χ_2) outliers. Deviations from ideal C_β positions are reported as outliers if they are $\geq$ .25 Å. When non-donor–acceptor atoms overlap by more than 0.4 Å, a clash is recorded, and the clash score reports the number of such clashes per 1000 atoms. Lastly, the overall *MolProbityScore* is a sum of log-transforms of the preceding scores, weighted to be highly correlated with reported resolution (Equation 16).

$$
\begin{aligned}
MolProbityScore = \; & 0.42574 * \log\left(1 + clashscore\right) \\
& + 0.32996 * \log\left(1 + \max\left(0, \%RotamerOutliers - 1\right)\right) \\
& + 0.24979 * \log\left(1 + \max\left(0, 100 - \%RamaFavored - 2\right)\right) \\
& + 0.5
\end{aligned}
$$

Equation 16 Formula for overall MolProbity Score

MolProbity was successfully run on 70,001 entries (nearly 100%).

4.2.8 ProCheck

ProCheck[84] evaluates aspects of a model's stereochemistry. The counts and percentages of φ-ψ angle pairs that fall within regions that are core (most favored), additionally allowed, generously allowed, and disallowed (due to steric clashes). Similarly, the rotamer analysis reports the number and percentage of side-chain torsion angle pair (χ_1-χ_2) outliers, as well as the standard deviations of χ_1 (gauche+, gauche-, trans, and pooled) and χ_2 (trans). The standard deviations of the ω torsion angle (evaluating peptide bond planarity) and the ζ notional torsion angle (using the tetrahedron of C_α, N, C, and C_β to evaluate chirality) are also reported.

The energetic favorability of the secondary structure characteristics of a model are evaluated by measuring the energy of intra-backbone hydrogen bonds. The whole-model standard deviation of these energies is reported in kcal/mol.

Covalent geometry can be evaluated by examining the reported counts and percentages of main-chain bond length and bond angle outliers. Counts and percentages of planar group outliers, in aromatic rings (Phe, Tyr, Trp, His) and planar end-groups (Arg, Asn, Asp, Gln, Glu) are also reported.

Log-odds scores for stereochemical parameters are reported as G-factors (‘G’ is for “geometry”). They reflect how “normal” a given score is with respect to observed distributions, with low scores indicating potentially problematic aspects of stereochemistry. These scores were omitted from this study because they are based on outdated reference distributions. However, they could easily be incorporated into follow-up studies.

ProCheck was successfully run on 69,870 entries (nearly 100%).

4.2.9 SFCheck

SFCheck[154] reports several indicators of diffraction data quality, model quality, and agreement between diffraction data and model. Aspects of diffraction data quality include high and low resolution limits; the number of reflections; the number of acceptable and unacceptable reflections; the number of negative intensities; the number of reflections with $I > \sigma(I)$; the number of reflections with $I > 3*\sigma(I)$; $R_{stand}(I) \equiv \langle\sigma(I)\rangle / \langle I\rangle$; $R_{stand}(F) \equiv \langle\sigma(F)\rangle / \langle F\rangle$; optical resolution (the approximate width of an atomic peak); expected optical resolution for complete data; $B_{overall}$; $B_{overall}$ by Patterson method; P_{add}; the number of possible reflections; the number of observed reflections; fractional completeness; effective resolution (nominal resolution divided by the cube root of fractional completeness); minimal estimated coordinate error; and the anisotropic distribution of structure factors (expressed as the ratio of eigenvalues for the major and minor axes of an ellipse).

Aspects of model quality and description include the number of atoms, the number of water molecules, average B-factor, standard deviation of B, Matthews coefficient, percent solvent, reported (nominal) resolution, reported R-factor (R_{work} if R_{free} is present), the number of chains, refinement program used, refinement resolution range (high and low resolution limits), and reported R_{free}.

Aspects of data and model agreement include recalculated R-factor (R_{work}), real space correlation factor, recalculated R-factor using structure factors with $I > 2*\sigma(I)$, the number of reflections with $I > 2*\sigma(I)$, recalculated R_{free}, recalculated R_{rest} (i.e., R_{work}), the number of free reflections, the number of rest (work) reflections, real space correlation factor for structure factors with $I > 2*\sigma(I)$, Luzzati coordinate error, Patterson scale factor, Patterson B_{temp}, anisothermal scaling terms, solvent correction parameter K_s, solvent correction parameter B_s, maximal estimated coordinate error, and diffraction precision index (DPI).

Of the PDB entries in the query described above, 58,828 (84%) also had structure factor files deposited and were successfully run through `SFCheck` and other validation tools. For the purposes of this study, only these entries were retained. Prior studies have indicated that structure models deposited without structure factors often have more suspicious validation scores[17].

4.2.10 WhatCheck

`WhatCheck`[155] reports several aspects of model description and quality. Aspects of the diffracted crystal include `CRYST` card details from the PDB file, the molecular weight of all polymer chains, the volume of the unit cell, space group multiplicity, reported and calculated Matthews coefficient, the number of amino acids in all chains, and the number of water molecules

Aspects of a model’s B-factors include M-factor (a measure of B-factor variation in a model), the percentage of buried atoms with B less than 5Å, B-factor RMS z-score, the number of bonds, and the average difference in B over a bond.

Aspects of stereochemistry include bond length RMS z-score, bond length RMS deviation, bond angle RMS z-score, bond angle RMS deviation, chirality average deviation, improper dihedral RMS z-score, τ angle RMS z-score, Ramachandran z-score, χ_1-χ_2 correlation z-score, backbone conformation z-score, total bump (steric clash/overlap) value, sum of all overlaps exceeding 0.4 Å, total bump value per residue, a count of bumps, total squared bump value, counts of bumps in bins of differing severity, inside/outside RMS z-score, average structural packing score, average packing scores/z-scores for interactions of backbone and/or side chain atoms, the number hydrogen bond donors and acceptors, the number of buried and acceptors donors with any hydrogen bond, the number of buried donors and acceptors with a very poor or poor bond, and the number of buried donors and acceptors with essentially no or no bond.

`WhatCheck` was successfully run on 65,021 entries (93%).

4.2.11 Constructed Validation Scores

In addition to the validation scores downloaded from the PDB and those extracted from various program outputs, I have constructed additional scores. They represent combinations of extracted scores and attempt to provide validation information not present in the scores as extracted.

4.2.11.1 Optical Resolution Estimation

In addition to scores directly extracted from `SFCheck` logs, and combinations of them, the expected value of optical resolution was calculated from nominal resolution using linear regression (Equation 17).

$$E\left[d_{optical} \mid d_{nominal}\right] = .600 * d_{nominal} + .393$$

Equation 17 Formula for the estimation of optical resolution from nominal resolution

4.2.11.2 B-factor (Atomic Displacement Parameter) Statistics

The B-factors listed in PDB entries are obtained by minimizing a complex residual. Regardless of the precise residual being minimized, they are all subject to the problem of multiple local or false minima, as well as human error. The refinement process and its pitfalls have been the subject of much discussion in the literature[156-159]. The salient point here is that the B-factors appearing in the PDB are effectively "empirical" parameters that have emerged from the refinement processes. Because PDB B-factors are obtained in this way, many circumstances other than atomic dynamics affect their values significantly, making them problematic to compare directly between models.

The following validation scores are based on the calculation of atomic displacement parameters (ADPs) from reported B-factors, related statistics, a regression fit of those ADPs and optical resolution, and ratios of ADP statistics for main and side chain atoms. Their development will be described in detail in a forthcoming paper.

The atomic displacement parameter, u_{atomic}, for each atom in a PDB entry (structure model) is defined by the following equation:

$$B_{atomic} = 8\pi^2 u_{atomic}^2$$
$$u_{atomic} = \sqrt{B_{atomic} / 8\pi^2}$$

Equation 18 Definition of atomic displacement parameters calculated from atomic B-factors

The model average atomic displacement parameter, u_{model}, is also defined.

$$u_{model} = \langle u_{atomic} \rangle = \left\langle \sqrt{B_{atomic}} / 8\pi^2 \right\rangle$$

Equation 19 Definition of mean model atomic displacement

I define $E[u_{model} \mid optical]$ as the value of u_{model} predicted by weighted least squares regression of all u_{model} values *vs.* optical resolution (using models deposition between 2007 and 2011). The optimal exponent in the weighting coefficient, $\left(1/d_{optical}\right)^{power}$, was found (via maximum likelihood) to be 3.7.

$$E[u_{model} \mid optical] = .439optical - .091$$

Equation 20 Regression equation for estimating mean atomic displacement from optical resolution

I define Δu as the difference between u_{model} and $E[u_{model}]$.

$$\Delta u = \frac{u_{model} - E\left[u_{model}\right]}{E\left[u_{model}\right]}$$

Equation 21 Formula for fractional delta u, based on a model's mean u and an estimate of u calculated from optical resolution

I define u_{main} and u_{side} as the average u_{atomic} values for main-chain and side-chain atoms, respectively. I define u_{ratio} as the ratio of u_{main} to u_{side}.

$$u_{ratio} = u_{main} / u_{side}$$

Equation 22 The ratio of main-chain to side-chain atomic displacement parameter means

I define σ_{main} and σ_{side} as the standard deviations of u_{atomic} u_{atomic} values for main-chain and side-chain atoms, respectively. We define σ_{ratio} as the ratio of σ_{main} to σ_{side}.

$$\sigma_{ratio} = \sigma_{main} / \sigma_{side}$$

Equation 23 The ratio of main-chain to side-chain atomic displacement parameter standard deviations

4.2.12 Missing Values

The attribute transformations described in section 5.5 require complete cases. Unfortunately, it is not uncommon for validation tools to leave some or all validation scores missing. Small numbers of score values may be missing due inapplicability. For instance, some of values produced by `WhatCheck` are reported for all evaluated model, but others are only reported if they exceed programmed thresholds. Cases of complete failure for particular tools are likely caused by at least one of four possible situations. Namely, there may be a bug in the tool, there may be "clerical" errors causing a model to not conform with established file format standards, a tool may predate changes to standard file formats (such as `Verify3D`), or there may be irregularities in a model that violate programmed expectations. This last case could form the basis of an anomaly detection method, and would be an interesting addition to the work presented in this study.

One more cause of missing values only affects obsoleted models. Several values are queried from the RCSB's PDB web site, which does not report data from obsoleted models. Those have to be extracted using one's own parser. I did not choose to do that, however. Since some of the obsoleted models may be erroneous or fraudulent, excluding them from analyses reduces the risk of introducing bias.

Future studies will likely examine the significant effects (if any) of excluding attributes that are frequently lacking values. These attributes can still be included in rule learning (section 6.0), however, since missing values can be accommodated there.

4.2.13 Missing Value Summaries and Complete Attribute List

Complete lists of missing value summaries[d] and validation score attribute names[e] can be found in the supplementary material.

[d] http://d-scholarship.pitt.edu/19601/10/SuppF_missing_PDB_value_summaries.txt

[e] http://d-scholarship.pitt.edu/19601/9/SuppE_complete_PDB_attribute_list.pdf

5.0 OUTLIER SCORING

5.1 LOCAL OUTLIER FACTOR

Outlier scoring was performed using Local Outlier Factor (LOF). LOF, introduced by Kriegel, et al., at Ludwig Maximilian University of Munich, is based on the assumption that anomalies lie in regions of feature space with lower density than neighboring regions[160].

It is non-parametric in the sense that no parametric distribution assumptions are required. However, there are in fact two parameters, one explicit, and the other implicit. The explicit parameter is the minimum cardinality of the neighborhood to which a point is compared (*MinPts*, or *k* in k-nearest-neighbors). The implicit parameter is the distance (or dissimilarity) measure used. Implementations of LOF typically have Euclidean distance (the L_2 norm) hard-coded in, and this is what I used. However, with a little code editing this can be easily changed to any symmetric and non-negative dissimilarity measure (which need not be a true metric).

$$\begin{aligned}&k\text{-}distance(A) \equiv \text{distance from } A \text{ to its kth nearest neighbor}\\&reachability\text{-}distance_k(A,B) \equiv \max\{k\text{-}distance(B), distance(A,B)\}\end{aligned}$$

Equation 24 Definition of reachability-distance

Reachability distance is defined as being to a point A from a point B (Equation 24). If A is within B’s k-neighborhood, the reachability distance from B to A is B’s k-distance. Otherwise, it is the distance between A and B. That is, the reachability distance to a point A from any given neighbor, B, is greater than or equal to the distance between B and its kth nearest neighbor.

The local reachability density of a point *A* is defined as the reciprocal of the average reachability distance to *A* from each of its *k* nearest neighbors (Equation 25). The local outlier factor of *A* is then defined as the ratio of the average local reachability density of *A*'s *k* nearest neighbors to the local reachability density of *A* itself (Equation 26). A ratio of 1.0 indicates that a point's local density is the same (on average) as that of its neighbors. Values much larger than 1.0 are outliers, but the threshold separating inliers from outliers varies according to *k* and particulars of different data sets, including dimensionality and distribution (see section 5.2 for details). Values less than 1.0 are potential hubs (see section 5.9).

$$
\begin{aligned}
&lrd(A):\ \text{local reachability density of } A \\
&\left|N_k(A)\right|:\ \text{cardinality of the k-neighborhood of } A \text{ (including ties)} \\
&lrd(A) \equiv 1 \Big/ \left\langle reachability\text{-}distance_k(A,B) \right\rangle_{B \in N_k(A)} \\
&lrd(A) \equiv \left|N_k(A)\right| \Big/ \sum_{B \in N_k(A)} reachability\text{-}distance_k(A,B)
\end{aligned}
$$

Equation 25 Definition of local reachability density. (Angle brackets indicate average.)

$$
LOF(A) \equiv \left\langle lrd(B) \right\rangle_{B \in N_k(A)} \Big/ lrd(A)
$$

Equation 26 Definition of local outlier factor (LOF)

To produce LOF scores, I used the Data Mining With R package, DMwR 0.3.1[161].

5.2 OUTLIER SCORE STANDARDIZATION

5.2.1 Introduction

Scores produced by many outlier scoring algorithms are not easily interpreted on their own, or directly compared between sets. This is because the distributions of scores are often highly dependent on the data dimensionalities, attribute value distributions, and distance functions[162]. This problem is directly relevant to my use of contextual subsets (see section 0). In order to merge multiple outlier scores calculated for the same case, scores calculated from different sets of attributes and neighboring cases must be directly comparable.

LOF scores are difficult to interpret because they are calculated from a ratio that has a lower asymptotic bound of zero and an infinite upper bound. Points in inlying clusters are expected to have ratio values near 1.0. However, the spread of scores around that value varies, for the reasons just described, and the threshold separating inliers and outliers is data-dependent. Generally speaking, the variance of LOF scores is proportional to data set cardinality and dimensionality (Figure 1).

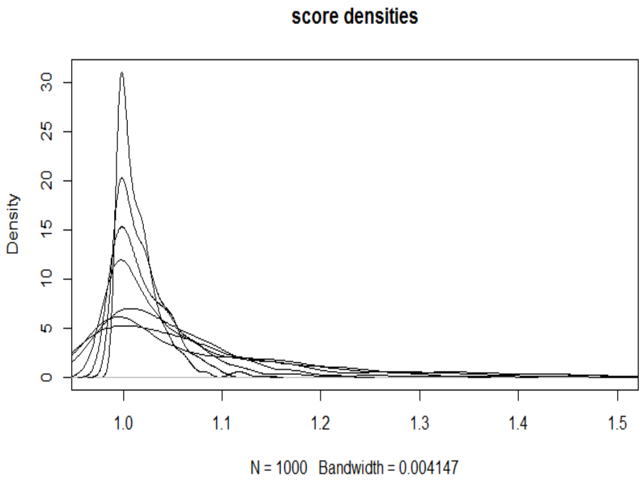


Figure 1 Distribution of LOF scores (*MinPts* = 100) for 1000 randomly generated Gaussian data points, with dimensionalities equal to powers of 2 from 1 to 7

LOF score distribution characteristics appear to be consistent with one of the curses of dimensionality, the concentration of distances/outlier scores[69]. Ideally, outliers should be well separated from inliers; scores should carry the same meanings between sets, and have a finite range. Unfortunately, the boundary between “true” outliers and inliers depends on the distribution of points in particular datasets. Therefore, I sought to fit probability families to distributions of scores, thereby facilitating their conversion to probability estimates.

5.2.2 Probability Distribution Fitting

5.2.2.1 Published Fits to LOF

Kriegel, et al., suggested fitting a Gaussian or gamma distribution to LOF scores. Both methods use sample statistics to estimate distribution parameters (Figure 2, Figure 3). Adequate (but not optimized) fits are reported. However, when I implemented the suggested fits, they were found wanting. These fits did not even pass visual inspections. However, I cannot in charity rule out either errata in the paper or in humility rule out misapplication of methods on my part.

$LOF_S(o)$: LOF score for any given case o in set S

$pLOF_S(o)$: probabilistic LOF score

$$pLOF_S^{Gauss}(o) := \max\left\{0, 2 * \mathrm{cdf}_S^{Gauss}(o) - 1\right\}$$

$$\mathrm{cdf}_S^{Gauss}(o) := \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{LOF_S(o) - \mu_S}{\sigma_S\sqrt{2}}\right)\right)$$

$$pLOF_S^{Gauss}(o) := \max\left\{0, \mathrm{erf}\left(\frac{LOF_S(o) - \mu_S}{\sigma_S\sqrt{2}}\right)\right\}$$

μ_S: mean LOF score for a set

σ_S: standard deviation of LOF score for a set

Figure 2 Calculation of probabilistic LOF scores according to a Gaussian distribution

$$pLOF_S^{gamma}(o) := \max\left\{0, \frac{\mathrm{cdf}_S^{gamma}(o) - \mu_{cdf}}{1 - \mu_{cdf}}\right\}$$

$$\mathrm{cdf}_{\mathrm{S}}^{\mathrm{gamma}}(o) := \frac{\gamma\left(k, LOF_S(o)/\theta\right)}{\Gamma(k)}$$

$$\hat{k} := \frac{\hat{\mu}^2}{\hat{\sigma}^2}$$

$$\hat{\theta} := \frac{\hat{\sigma}}{\hat{\mu}^2}$$

$$\mu_{cdf} = \mathrm{cdf}_{\mathrm{S}}^{\mathrm{gamma}}(\mu_S)$$

Figure 3 Calculation of probabilistic LOF scores according to a gamma distribution

Since `LOF` scores are ratios, the authors suggest alternative fits based on distributions for random variables formed from ratios of other random variables, and they specifically mention Cauchy (Figure 4) and Fisher-Snedecor (aka F) (Figure 5) distributions. However, maximum likelihood estimation (MLE), using the `fitdistrplus` package[163] in `R`, was unable to obtain an acceptable fit using either distribution family.

$$
\begin{aligned}
&X,Y \sim \mathrm{N}(0,1) \\
&X/Y \sim \mathrm{Cauchy}(0,1) \\
&f(x \mid x_0,\gamma) = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right] \\
&F(x \mid x_0,\gamma) = \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}
\end{aligned}
$$

Figure 4 Probability and cumulative density functions for a Cauchy distribution

$$
\begin{aligned}
&X \sim \chi^2_{d_1} \quad Y \sim \chi^2_{d_2} \quad \frac{X/d_1}{Y/d_2} \sim \mathrm{F}(d_1,d_2) \\
&X \sim \mathrm{Beta}\left(\frac{d_1}{2},\frac{d_2}{2}\right) \quad \frac{d_2 X}{d_1(1-X)} \sim \mathrm{F}(d_1,d_2) \\
&f(x \mid d_1,d_2) = \frac{\sqrt{\dfrac{(d_1\,x)^{d_1}\;d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\dfrac{d_1}{2},\dfrac{d_2}{2}\right)} \\
&F(x \mid d_1,d_2) = I_{\frac{d_1 x}{d_1 x+d_2}}\left(\tfrac{d_1}{2},\tfrac{d_2}{2}\right)
\end{aligned}
$$

Figure 5 Probability and cumulative density functions for an F-distribution. B is a beta function, and I is a regularized incomplete beta function.

5.2.2.2 MLE Fits to LOF and LOF-1

MLE was used to fit several additional distributions to LOF_S. Several distribution fits from the Amoroso family of distributions were attempted (Equation 28). The Amoroso distribution is a four parameter generalized gamma distribution[164-166]. At the time, I did not realize that the distributions I attempted to fit were part of a superfamily. Rather, I sought distributions with large positive skews, including Rayleigh, χ, χ^2, Maxwell-Boltzmann, gamma, inverse χ^2, inverse gamma, and Weibull. Fits were attempted with both LOF_S and $LOF_S - 1$. Figure 6 shows a Cullen and Frey plot[167] of 100 bootstrapped skew and kurtosis values for LOF_S scores against values associated with certain common distribution families. This plot is applicable to $LOF_S - 1$ as well, because skew and kurtosis calculations do not vary with simple shifts.

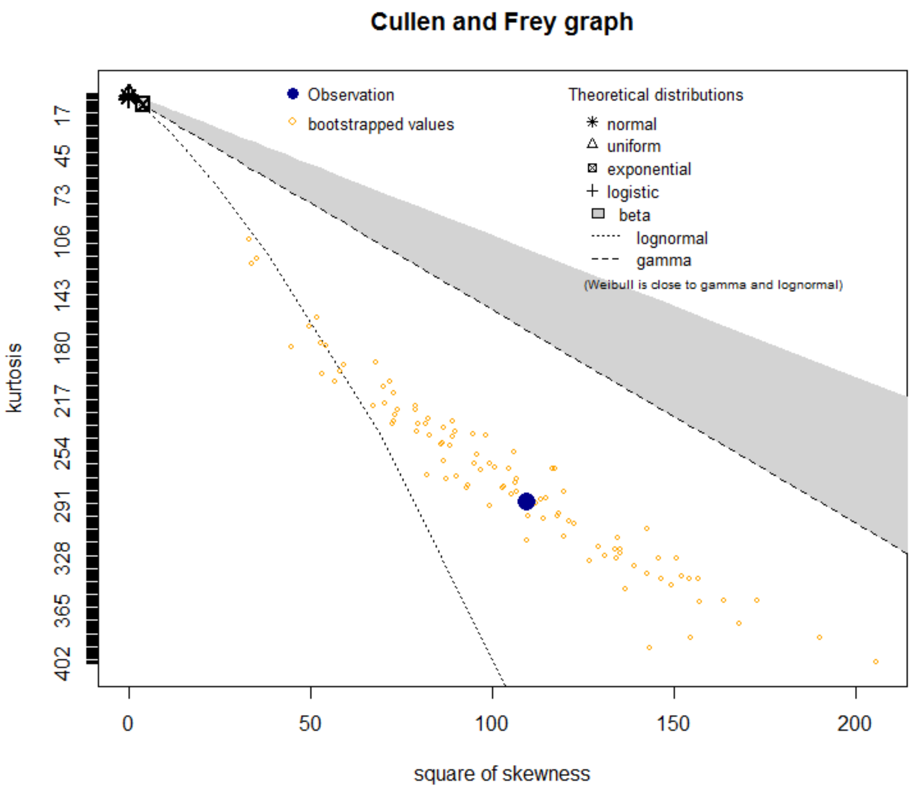


Figure 6 Cullen and Frey plot for LOF in the improper subset of the PDB data

The decision to explore the use of $LOF_S - 1$ was based on the threshold cluster membership, at which the ratio of the average neighborhood local density and the local density about a point is unity. My conjecture was that this value introduced a sort of bias to all LOF_S distributions, and furthermore, removing that bias would facilitate the fitting of parent distributions bounded at zero.

$$f\left(x\,|\,a,\theta,\alpha,\beta\right)=\frac{\left|\beta/\theta\right|}{\Gamma\left(\alpha\right)}\left(\frac{x-a}{\theta}\right)^{\alpha\beta-1}e^{-\left(\frac{x-a}{\theta}\right)^{\beta}}$$

$$F\left(x\,|\,a,\theta,\alpha,\beta\right)=\begin{cases}1-Q\left(\alpha,\left(\dfrac{x-a}{\beta}\right)^{\beta}\right) & ,\dfrac{\beta}{\theta}>0\\ Q\left(\alpha,\left(\dfrac{x-a}{\beta}\right)^{\beta}\right) & ,\dfrac{\beta}{\theta}<0\end{cases}$$

$$Q\left(\alpha,x\right)=\Gamma\left(\alpha,x\right)/\Gamma\left(\alpha\right)$$

$$x,a,\theta,\alpha,\beta\in\mathbb{R}$$

$$\alpha>0$$

$$x\geq a,\ \text{if}\ \theta>0$$

$$x\leq a,\ \text{if}\ \theta<0$$

Equation 27 Probability and cumulative distributions of the Amoroso distribution

Distribution	a	θ	α	β
Stacy (generalized gamma)	0	.	.	.
χ (χ^2)	0	$\sqrt{2}$ (2)	k/2	1
Rayleigh	0	.	1	2
Maxwell-Boltzmann	0	.	3/2	2
gamma	0	.	.	1
half-normal	0	.	1/2	2
log-normal	.	.	$\lim_{\beta\to 0}1/\left(\beta\sigma\right)^2$	.
(generalized) Weibull	.	.	(n) 1	>0
generalized extreme value (GEV)	.	.	1	.

Table 1 Amoroso distribution parameter values for a sample of members of the family

Goodness-of-fit statistics are reported in Table 2. Values reported are from the Kolmogorov-Smirnov, Cramer-von-Mises, and Anderson-Darling tests, all of which are methods to test whether two samples are drawn from the same distribution or a sample is drawn from a given hypothesized distribution. If the test statistic is less than the critical value for a given distribution, the null hypothesis of common distribution is accepted; othe[168]rwise, it is rejected. In none of the fits I produced was the null hypothesis accepted. So, instead of seeking statistical significance at some arbitrary level, I sought minimize all of the test statistics. Using the `fitdistrplus` package, one can fit parameters using maximum goodness-of-fit. However, I chose to use maximum likelihood for fitting and goodness-of-fit for evaluation as a way to avoid over-fitting.

The inability to find statistically significant parameter fits is not unexpected or even undesirable, since the disparities appear to be mostly limited to the upper tails (Figure 7, Figure 9). Since outliers are assumed to be generated by different phenomena, processes, and parent distributions, it is reasonable to expect that LOF score distributions containing true outliers would have heavier tails than a hypothetical outlier-free distribution. Put another way, distribution parameters are fit to inliers, and it is not surprising to find that outliers are not fit well.

In addition to objective goodness-of-fit statistics, fits were subjectively assessed by visual inspection of probability density function, cumulative density function, percentile-percentile (P-P), and quantile-quantile (Q-Q) plots. The first two compare empirical distribution functions to those hypothesized fits. P-P plots compare empirical and hypothetical distribution function values for each point in a sample. If the two distributions are equal, the points will fall on a line from (0, 0) to (1, 1). Similarly, Q-Q plots compare empirical and hypothetical quantiles, such that the points all fall on a 45° if the distributions are equal. Heavy right tails manifest as points above the line. Other deviations signal unequal skew, kurtosis, or dispersion[169].

The best fit of the distributions attempted at the time was with the Weibull distribution fit to $LOF_S - 1$ (Table 3). However, after the benchmark experiments were complete, another attempt was made to fit a gamma distribution, using MLE instead of the suggested parameterization described above. The goodness-of-fit statistics indicate that this gamma fit for $LOF_S - 1$ was marginally better than the Weibull (Table 2, Table 3). Since this was not known at the time, the gamma fit was not used for any experiments.

Distribution	Kolmogorov-Smirnov	Cramer-von-Mises	Anderson-Darling
Cauchy	0.1391	100.3	768.4
Normal	0.1420	205.6	Infinite
Gamma	0.0837	86.99	Infinite
Log-Normal	0.0682	58.02	Infinite
F	0.4517	2356	10770
Weibull	0.3109	689.0	Infinite
GEV	0.0187	3.9	27.5

Table 2 Goodness-of-fit scores for LOF using maximum likelihood estimation to fit parameters

Distribution	Kolmogorov-Smirnov	Cramer-von-Mises	Anderson-Darling
Cauchy	0.1452	104.7	791.9
Normal	0.1499	212.5	Infinite
Gamma	0.0246	7.518	Infinite
Log-Normal	0.0822	74.47	453.5
F	0.4183	1810	8185
Weibull	0.0368	15.49	Infinite

Table 3 Goodness-of-fit scores for LOF-1 using maximum likelihood estimation to fit parameters

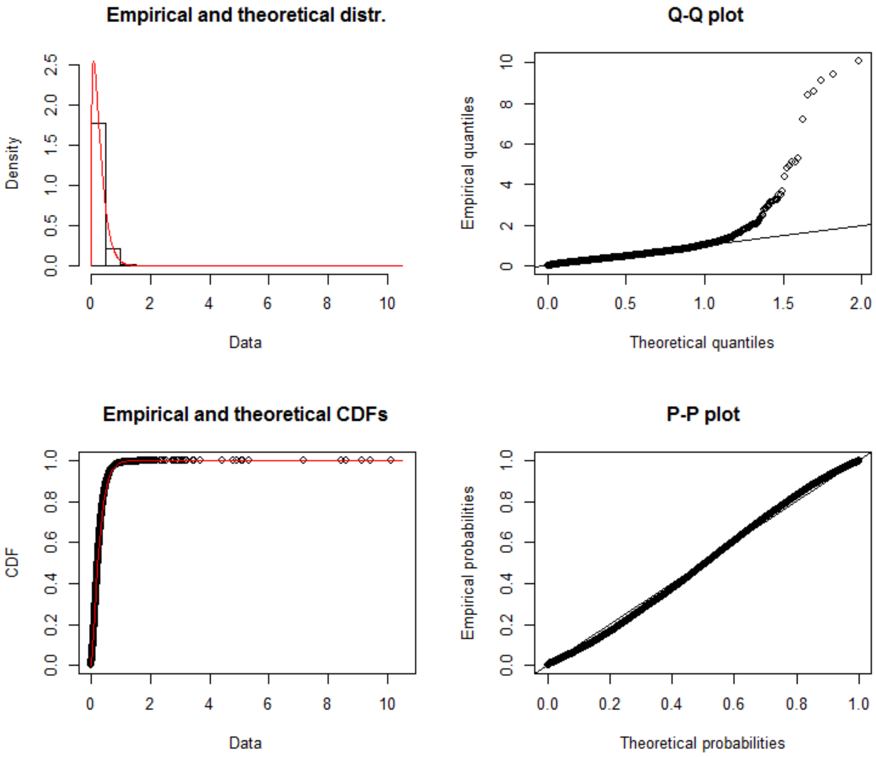


Figure 7 Weibull distribution fit plots for LOF-1 in the improper subset of the PDB data

5.2.2.3 MLE Fits to 1-1/LOF

The Weibull fit was used for all of the benchmark experiments. However, prior to beginning the PDB experiments an idea for a tighter fit came to me. It occurred to me that part of the difficulty may be due to the very low outlier score variance for very large, high-dimensional data sets (such as the PDB). This low variance, a manifestation of a curse of dimensionality, clusters scores very tightly around the inlier threshold of 1.0. Score values are approximately bounded by 1.0 on one end, and unbounded on the other. By inverting the ratio and subtracting from 1.0, values are approximately bounded at 0.0 on one end and bound at 1.0 on the other. The distribution of these values has much broader variance, and the range (0, 1) corresponds well to the beta family of distributions.

To test the viability of this change, I used MLE to find a fit for a Beta distribution and a similar distribution, Kumaraswamy, with Weibull as the control. A Kumaraswamy distribution is directly convertible to a beta distribution, but unlike a beta distribution, its probability density function has a closed form[170,171] (Equation 28). The Kumaraswamy distribution provided the best fit, both visually and according to statistics from Kolmogorov-Smirnov, Cramer-von-Mises, and Anderson-Darling goodness-of-fit tests. None of the tests showed distribution similarity to the p=.05 significance level. Nonetheless, the Kumaraswamy fit produced test statistics closest to that significance level (Table 4). It should be noted that for both $LOF_S - 1$ and $1 - 1/LOF_S$, values less than or equal to zero were omitted, in order to satisfy support restrictions for several candidate parent distributions.

Additionally, during the revision process a fit was made to the generalized extreme value distribution (Table 1) using LOF_S. This was found to be the best fit of all (Table 2). As with the gamma fit to $LOF_S - 1$ described above, it was not found in time to be used for this study. However, it will likely be applied in future work.

$$\begin{gathered}
f\left(x \mid a,b\right) = abx^{a-1}\left(1 - x^{a}\right)^{b-1} \\
F\left(x \mid a,b\right) = 1 - \left(1 - x^{a}\right)^{b} \\
X \sim \text{Beta}\left(1,b\right) \quad \Leftrightarrow \quad X \sim \text{Kumaraswamy}\left(1,b\right) \\
X \sim \text{Beta}\left(a,1\right) \quad \Leftrightarrow \quad X \sim \text{Kumaraswamy}\left(a,1\right) \\
\text{If } X \sim Beta\left(1,b\right), \text{ and } Y \sim Kumaraswamy\left(a,b\right), \\
\text{then } X^{1/a} = Y
\end{gathered}$$

Equation 28 Probability and cumulative distribution functions for the Kumaraswamy distribution, and the relationship between Kumaraswamy and beta distributions

Distribution	Kolmogorov-Smirnov	Cramer-von-Mises	Anderson-Darling
Normal	0.0385	17.89	124.4
Gamma	0.0604	35.67	207.6
Log-Normal	0.1077	133.9	800.8
Weibull	0.0286	6.908	46.69
Beta	0.0420	16.47	97.82
Kumaraswamy	0.0262	5.755	37.45

Table 4 Goodness-of-fit scores for 1-1/LOF using maximum likelihood estimation to fit parameters

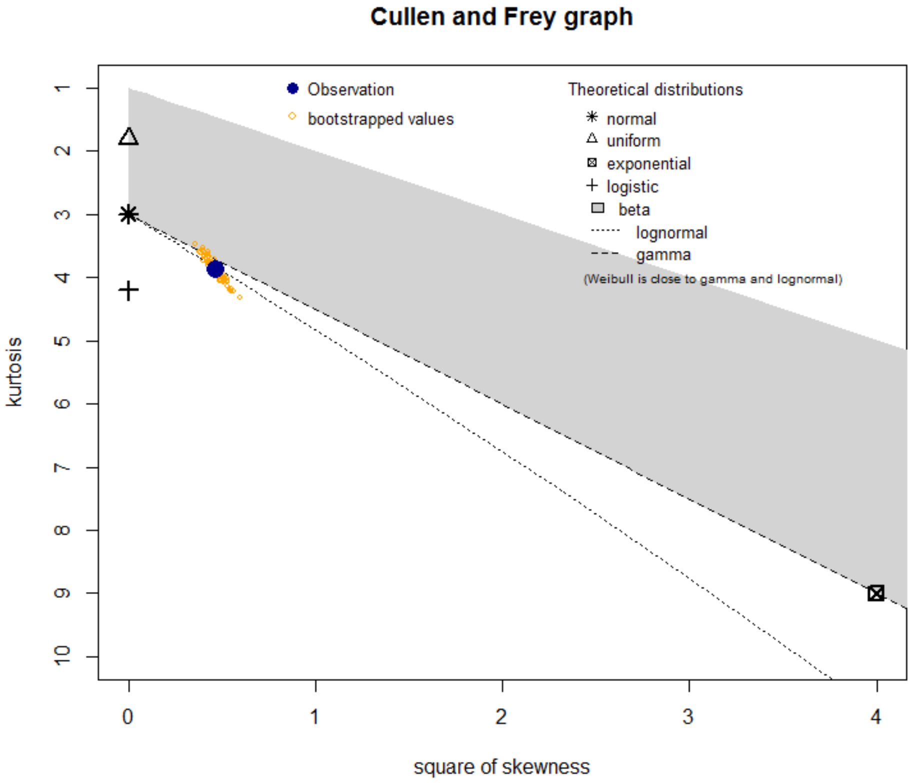


Figure 8 Cullen and Frey plot for 1-1/LOF in the improper subset of the PDB data

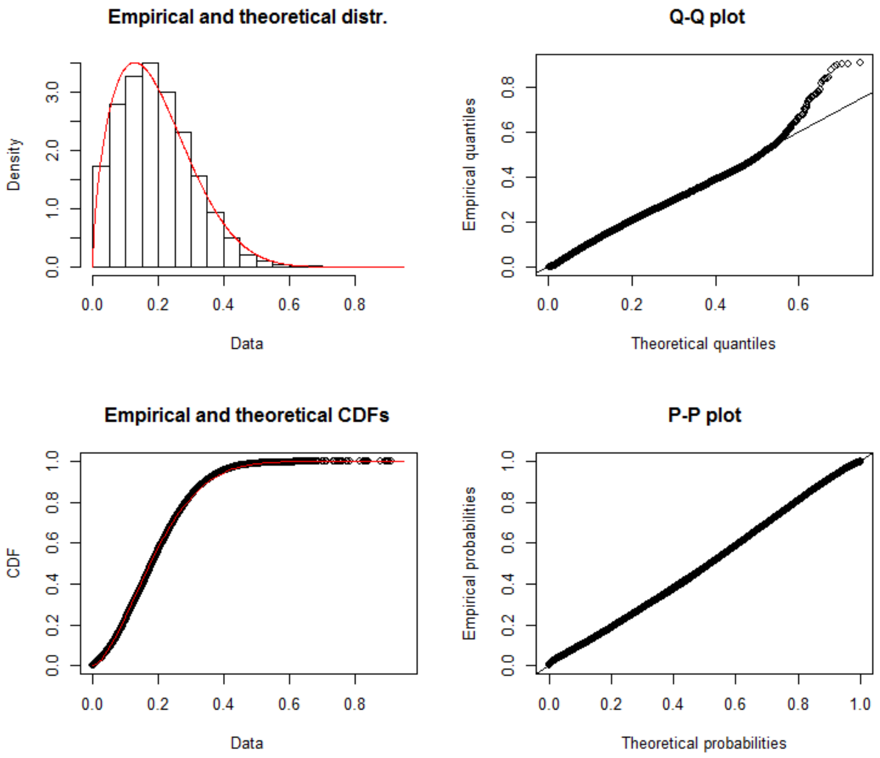


Figure 9 Kumaraswamy distribution fit plots for 1-1/LOF in the improper subset of the PDB data

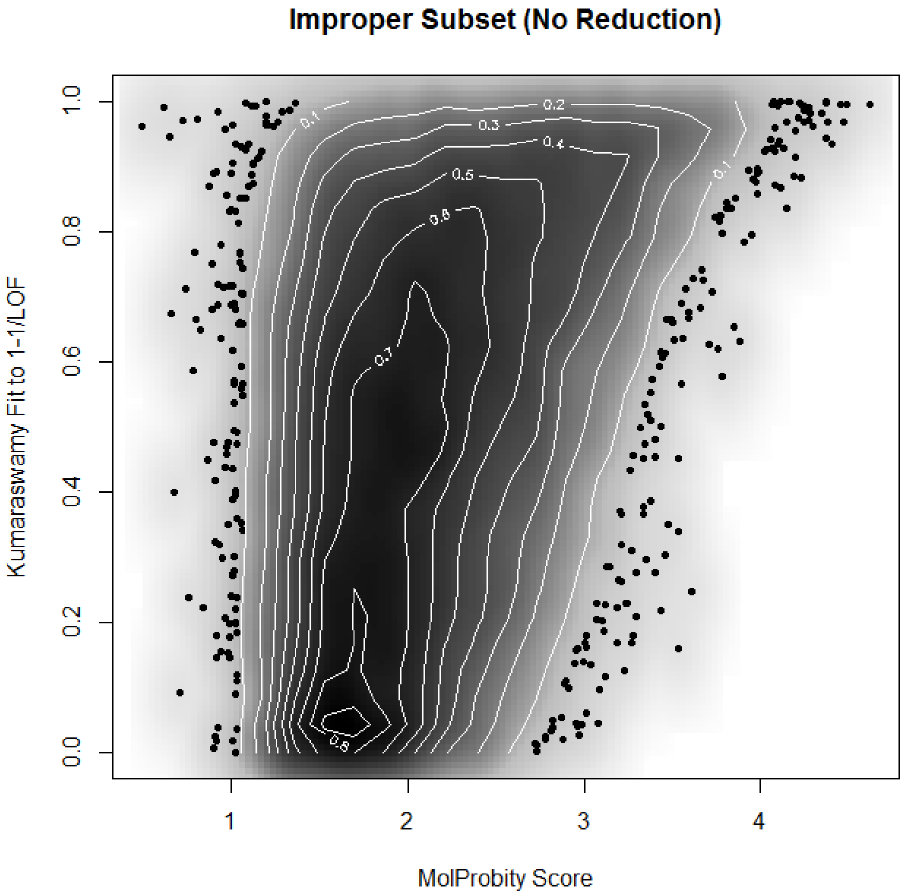


Figure 10 Combined scatter and contour plot of MolProbity score and *pLOF* using the 1-1/*LOF* fit to a Kumaraswamy distribution

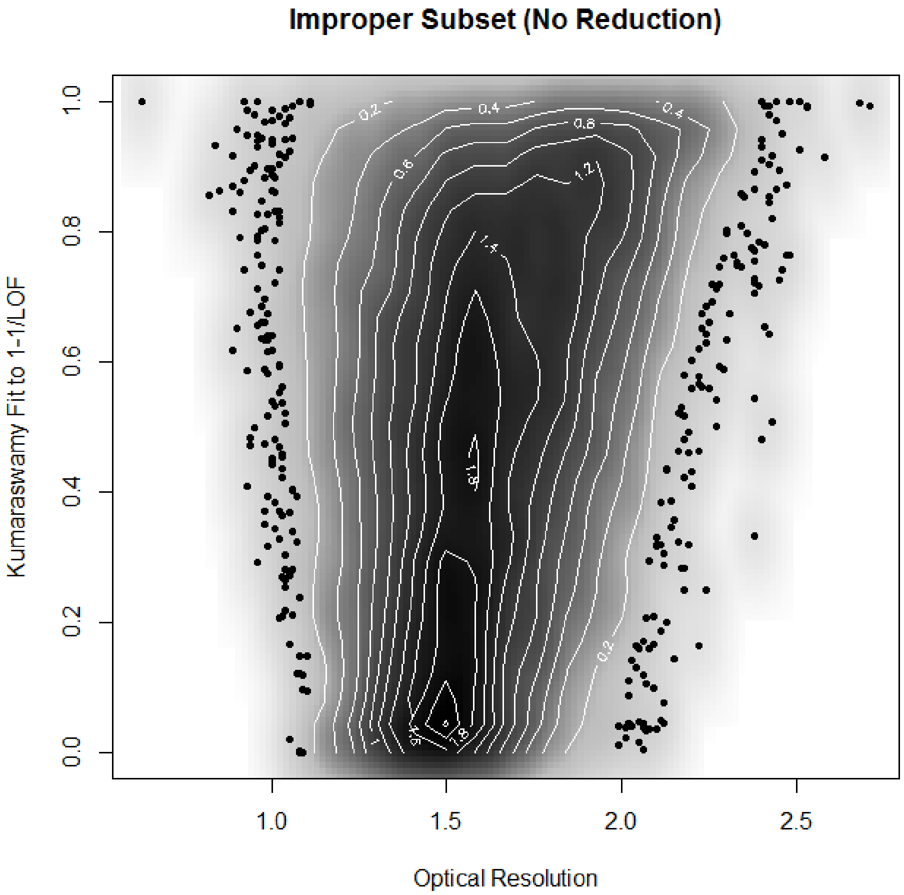


Figure 11 Combined scatter and contour plot of optial resolution and pLOF using the 1-1/LOF fit to a Kumaraswamy distribution

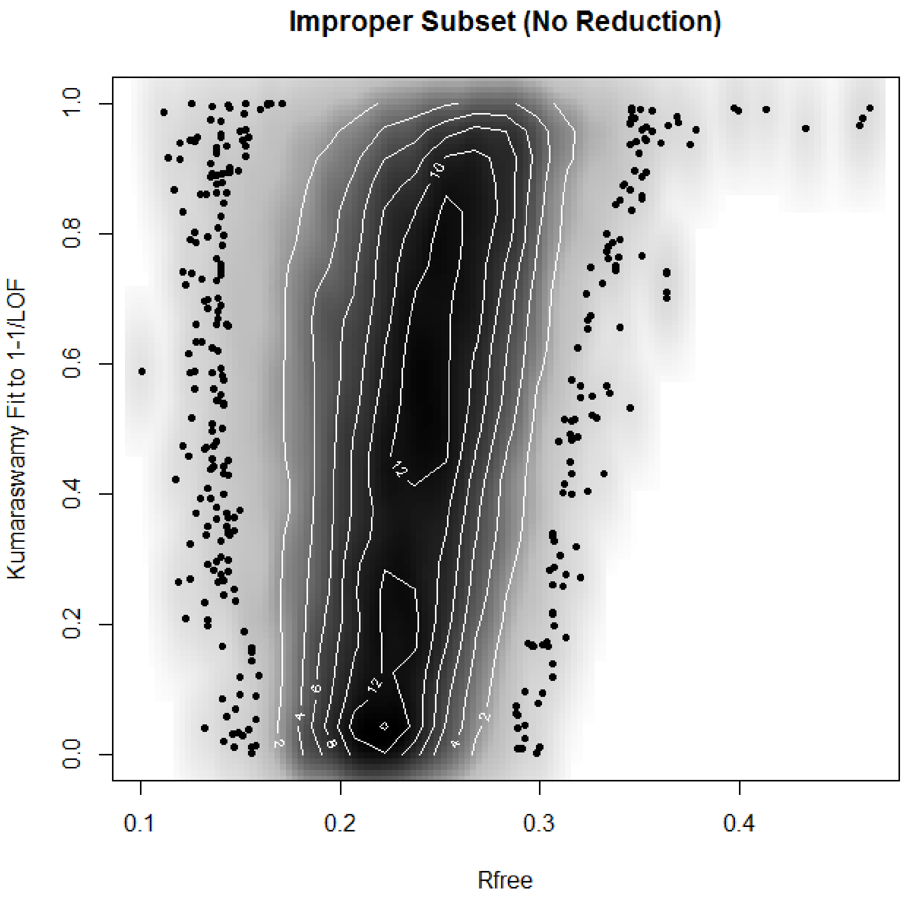


Figure 12 Combined scatter and contour plot of R_{free} and pLOF using the 1-1/LOF fit to a Kumaraswamy distribution

5.3 OUTLIER SCORING EVALUATION CRITERIA

5.3.1 Loss Function

In these data, the target class is binary, with inliers and outliers associated with 0 and 1, respectively. Estimations are in the continuous range [0, 1). False positives are defined as outlier scores greater than 0 for known inliers, and false negatives are defined as outlier scores less than 1 for known outliers. To evaluate the performance of outlier scoring for my benchmark tests, I needed to select a loss function. A loss function is an indication of the cost associated with an incorrect prediction or estimation. There are countless possible loss functions. I chose squared error because it is strongly affected by outliers. That is, large errors are treated as more costly than small errors. Thus, an outlier score of 0.8 for a known outlier is not twice as bad as a score of 0.9 for the same, but four times worse.

$$\hat{f}(X) := \text{ estimate or prediction of Y}$$

$$Y = \begin{cases} 1 & \text{, true outlier} \\ 0 & \text{, true inlier} \end{cases}$$

$$L_{SE}\left(Y, \hat{f}(X)\right) = \left(Y - \hat{f}(X)\right)^2$$

$$MSE = {}^{1}\!/_{N} \sum_{i=1}^{N} \left(Y_i - \hat{f}(X_i)\right)^2$$

Equation 29 Square error loss function and mean square error

5.3.2 Expected Cost

From these squared losses, mean squared error (MSE) was calculated separately for inliers and outliers (Equation 29). I did not calculate MSE for inliers and outliers together due to class imbalance. Almost by definition, inliers greatly outnumber outliers. If this class imbalance is not accounted for, very low error rates could theoretically be achieved by simply scoring all cases as inliers. Additionally, I calculated expected cost as a biased average of inlier and outlier MSE by making false negatives more costly. An unbiased average would simply be the arithmetic mean.

$$E\left[cost\right]_{unbiased} = .5MSE_{inlier} + .5 * MSE_{outlier}$$
$$E\left[cost\right]_{biased} = \left(N_{outliers}/N_{total}\right) * MSE_{inlier} + \left(N_{inliers}/N_{total}\right) * MSE_{outlier}$$

Equation 30 Unbiased and biased estimated costs based on means squared errors

The bias I chose was somewhat arbitrary and could be easily replaced with another; the point was to favor false positives (type I errors) over false negatives (type II errors). MSE for inliers was weighted according the fraction of cases that are known outliers, and MSE for outliers was weighted according to the fraction of cases that are known inliers.

5.3.3 Randomization Test

Once expected costs were calculated from benchmark experiments, I needed to determine whether they were better than those generated by random assignment of outlier scores. To achieve this, I uniformly reassigned known outlier and inlier labels (without replacement), calculated MSE values, and calculated expected cost. I repeated this process 100 times and calculated the average cost. By keeping the outlier scores fixed while re-assigning labels, I was able to observe whether sensitivity to outliers was a spurious product of the generating distributions.

The use of ROC (receiver operating characteristic) plots and calculation of the AUC (area under the curve) were considered. However, the AUC calculation (also known as the Mann-Whitney statistic) is insensitive to the relative sizes of the predicted class (i.e., their prevalence). When the target attribute is distributed with strong skew (typically with far more negatives than positives), an ROC curve will not differentiate between models that are more or less sensitive to skew.

Thus, a model with high recall (sensitivity, true positive rate) and low precision (positive predictive value) may not be distinguishable from one with lesser recall and greater precision with respect to AUC performance[172]. This is especially problematic for anomaly detection, which typically involves sets of true outliers that are vastly outnumbered by true inliers[173-175].

Calculating MSE for each class is just one of many possible alternatives to AUC. It was chosen for its simplicity and sensitivity to large deviations of estimated outlier probability from ground truth. In contrast to the latter trait, AUC does not take into account *how much* a continuous outlier score deviates from a true label, with deviations of all sizes treated equally. Even so, future studies may make use of alternatives to ROC curves, such as cost curves and precision-recall curves[174,176].

5.4 NEIGHBORHOOD SIZE EXPERIMENTS

The number of nearest neighbors queried, k (aka MinPts) used in most prior uses of LOF is less than or equal to 50, regardless of data set characteristics[177]. This seems counterintuitive, since neighborhood size effectively establishes the minimum cardinality of inlying clusters160. If it is too low, the number of false negatives will likely be inflated for outliers that cluster together. If it is too high, the number of false positives will likely be inflated for inliers in small clusters. It seems reasonable to posit that as cardinality of a dataset increases, so will the cardinality of inlying clusters, which suggests that neighborhood size should be proportional to dataset size.

5.4.1 Hypotheses

Hypothesis 3 Optimal Neighborhood Size

Fixed neighborhood sizes lower than needed for minimal mean-squared errors and expected cost in large datasets, and optimal neighborhood sizes should be proportional to dataset size[69,177,178].

5.4.2 Methods

To test these hypotheses, I varied the value of k by values of $10*2^i$, where *i* increments from 0 to the largest integer less than or equal to ten percent of the cardinality of a set. `LOF` was then run with each of those values of k for `MinPts`.

5.4.3 Results

Contra my hypothesis, the optimal value of `MinPts` appears to be in the range of 10 to 40 for most of the sets, regardless of data set size (Table 5). It should be noted that this conclusion is based on my chosen bias, which is to prefer false positives to false negatives. The variance of `LOF` scores demonstrably decreases with increasing values of `MinPts`, suggesting that sensitivity is traded for specificity. In a scenario demanding few false alarms, larger values of `MinPts` may be preferred, and this preference can be expressed in the form of higher weight placed on MSE for true inliers.

The only exception to this trend is the cardiotocography set. No experiments were performed to determine the cause(s) of this deviation. One conjecture would be that these data are comprised of multiple small and distinct clusters of inliers. In such a case, small neighborhood values would tend to produce high false positive rates. Alternately, the inliers may not be well clustered at all.

	k = 10	k = 20	k = 40	k = 80	k = 139	k = 160	k= 190	k = 213	k = 320	k = 640	k = 650
Wine Quality	0.15	0.12	0.12	0.14		0.16			0.18	0.19	0.19
SECOM	0.26	0.28	0.29	0.31	0.32						
Violent Crime	0.28	0.29	0.30	0.31		0.33	0.34				
Nonviolent Crime	0.20	0.22	0.23	0.23		0.25	0.25				
Total Crime	0.21	0.22	0.23	0.23		0.25	0.25				
Cardiotocography	0.33	0.27	0.24	0.22		0.18		0.17			

Table 5 Expected costs for the unreduced ICA run for each benchmark set. The complete set of tables can be found in Appendix A Error! Reference source not found.**.**

5.5 ATTRIBUTE TRANSFORMATION AND DIMENSIONALITY REDUCTION EXPERIMENTS

5.5.1 Introduction

Contrary to long-held concerns, recent research suggests that high dimensionality is not a curse, per se, but high *irrelevant* dimensionality, i.e., uncorrelated noise, is[69]. A common method of dimensionality reduction is to project multidimensional data into a transformed space of reduced dimension. Principal component analysis (PCA) is typically performed, and the eigenvectors associated with the top few eigenvalues are kept. Ironically, such stopping criteria (the heuristics guiding components rejection and retention) may be counterproductive, since outliers may be most recognizable in dimensions with low inlier variance. The components with the smallest eigenvalues account for the least proportion of variance, and as the variance for a component decreases to zero in the limit, the distribution of data in that dimension collapses to a delta function. This does not happen in real data, but if a component's inliers are very tightly distributed, outliers are likely to stand out very strongly. Even so, such low-variance components would typically be discarded, since common stopping criteria were formulated with characterization of inliers in mind, not outliers. Some recent preliminary work has explored the selection of components for the explicit purpose of outlier detection, such as by finding the dimensions of greatest contrast on a case-by-case basis[179].

Instead of PCA, I have chosen Independent Component Analysis (ICA). ICA is a linear transformation technique that performs blind source separation[180]. In a blind source separation problem, signals are mixed in some unknown way, and the task is to recreate the original unmixed signals. An example is the "cocktail effect" problem, in which multiple voices are mixed in a noisy environment and the original voices are sought. ICA is also related to projection pursuit, a statistical method that projects data into transformed dimensions that maximize some functional definition of interestingness[181]. In this case, interestingness is defined in terms of (approximately) maximizing statistical independence. Mathematical details of the method can be found below.ICA's emphasis on independence is desireable from a domain perspective since it would be interesting and informative to discover latent dimensions of anomalousness (a goal for future research). PCA only produces uncorrelated variables on orthogonal axes (second order independence), ICA produces (approximately) statistically independent variables without requiring orthogonality. The advantage in this is that independent variables are guaranteed to have zero correlation, but not all uncorrelated variables are independent (e.g., y_1=x and y_2=x^2).

Another advantage of ICA is that the transformed variables are maximally non-Gaussian. Recent research suggests that some attribute transformations may be resistant to the concentration of distances aspect of the curse of dimensionality. In particular, those that maximize kurtosis, the fourth central moment of a distribution (Equation 31) seem to be ideal[182]. Leptokurtic distributions have narrow peaks about their means and “fat” tails. Platykurtic distributions have broad peaks and “skinny” tails. Some ICA methods explicitly maximize positive kurtosis, and others tend to enhance kurtosis by maximizing negative entropy (both of which are zero only for Gaussian distributions). This tail-stretching is a very desirable quality for anomaly detection. PCA tends to produce spherical projections, keeping potentially outlying points close to the bulk of inlying points. In contrast, ICA’s non-Gaussian projections tend to push outliers deep into the tails, making them easier to identify.

$$\beta_2 = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^2} = \frac{\mu_4}{\sigma^4}$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

Equation 31 Classical (top) and modern (bottom) definitions of kurtosis. The classical definition is in terms of central moments. The modern definition is in terms of cumulants. The latter is actually excess kurtosis, with the kurtosis of a Gaussian distribution subtracted.

A third advantage of ICA is that PCA’s assumption of Gaussian sources and preference for maximizing variance can be detrimental to clustering. Since outliers are likely to lie outside of well-defined clusters or in small clusters distinct from larger inlier clusters, PCA’s assumptions may therefore be detrimental to outlier detection. The dimension of greatest discrimination is often not the dimension of greatest variance. For instance, in the case of prolate ellipsoidal clusters with parallel major axes, the axis of greatest variance is orthogonal to the axis of greatest cluster separation.

5.5.2 Hypotheses

Hypothesis 4 ICA Transformation

I posited that a kurtosis-emphasizing transformation, independent component analysis, would improve outlier detection in high-dimensional data sets. Support would be indicated by lower expected costs for ICA-transformed data than simply scaled data (the control condition).

Hypothesis 5 Simple PCA Dimensionality Reduction

If PCA-based dimensionality reduction is harmful to outlier detection, then expected cost would be higher for PCA-reduced data versus merely scaled data.

Hypothesis 6 ICA From Simple PCA Dimensionality Reduction

Proceeding from PCA to ICA may not only reverse the damage but also enhance outlier detection (as reflected in expected costs).

5.5.3 Methods

5.5.3.1 Scaling

For the control condition, attribute values were converted to z-scores, having zero mean and unitary variance. In the case of contextualized PDB data, means and standard deviations were calculated with respect to contextual subsets. Standardization was performed by running the `scale` command (of the `base` package of `R`[183]) on each attribute column.

5.5.3.2 PCA

For the PCA and ICA conditions, a Pearson correlation matrix was calculated for all pairwise complete cases using the `cor` command from the `base` package. “Pairwise complete” means that when calculating the correlation coefficient for a pair of attributes, cases are excluded only if they are missing a value for one of the attributes in the pair. In this way, incomplete cases contribute to the correlation matrix.

This correlation matrix is used to perform singular value decomposition using the `svd` command from the `base` package. SVD is a factorization of the form $\mathbf{M} = \mathbf{U\Sigma V}^*$, where the left-singular vectors of $\mathbf{M}$, the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{MM}^*$, the principal components, and the diagonal values of $\mathbf{\Sigma}$ are the singular values (the square roots of the eigenvalues). SVD and eigenvalue decomposition are equivalent when SVD is performed on a symmetrical square matrix. The developers of `R` recommend using their SVD function over their eigenvalue decomposition function for performing PCA, because the latter is more prone to rounding errors.

The loadings matrix, $\mathbf{U}$, was used to complete the PCA linear attribute transformation. The unreduced principal components were not used for outlier scoring, because PCA performs an isometric transformation, so distances calculated for nearest neighbors would not differ from those in the scaled condition.

For the reduced dimensionality benchmark conditions, eigenvectors (already sorted by decreasing eigenvalues) were retained or trimmed away according to the “broken stick model”. There is no uniquely correct or optimal method to reduce the number of components, but some studies have shown that the broken stick model comes closest to finding the “true” or “intrinsic” dimensionality of a feature space among simplistic methods[184,185]. It was originally developed as an ecological niche apportionment model[186,187], and gets its name from the analogy of breaking a stick in multiple places with uniform probability.

According to the broken stick model, the null hypothesis for the eigenvalue spectrum is that eigenvalues are generated from the sums of order statistics from a uniform distribution. Only components with eigenvalues greater than the expected null hypothesis values are retained. Order statistics are a series of random variables $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ that are the smallest, next smallest, etc., values drawn from independent identically distributed (IID) random variables. The length of the stick is equal to the total variance to be accounted for and is equal to the number of variables when PCA is performed using a correlation matrix.

Let $Y_1, Y_2, \ldots, Y_N$ be random variables representing the N null hypothesis eigenvalues. Since they are sums of IID variables, their expected values may be expressed as sums (Equation 32).

$$\mathrm{E}(Y_i) = \sum_{j=i}^{N} \frac{1}{j}$$

Equation 32 Expected value of each eigenvalue under the broken stick null hypothesis

As indicated in the following results section, simple dimensionality reduction did not benefit outlier scoring for the benchmark data. In fact, it degraded performance. Conjecturing that the Broken Stick Model was to strict and discarded too many components, a slight modification was made for the PDB data experiments. For those data, the Broken Stick Model was combined with the Guttman-Kaiser Criterion, which retains eigenvectors with eigenvalues greater than or equal to one[188] (assuming a correlation matrix is diagonalized, rather than a covariance matrix). The heuristic is that such eigenvectors carry as much or more signal than a single original attribute. This criterion has been criticized for retaining too many components, but it is possible that other methods retain too many.

5.5.3.3 ICA

In addition to being an end unto itself, PCA can also be a useful preprocessing step for ICA algorithms that perform better on pre-whitened data. “Whitened” means the covariance matrix for the data has been diagonalized by SVD to equal the identity matrix. Once my data were whitened through PCA, the dimensionality was reduced as explained above. The dimensions remaining after reduction are already uncorrelated and need only be rotated to maximize statistical independence. This approach to ICA puts it in the same class of factor rotation methods as varimax, quartimax, and equamax, except ICA does not seek maximal interpretability of components.

ICA assumes are a set of independent components are linearly combined to produce mixed signals, and the goal is to find an unmixing to recover the independent components. To describe this process mathematically, let us assume that we have *n* random variables corresponding to the mixed signals (and the same number of independent components). Each case in a set is therefore a vector of samples from each random variable, x_i (Equation 33).

$$x_i = a_{i,1}s_1 + \ldots + a_{i,k}s_k + \ldots + a_{i,n}s_n$$

Equation 33 Signals as linear combinations of independent components

$$\begin{aligned} x &= \left(x_1, \ldots, x_m\right)^T & s &= \left(s_1, \ldots, s_n\right)^T \\ a_k &= \left(a_{1,k}, \ldots, a_{m,k}\right)^T & A &= \left(a_1, \ldots, a_n\right) \\ x &= \sum_{k=1}^{n} s_k a_k & x &= As \end{aligned}$$

Equation 34 Signal mixing in vector and matrix notations

$$\begin{aligned} W &= A^{-1} \\ s_k &= \left(w^T * x\right) \end{aligned}$$

Equation 35 Unmixing matrix, W, as the inverted mixing matrix, A

For two random variables (the original signals in this case) to be independent, their joint probability must equal the product of their marginal probabilities. Independent variables are linearly uncorrelated, but the reverse may not be (and often is not) true. Two variables are uncorrelated if their covariance (Equation 36) is equal to zero. If they are independent, any function of the random variables will also have zero covariance.

$$\begin{aligned} &Cov\left(X,Y\right) = E\left[XY\right] - E\left[X\right]E\left[Y\right] \\ &Cov\left(h_1\left(X\right), h_2\left(X\right)\right) = E\left[h_1\left(X\right)h_2\left(X\right)Y\right] - E\left[h_1\left(X\right)\right]E\left[h_2\left(Y\right)\right] \end{aligned}$$

Equation 36 Definition of covariance

Unfortunately, since the original signals are unknown, their marginal and joint distributions are also unknown. Therefore, independence must be estimated. One class of methods minimizes the mutual information of components. Another class maximizes the non-Gaussianity of components, and that class was utilized in this study. The basis for maximizing non-Gaussianity is two-fold. First, ICA is impossible with more than one Gaussian independent component, because uncorrelated Gaussian random variables are also independent, and no unmixing matrix is identifiable. Second, according to Central Limit Theorem sums of independent random variables (such as in the mixing of signals) have Gaussian-like distributions. Therefore, the less Gaussian random variables are, the less likely they are to be formed from sums of other random variables. Greater mathematical detail can be found in "Independent Component Analyis and Applications"[180] and related publications.

One way to maximize non-Gaussianity is to maximize absolute excess kurtosis, since excess kurtosis is only zero for Gaussian distributions, by definition. Older ICA algorithms took this approach. It was later abandoned, however, because kurtosis calculations are strongly biased by outliers. This would seem to be an advantage for outlier detection, and follow-up studies will likely to exploit it. However, the computational speed was important, and the fastest ICA algorithms do not maximize kurtosis.The `R` package I used, `fastICA`[189,190], maximizes negative entropy (**Error! Reference source not found.**), which is also only zero if and only if a distribution is Gaussian. Negentropy is calculated from the information theoretic quantity of differential entropy (i.e., the entropy of a continuous random variable) (Equation 37). Gaussian random variables have the maximum entropy among random variables of equal variance. The negentropy of a random varianble is therefore defined as the negative of the "excess" entropy with respect to a Gaussian random variable of the same variance (**Error! Reference source not found.**).

$$S(p_x) = -\int p_x(u) \log p_x(u) du$$

Equation 37 Differential entropy of a probability distribution, *p_x*

$$J(p_x) = S(\phi_x) - S(p_x)$$

Equation 38 Negative entropy (aka negentropy) of a probability distribution, *p_x*, defined in relation to a Gaussian distribution of the same mean and variance, *ϕ_x*

In practice, however, negentropy is not directly calculated, because doing so would require knowledge of the distributions underlying the independent components. Instead, it is estimated according to Equation 39. The `fastICA` package is capable of using either of the functions in Equation 40, with the first being the default that I used.

$$J(p_x) \approx \left(E\left[G(p_x)\right] - E\left[G(\phi)\right]\right)^2$$

Equation 39 General form of a negentropy estimator

$$G(u) = \frac{1}{\alpha} \log\cosh(\alpha u)$$
$$G(u) = -e^{u^2/2}$$

Equation 40 Functions for use in negentropy estimation

The limitations of ICA should be noted. First, the variances (amplitudes or energies) of the independent components cannot be determined. Because both the mixing matrix and the independent components are unknown a priori, any scalar multiple of a component could be cancelled out in the mixing matrix. Consequently, the variances are typically restricted to equal 1. Second, the signs of the components are unknown, so components that are reflections of each other are indistinguishable. Third, the order of components cannot be known. Again, because both the mixing matrix and independent components are unknown, the any permutation of components can be accommodated by a permutation of the matrix columns. Furthermore, since the variances cannot be known, they cannot be used to establish a unique ordering. Fourth, the number of unique independent components cannot be known a priori. As implied above, the number of components is almost universally assumed to be equal to the number of mixed signals. Obviously, this need not be true, since redundant observations can make the problem over-determined. Lastly, as mentioned earlier, Gaussian components cannot be identified as independent. However, as long as no more than one component is Gaussian the mixing matrix (and therefore the unmixing matrix) is identifiable.

In the unreduced ICA experimental condition, slight dimensionality reduction did occur in some instances. This happened when a diagonalized correlation matrix was computationally singular. I singular matrix has no inverse, having a determinant equal to zero, and is due to a eigenvalue equal to zero in this case. Computational singularity is a result of finite precision and arises when an eigenvalue is vanishingly small or the ratio of the smallest (non-zero) to the largest eigenvalue is too small to store as anything but zero. To avoid this scenario, principal components with eigenvalues less than 0.01 were removed prior to ICA calculation.

5.5.4 Results

For all but the cardiotocography set, outlier MSE and expected cost were less for the ICA-transformed condition. Indeed, as the following section on dimensionality reduction reveals, unreduced ICA produced the least cost of all the conditions. In the case of the cardiotocography data, though, unreduced ICA produced the worst results of all. Further study is required to be certain of the cause. However, one may speculate that the source data are actually Gaussian in distribution, which would violate a fundamental assumption of ICA. Interestingly, the cardiotocography set was the only one actually intended for anomaly detection, having ground truth labels generated using additional clinical data. Clearly, this bears further scrutiny, and future studies will likely explore this issue in greater depth, as well as utilize more data sets with independent anomaly labels (rather than labeled using simplistic statistical criteria).

	Scaled	Unreduced ICA	Reduced PCA	Reduced ICA
Outlier MSE	0.25	0.20	0.24	0.22
Inlier MSE	0.28	0.31	0.26	0.26
E[cost]	0.25	0.20	0.24	0.22

Table 6 Example of benchmark experiment results, showing MSE and expected costs for nonviolent crime at k=10 for each attribute transformation experiment. The complete set of tables can be found in Appendix A.

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 320	k = 640	k = 650
Wine Quality	Outlier MSE	0.15	0.12	0.12	0.14	0.16	0.18	0.19	0.19
35 Outliers	Inlier MSE	0.28	0.28	0.28	0.27	0.27	0.26	0.25	0.25
6462 Inliers	E[cost]	0.15	0.12	0.12	0.14	0.16	0.18	0.19	0.19
ICA	Rand Out MSE	0.43	0.42	0.43	0.42	0.43	0.45	0.46	0.46
	Rand In MSE	0.28	0.28	0.28	0.28	0.27	0.26	0.25	0.26
	Rand cost	0.43	0.42	0.43	0.42	0.43	0.45	0.46	0.46

Table 7 Example of benchmark experiment results, showing MSE and expected costs for wine quality data in the unreduced ICA condition

5.6 SEMANTIC DIMENSIONALITY REDUCTION EXPERIMENT

5.6.1 Introduction

As indicated by the simple dimensionality reduction experiment above, it is typical to transform all attributes at once and apply a single criterion for dimensionality reduction to all of the transformed attributes. In the semantic dimensionality reduction I have developed, attributes are grouped according to shared domain semantic characteristics, PCA is performed for each grouping, and dimensionality is reduced within each group.

This technique is premised on the idea that explicitly reducing or eliminating redundancy would counter aspects of the curse(s) of dimensionality and improve outlier detection. By reducing a feature space to semantically distinct (and hopefully poorly correlated) concepts, the concentration of distance and combinatorial explosion curses of dimensionality might be abated.

On the other hand, eliminating redundancy may be counterproductive, and noise elimination may be a more important concern.

Recent research on intrinsic dimensionality and fractal dimensionality has suggested that far from being harmful, redundancy is beneficial in clustering and nearest-neighbors scenarios[191-205]. The key concepts are "self-similarity" and "ultrametricity". In brief, as dimensionality increases, feature space becomes increasingly hierarchically structured, taking on an ultrametric (fractal) nature. "Ultrametric" means that the data obey a stricter form of the triangle inequality that forces all triangles to be equilateral or narrow isosceles.

If a feature space is unclustered or only weakly clustered, increasing dimensionality collapses all cases to a single point (in the limit). However, if there are distinct clusters, points in the same cluster should be self-similar with each other, but not points from other cluster. Adding redundant dimensions that are highly correlated with each other should cause intra-cluster distances should collapse to zero, but inter-cluster distances should increase. Furthermore, if there are hierarchical clusters present in the data, points should tend to collapse with increasing redundant dimensionality in fashion similar to hierarchical agglomerative clustering.

5.6.2 Hypothesis

Hypothesis 7 Semantic Dimensionality Reduction (Outlier Scoring)

If redundancy is problematic, semantic reduction should improve outlier detection. Conversely, research into ultrametricity has correctly indicated the benefits of redundancy, semantic reduction should harm outlier detection. Without ground truth labels, this technique is difficult to evaluate, though, since no comparison can be made between known and learned scores. However, some indication of performance can be obtained from rule extraction.

Hypothesis 8 Semantic Dimensionality Reduction (Rule extraction)

If semantic reduction facilitated the identification of robust outlier score patterns and trends, this would be reflected in better rule learning performance (with respect to other experimental conditions, especially no reduction).

5.6.3 Methods

Validation score attributes were grouped into semantically related groups. The groups[f] and the key to attribute name meanings[g] can found in supplementary material.

PCA was performed in each group, and dimensionality was reduced using the Broken Stick Model and Guttman-Kaiser Criterion (as specified in section 5.5.3). Once dimensionalities were reduced, the remaining components from each group were merged, another round of PCA was performed (without reduction), and finally ICA was performed on the PCA-whitened data.

5.6.4 Results

This experiment is included in the section on outlier scoring because it dealt with the transformation of attributes and reduction of dimensionality, which in turn affected the pairwise

[f] http://d-scholarship.pitt.edu/19601/11/SuppG_PDB_attribute_semantic_groupings.txt

[g] http://d-scholarship.pitt.edu/19601/9/SuppE_complete_PDB_attribute_list.pdf

distance matrix and therefore outlier scores. However, because the efficacy of this technique could only be assessed according to its effect on rule learning accuracy, the results are reported in section 6.4.4.

5.7 CONTEXTUALIZATION

5.7.1 Introduction

Recent studies have offered some suggestions for improving outlier detection by scoring cases with respect to different contexts/conditions[61,206,207]. Whether conditional outlier scoring is more sensitive and/or specific than scoring on an undifferentiated data set is a testable hypothesis, but in this domain, it is a necessary conjecture. There are multiple contexts to consider in evaluating these data. In addition to whether a model is an outlier with respect to the whole PDB, there is value in knowing if it is an outlier with respect to year of deposition or resolution. Those are universally available proxies for structure modeling paradigm and experimental data quality, respectively. To account for these contexts, I have subdivided the data into multiple subsets for each variable and sought outliers in each.

Differences in consensus best practices and technologies used for model building and refinement in different eras are known to produce differences between validation score distributions in different eras[14,17,72]. It is reasonable to conjecture that models that were not built and refined according to consensus best practices during a given era will tend to receive higher outliers than those that were.

Resolution gives an approximate measure of the information content of a diffraction pattern. Models built and refined from data with similar resolutions are expected to be of similar quality, and those that are not are anomalous.

Because trends and clusters do not obey strict bin boundaries, bin boundaries were chosen such that neighboring bins overlap slightly. In the ranges reported below, parentheses represent open interval endpoints and brackets indicate closed endpoints.

5.7.2 Contextual Attribute Preparation

Two attributes were chosen to generate contexts for outlier detection: optical resolution and year of deposition. They were chosen to be broadly discriminative for differing diffraction data quality and tools/techniques of model refinement. For these attributes, continuous values were discretized into equal-width bins (except for end bins), and each bin was treated as a separate context. Discretized bins overlap both of their neighbors in a “shingled” fashion[12].

5.7.3 Year of Deposition

Years of deposition were grouped into six ranges: [1972, 1992], [1992, 2000], [2005, 2005], [2005, 2010], and [2010, 2011].

The span of 1972 to 1992 represent the early, formative years of protein crystallography and structure determination. Two major changes to methodology occurred at the end of this era. First, covalent geometry constraints and restraints were standardized according to the distributions reported in 1991 by Engh and Huber[54]. Prior to that, each refinement program had its own set of parameters. Second, a cross-validating residual score, the free R-factor, was introduced in 1992[55]. Such as score was found to be necessary when the conventional R-factor was found to be frequently over-fit[8].

Unfortunately, it was necessary to exclude this context from the final set of models with probabilistic scores. This is because the raw score distribution was strongly bimodal, which made any distribution fits unreliable.

For the entirety of the period from 1992 to 1995, the frequency of R_{free} report remained below 10%. However, models steadily improved, thanks to the Engh and Huber parameters for ideal covalent geometry. The frequency of R_{free} use in 1996 was nearly 1 in 3 depositions. The sudden and steep change was likely provoked by papers highlighting some serious errors in published models[20,21].

The raw data for this context lacked clear unimodality and had only weak resemblance to the Amoroso family of distributions, but we conjectured that the fit was “good enough”. We speculate that in first two contexts, the main problem is the presence of multiple disparate subsets with low cardinality and high dispersion. That is, the slow rate of deposition in the early years of the PDB, coupled with changes to tools and techniques for model building, refinement, and validation, led to weakly clustered data and blurry boundaries between inliers and outliers. Future studies may combine the first two contexts in an attempt to reduce sparseness and increase the likelihood of finding a close parent distribution for score standardization.

The period of 1995 to 2000 was transitional in nature. The frequency of structure factor deposition in 1995 was around 1 in 3. By 2000, the frequency was better than 1 in 2. The frequency of R_{free} use in 2000 was over 85%.

In 2000, the Protein Structure Initiative (PSI), an effort to populate the space of protein folds by using high-throughput structure determination pipelines, began[208-211]. In 2005, the first phase of the PSI ended, and efforts began to reassess to the goals and effectiveness of the program[212-216].

The second phase of the PSI ended in 2010. This phase was mainly a continuation of successful methods devised during the first. In 2005, the frequency of structure factor deposition was nearly 85%.

Phase three of the PSI began in 2010[217]. The focus of phase three is biological relevance, rather than high throughput. This may significantly affect the types of protein structures determined and their relative complexities. My collected data only extend to the end of 2011. This research began in 2012, so I decided to make the cutoff at the end of the last full year.

5.7.4 Optical Resolution

Optical resolutions (as computed by SFCheck) were grouped into six ranges: (0, 1.1], [1.1, 1.5], [1.4, 1.7], [1.6, 1.9], [1.9, 2.4], [2.3, max]. These ranges are based on a mix of population statistics and domain knowledge. They approximately coincide with a regression-based conversion from nominal to optical resolution. The corresponding nominal resolution ranges are (0.0, 1.2], [1.2, 1.8], [1.8, 2.1], [2.1, 2.5], [2.5, 3.3], and [3.3, 4.0].

A nominal resolution of 1.2Å is approximately the boundary of atomic resolution, at which point atoms begin to appear as discrete "balls" of high density in electron density maps. The range of 1.2Å to 1.8Å is considered "high" resolution because information content of diffraction data at those resolutions allows a high degree of over-determination (i.e., the available number data points exceeding the number of adjustable parameters). Resolutions in the range 1.8 Å to 2.1Å are considered "moderate" to "high" and lie on the higher side of the global mean. The range 2.1Å to 2.5Å corresponds to "moderate" to "low" resolutions, and lies on the lower side of the global mean. Resolutions in the range 2.5Å to 3.3Å are "low" to "very low". The last interval, from 3.3Å to 4.0Å covers "very low" resolution. Models with resolutions in this range are very likely to contain accuracy errors. Indeed, WhatCheck regards resolutions beyond 3.5Å as sufficiently problematic that it will not even produce validation reports for them.

5.7.5 Merging Contextual Scores

Because contextual subset bins are bounded as overlapping "shingles", some of the cases earned an outlier score from two subsets of the same variable, whereas only one is desired. To resolve this conflict, the lower score of the pair is retained. The conjecture driving that decision is that the subset in which a case receives a lower outlier score is the one a case is more "at home" in. In other words, such a subset provides a more appropriate neighborhood of cases for comparison.

Once each case had one score from each of the two contextual attributes, plus one from the improper subset, I needed to combine them into a single score for each case. I had originally desired to use a component failure model, in which the join probability of each case being an outlier would be estimated[218]. This idea was inspired by the idea that each contextual outlier score represented a component, connected to the other components in series. That way, if any one component failed (had a high outlier score), they whole system would fail.

$$pLOF_{PDB} = 1 - \left(1 - pLOF_{res}\right)\left(1 - pLOF_{year}\right)\left(1 - pLOF_{improper}\right)$$

Figure 13 Failed idea for an overall outlier score as components connected in series

This idea did not succeed, because the resulting scores were tightly compacted near a probability of 1.0. I suspect that this is due to the component scores being too highly correlated, but further research is needed to be certain.

As an alternative, I simply took the arithmetic average of all three scores. This was a satisficing[137] solution, because the plots of “gold standard” attributes, such as MolProbity Score, had an acceptable appearance (Figure 14, Figure 15). By “acceptable”, I mean that extreme validation score values are associated with extremely high outlier score values, and there are not too many cases with extreme validation scores and low outlier scores.

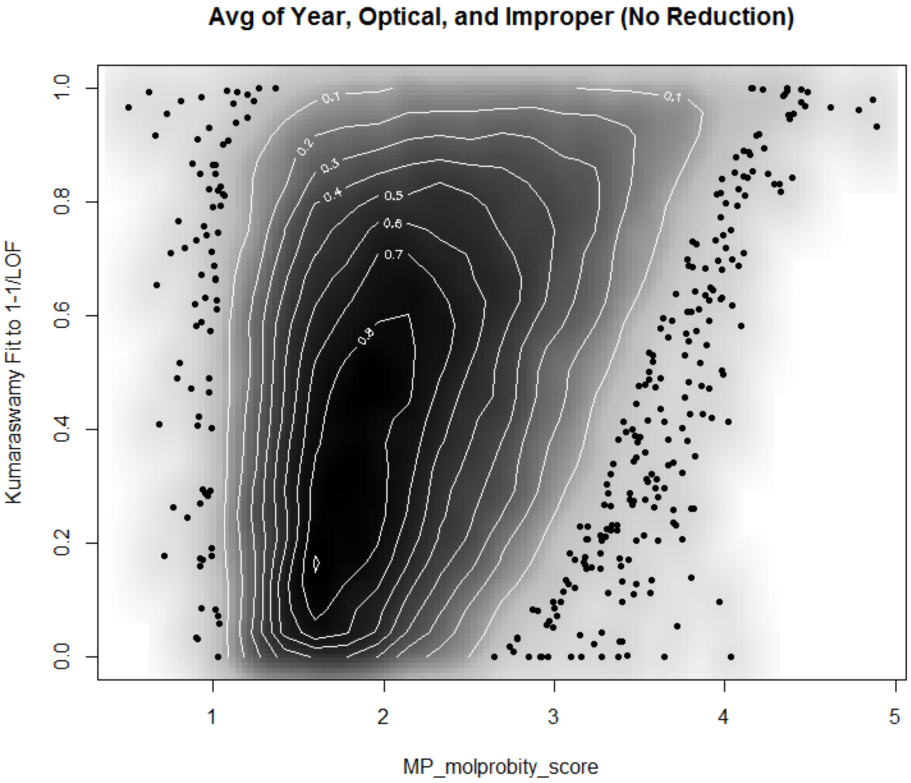


Figure 14 Plot of MolProbity Score and final *pLOF* score on the x and y axes, respectively

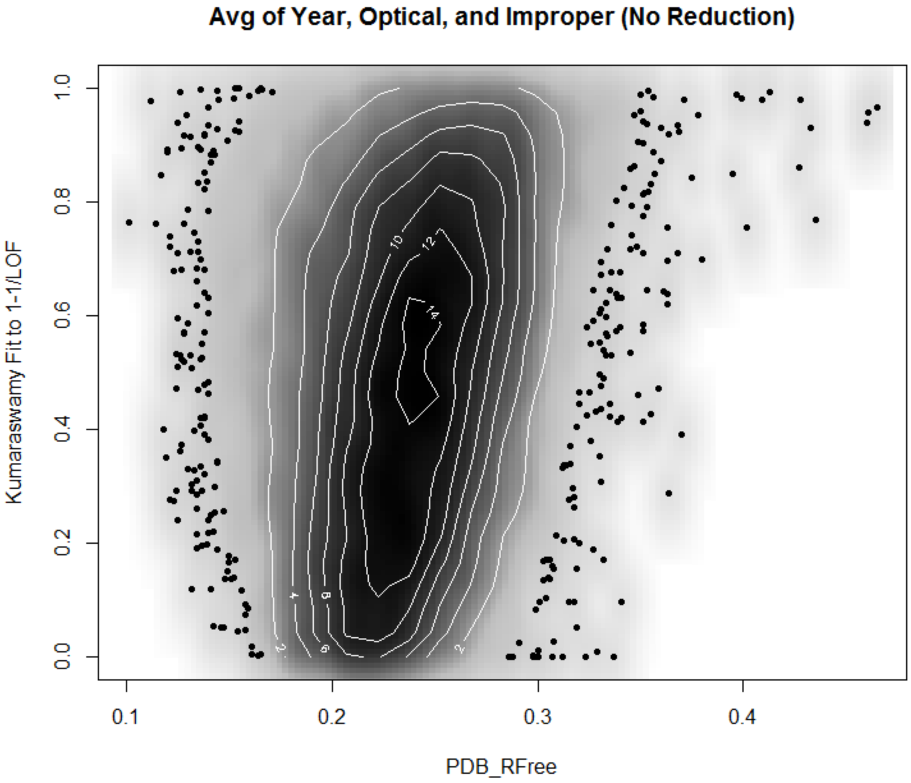


Figure 15 Plot of reported Rfree and final *pLOF* score on the x and y axes, respectively

These plots serve an important function in detecting outlier among PDB data. Without "gold standard" or ground truth labels to indicate known inliers and outliers

5.8 MAXIMAL OUTLIER ANALYSIS

5.8.1 Introduction

A basic assumption of anomaly detection is that inliers and outliers are generated by different population distributions. Outlier detection may be expressed as finding a separating boundary between those distributions for classification purposes. It is not an aim of this research to classify entries definitively as outliers or inliers, but to score entries with estimates of outlier probabilities. However, such a boundary could still be useful for identifying particularly interesting outliers for deeper analysis.

5.8.2 Conjecture

Examining the most extreme outliers might reveal interesting commonalities among them. Specifically, it is reasonable to expect that some extreme outliers represent models of exceptionally high quality, built from data near current technological limits, and some represent exceptionally poor quality, with evidence of physical absurdities. Others may be indicative of validation tool or batch processing glitches, or rare structural features of structural anomalies (see section 2.4).

5.8.3 Methods

All of the scores described here were generated in the unreduced ICA experimental condition.

In the process of fitting a distribution family to LOF_S and $1-1/LOF_S$ scores, inliers were fit very tightly, but a departure was observed for entries deep in the outlier score tail (see section 5.2). A visual inspection of the Q-Q plot for a Kumaraswamy distribution fit to scores in the improper subset indicated that the departure begins at a $pLOF_{PDB}$ of approximately 0.997, and encompasses 98 entries – a rather large set to be evaluated manually.

Another threshold was found by examining entries associated with a known fraud: 1Y8E and 2A01 (see section2.5.4). These entries received scores of 0.99994 and 0.99995, respectively. Setting the threshold at 0.9999 reduced the number of entries to 26 – a manageable number for manual evaluation.

Once a threshold was set, the raw validation scores of entries meeting or exceeding it were tabulated and analyzed. Focus was on values that were well outside of the typical range for ordinary structure models – good, bad, or just peculiar.

5.8.4 Results

The following table reports a series of threshold values and the number of entries have outlier scores greater than or equal to them. The complete list of PDB IDs for models with scores greater than or equal to 0.99 can be found in the supplementary material[h].

Threshold	0.9900	0.9910	0.9920	0.9930	0.9940	0.9950	0.9960	0.9970	0.9980	0.9990
Count	193	181	171	159	147	129	117	98	85	60
Threshold	0.9991	0.9992	0.9993	0.9994	0.9995	0.9996	0.9997	0.9998	0.9999	
Count	58	55	53	53	47	44	34	30	26	

Table 8 Counts of entries with outlier scores meeting or exceeding a series of threshold values

[h] http://d-scholarship.pitt.edu/19601/13/SuppI_extreme_PDB_outlier_ids.txt

The 26 entries examined are identified by PDB IDs 15c8, 1a8v, 1j78, 1sbm, 1tzb, 1w48, 1y8e, 2a01, 2akf, 2f80, 2q23, 2qgh, 2qz5, 2uwf, 2xi5, 2y8v, 2z2q, 2zhi, 3ai3, 3cji, 3e3z, 3fk0, 3iqv, 3k5d, 3kh2, and 3iqv.

Analysis of the 26 most extreme ("maximal") outliers indicated several types of outliers, but an unambiguous diagnosis is not possible for many of them. Further study is clearly required.

In the following subsections, the maximal anomalies are triaged based on diagnostic characteristics. Note that descriptive assessments like "good" and "poor" refer to the connotations of particular validation score values, not to whole models. Most validation scores have clear directionality, with "very good" and "very poor" extremes (see section 2.3).

As incomplete and preliminary as these analytical summaries these are, producing them was a very time-consuming process. If these maximally outlying 26 models had been analyzed non-stop, without interruption, it probably would have taken at least a week to complete. There are roughly 200 models in the extreme (greater than or equal to 0.99) category – eight times as many the maximals. A complete analysis of those would take two to four months. Automation, such as clustering, could be used to perform initial triage, but statements of causality require great caution, since reputations are on the line.

The methods I have chosen and employed in this study have fruitfully directed attention to the models that are most certainly anomalous from a validation standpoint. These methods form an effective screen, but screens are not diagnostic. Future work might include methods for automatically framing hypotheses related to diagnosis.

5.8.4.1 Documented Fraud

Models 1y8e and 2a01 are among the fraudulent Murthy structures described in section 2.5.4. It should be noted that no explicit effort was made to tune my methods to detect fraudulent models. Indeed, no case-specific tuning whatsoever was performed, and all detected phenomena are "emergent" and based solely on the distances between data points. Nonetheless, it is very reassuring, and frankly exciting, to see models known to be marked with physically absurd features highlighted by an exceptionally high outlier score.

1y8e[i] has the following characteristics.

- Very poor stereochemistry, especially for the given resolution
- Relatively high Matthews coefficient
- Relatively high percentage of volume solvent
- Very high number of steric clashes and bumps
- Very high percentage of covalent angle outliers
- Very high main chain h-bond energy standard deviation
- Negative difference between maximal and minimal estimated coordinate error

[i] http://eds.bmc.uu.se/cgi-bin/eds/uusfs?pdbCode=1y8e

2a01[j] has the following characteristics.

- Low 2σ correlation coefficient
- SFCheck recomputed R-factors are higher than reported R-factors
- Very poor stereochemistry, especially for the given resolution
- Relatively high Matthews coefficient
- Relatively high percentage volume solvent
- Very high number of steric clashes and bumps
- High percentage of covalent angle outliers
- Very negative difference between maximal and minimal estimate coordinate error
- Very high noise-to-signal ratios for measured

5.8.4.2 Viruses With Misreported NCS

Three of the maximal outliers, 2z2q, 2q23, and 3cji are virus structures produced by the same lab. They share several anomalous validation scores, which my advisor and I believe to be produced by incorrectly reported non-crystallographic symmetry (NCS). This is a testable hypothesis that I may pursue after graduation by attempting to generate the correct NCS operators and add them to the deposited model files. I would then re-run the validation tools to see if validation scores improved.

[j] http://eds.bmc.uu.se/cgi-bin/eds/uusfs?pdbCode=2a01

Interestingly, 2z2q and 2q23 were obsoleted in 2012 and replaced with 4ftb and 4fts, respectively. I do not have validation score profiles or outlier scores for those models, because by data are only complete through 2011.

The validation profiles of each are as follows.

- 2z2q
    - Very high all-data and 2σ R-factors
    - Very low R-factor correlations
    - very high recomputed R-factors
    - High $\langle u_{atomic} \rangle$, high $\sigma(u_{atomic})$
    - Large differences for B-factors of neighboring atoms
    - Negative Patterson $B_{obs} - B_{calc}$
    - High Errat score
    - OK stereochemistry
    - VERY high Matthews coefficient
    - Many (mild) bumps
    - High Luzzati error
    - High variance of coordinate displacement from electron density peaks
- 2q23
    - Very high all-data and 2σ R-factors
    - Very low R-factor correlations
    - Very high recomputed R-factors
    - Very high coefficient of variation for water ADPs

- Poor ProSA pairwise energy score
- Very high Errat score
- So-so stereochemistry
- VERY high Matthews coefficient
- 98% solvent content
- Very many (mostly mild) bumps
- Very high σ(h-bond energy)
- High Luzzati error
- High variance of coordinate displacement from electron density peaks

- 3cji
  - Very high all-data and 2σ R-factors
  - Very low R-factor correlations
  - Very high recomputed R-factors
  - Poor ProSA pairwise energy score
  - High Errat score
  - High TAP score
  - So-so stereo
  - VERY high Matthews coefficient
  - High Luzzati error
  - 74% complete diffraction
  - Relatively high maximal estimated coordinate error
  - Relatively high mean and variance of atomic electron densities

5.8.4.3 Possibly Problematic

The following models have one or more score values in a physically absurd or exceptionally rare range. For instance, maximal estimated coordinate error should always be greater than minimal estimate coordinate error. If it is not, this means that somehow observed and calculated structure factors are in agreement to greater precision than the measurement error associated with the measured reflections.

- 15c8
  - good R-factors
  - High $\sigma\left(u_{atomic}\right)$
  - Large differences for B-factors of neighboring atoms
  - High Errat score
  - Very poor stereo
  - Negative difference between maximal and minimal estimated coordinate error
- 1j78
  - Good R-factors
  - High $\left\langle u_{atomic}\right\rangle$, somewhat high $\sigma\left(u_{atomic}\right)$
  - Main chain B-factor variance greater than side chain variance
  - Large differences for B-factors of neighboring atoms
  - Good stereochemistry
  - Many (mild) bumps
  - Highly negative difference between maximal and minimal estimated coordinate error

  - <σ(F)>/<F> greater than 1 (more reflection noise than signal)

- 2qgh
  - High 2σ R-factor
  - Very high recomputed R-factors
  - Good stereochemistry
  - Negative difference between maximal and minimal estimated coordinate error
  - <σ(I)>/<I> greater than 1, <σ(F)>/<F> greater than 1
- 3ai3
  - Somewhat poor R-factors for given resolution
  - Much lower than expected B-factors
  - High percentage buried B-factors < 5Å^2
  - Very good Verify3D score
  - Good overall stereochemistry
  - Very many bumps (some as bad as Murthy's frauds)
  - WhatCheck reports possible pseudo-symmetry
- 3n7x
  - Very high all-data and 2σ R-factors
  - Very low R-factor correlations
  - very high recomputed R-factors
  - Strongly negative Patterson $B_{obs} - B_{calc}$
  - Very poor ProSA pairwise energy score
  - Poor Errat score
  - Relatively good Ramachandran plot

- Surprisingly excellent rotamers
- Large (74%) volume not model
- SFCheck Matthews coefficient is enormous; disagrees with WhatCheck
- 98% solvent content
- High Luzzati error
- High variance of coordinate displacement from electron density peaks
- 63% completeness
- Relatively strong anisotropy
- EDS and PDB REDO could not reproduce R-factors
- WhatCheck reports atoms at special positions with too high occupancy and atoms too close to a symmetry axis

5.8.4.4 Obvious Glitch

One maximal outlier is clearly the result of a data processing error. For 2xi5, my version of `WhatCheck` reports a physically absurd Matthews coefficient, but it is not corroborated by `SFCheck` or PDB REDO's server version of `WhatCheck`.

5.8.4.5 Strong Anisotropy

Some of the maximal outliers appear to exhibit strong diffraction anisotropy. Diffraction anisotropy produces different degrees of quality along the three principal axes, which in turn produce uneven electron density quality. Electron density maps of diminished quality are harder to accurately and precisely thread amino acid residue chains through, which can lead to poorer models than expected for the reported resolution (which may be the resolution of the best dimension).

- 2uwf
  - Very high recomputed R-factors
  - Very good ProSA pair energy score
  - Good stereochemistry
  - HUGE maximal estimated error
  - High DPI
  - 10% completeness?!
  - Extreme anisotropy
- 1w48
  - Very good R-factors
  - Somewhat high $\sigma\left(u_{atomic}\right)$
  - Large differences for B-factors of neighboring atoms
  - Good stereochemistry
  - VERY high bond ratio
  - Large number of planar group outliers
  - Very many (mostly mild) bumps
  - Strong anisotropy
- 2qz5
  - Very high recomputed Rfree
  - High coefficient of variation for water ADPs

  - Much lower than expected B-factors
  - Negative Patterson $B_{obs} - B_{calc}$
  - High ProSA pairwise energy score
  - Somewhat high Errat score
  - High TAP score
  - Somewhat high clash score
  - High σ(h-bond energy)
  - Large number of unacceptable (unusable) reflections
  - Strongly anisotropic

- 3e3z
  - High all-data and 2σ R-factors
  - High recomputed R-factors
  - High variance for water ADPs
  - Much lower than expected B-factors
  - High percentage of buried atoms with B-factors < $5Å^2$
  - Very strongly negative Patterson $B_{obs} - B_{calc}$
  - 97.3% volume not model
  - High maximal estimated error
  - VERY large atomic electron densities coefficient of variation
  - VERY low real space correlation
  - strong anisotropy

5.8.4.6 Unclear or Ambiguous

For the remainder of maximal outliers, a clear potential diagnosis was not possible without considerable manual analyses. However, we speculate that the problem is related to very high resolutions paired with disproportionately poor validation scores. Generally speaking, we could identify *what* made a model anomalous, but necessarily *why* one or more scores were anomalous (i.e., their underlying causes).

- 3kh2
    - Good R-factors
    - VERY high $\sigma\left(u_{atomic}\right)$
    - VERY high coefficient of variation for waters ADPs
    - Very good ProSA pairwise energy score
    - Decent stereochemistry
    - Large number of unacceptable reflections
    - mild anisotropy
- 1tzb
    - Atomic resolution
    - Excellent R-factors
    - Excellent stereochemistry
    - Only 19% volume not model (many explicit waters?)
    - Very high mean and variance of atomic electron densities
    - VERY large intensity scale factor

- 2akf
  - Atomic resolution
  - Good R-factors
  - Somewhat large B-factors that expected
  - Strongly negative Patterson $B_{obs} - B_{calc}$
  - High ProSA surface energy score
  - Superlative stereochemistry
  - Very high mean and variance of atomic electron densities
  - VERY large intensity scale factor
  - somewhat anisotropic
- 1sbm
  - Near-atomic resolution
  - Very high recomputed R-factors
  - Low diffraction correlation
  - Excellent stereochemistry
  - Somewhat anisotropic?
- 2f80
  - Near-atomic resolution
  - Very high recomputed Rfree
  - Large differences for B-factors of neighboring atoms
  - Very good stereochemistry
  - Very good Verify3D
  - Very good backbone z-score

- 2y8v
    - Good R-factors
    - Good stereochemistry
    - WhatCheck and SFCheck slightly disagree on Matthews coefficient
    - High C_{beta} deviation count
    - High maximum deviation
    - ENORMOUS bond ratio
    - Very many (mostly mild) bumps,
    - mild anisotropy
- 2zhi
    - Near-atomic resolution
    - Very high recomputed Rfree
    - Large differences for B-factors of neighboring atoms
    - Very good ProSA pairwise energy score
    - Poor MolProbity score for the given resolution
    - OK stereochemistry
    - Only 17% volume not model (many explicit waters?)
    - High maximum deviation
- 3fk0
    - Near-atomic resolution
    - Excellent R-factors
    - Very good ProSA pair energy score
    - Very good Verify3D score

  - Superlative stereochemistry
  - Large number of unacceptable reflections
  - Mild anisotropy
- 3iqv
  - Atomic resolution
  - Very high all-data and 2σ R-factors
  - Very high recomputed R-factors
  - Very low R-factor correlation
  - High coefficient of variation for B-factors
  - Strongly negative Patterson $B_{obs} - B_{calc}$
  - Very good ProSA pairwise energy score
  - High ratio of waters to amino atoms
  - Superlative stereochemistry
  - Very low (good) σ(h-bond energy)
  - ENORMOUS variance of coordinate displacement from electron density peaks
  - Low real space correlation
- 3k5d
  - Relatively high recomputed R-factors
  - High mean B-factor
  - Large differences for B-factors of neighboring atoms
  - Somewhat strongly negative Patterson $B_{obs} - B_{calc}$
  - Poor ProSA pairwise energy score
  - High Errat score

- Decent stereochemistry
- Very many (mostly mild) bumps
- ENORMOUS DPI
- Somewhat high mean and variance for atomic electron densities
- Mildly anisotropic

5.8.4.7 Exceptionally High Quality

Interestingly, one expected class of outliers was not observed among the set of maximal outliers, namely those of extremely high quality. This may be for multiple reasons. Validation score values for very good models may tend to fall along inlying trends, rather than in regions of locally low (i.e., outlying) density. Alternately, they may appear in proximate groups large enough to be regarded as inlying clusters according to the neighborhood size chosen. It is also possible that bad outliers are more greatly separated from the nearest inliers than good outliers are, making the density concentrations much stronger for the former. Additionally, validation scores were designed to highlight bad outliers, not characterize "well-behaved" models. This bias shows up in compression on the "good" end of score value spectra.

5.9 DEEP INLIER ANALYSIS

5.9.1 Introduction

Since this was a study of outliers, I had not set out to examine deep inliers. However, doing so came about as a logical consequence and "emergent" property of the outlier scoring algorithm used. Indeed, it was fortuitous to do so, as it provided supporting evidence for one of my central hypotheses.

As specified in section 5.1, LOF_S is calculated as a ratio. The denominator is the local density in which a given point is embedded, and the numerator is the average density of that point's neighbors. These quantities are only comparable when a point's density does not differ much from its neighbors.

A key principle of LOF's design is that points in clusters (inliers) are expected to have similar local densities, and therefore their LOF_S scores should be close to one. Points in regions of locally low density have LOF_S scores greater than 1.0, and outliers can have LOF_S scores many times greater than 1.0. However, it is also possible for a point to have an LOF_S score less than 1.0. With a little reflection, one can see that these scores represent peaks in the density of a feature space.

It is possible that peaks in the density of a feature space are hubs, a phenomenon described in recent anomaly detection literature related to the curse(s) of dimensionality[219-222]. Hubs are cases that appear in many k-neighborhoods of points in a data set. That is, they are "popular" neighbors. They arise because points in high-density regions of feature space typically do not have low-density points in their k-neighborhoods, but points in low-density regions must include high-density points in order to accumulate k neighbors. Consequently, high-density points appear in more neighborhoods than low-density points. Conversely, outliers are anti-hubs. Some work has been done to exploit hubs as cluster prototypes[222].

5.9.2 Hypothesis

I hypothesize that deep inliers (i.e., feature space density peaks), defined as LOF scores less than one, correspond to a type of hub. These hubs are prototypical cases in a data set, exemplars of the strongest modes, around which clusters form. In the context of this domain, I specifically posited that deep inliers are indicative of standardized model building, refinement, and validation practices. This was tested by analyzing score distributions for "gold standard" validation scores. My judgment, aided by an expert structural biologist (namely my advisor), determined whether deep inlier score distributions are more representative of state-of-the-art model building, refinement, and validation standards.

5.9.3 Methods

After LOF_{PDB} scores have been standardized as specified in section 5.2, LOF_{PDB} scores less than 1.0 become $pLOF_{PDB}$ (i.e., probabilistic LOF) scores equal to zero. Entries that have scores of zero in all three contexts are potential hubs with respect to the improper subset, their relevant deposition eras, and their relevant optical resolution ranges. The PDB IDs of all cases with $pLOF_{PDB}$ scores equal to zero were extracted. The minimum, maximum, and quartile values of select validation scores were computed for those entries and compared to the same for all scored entries.

5.9.4 Results

The complete list of 184 PDB entries extracted as possible hubs can be found in the supplementary material[k].

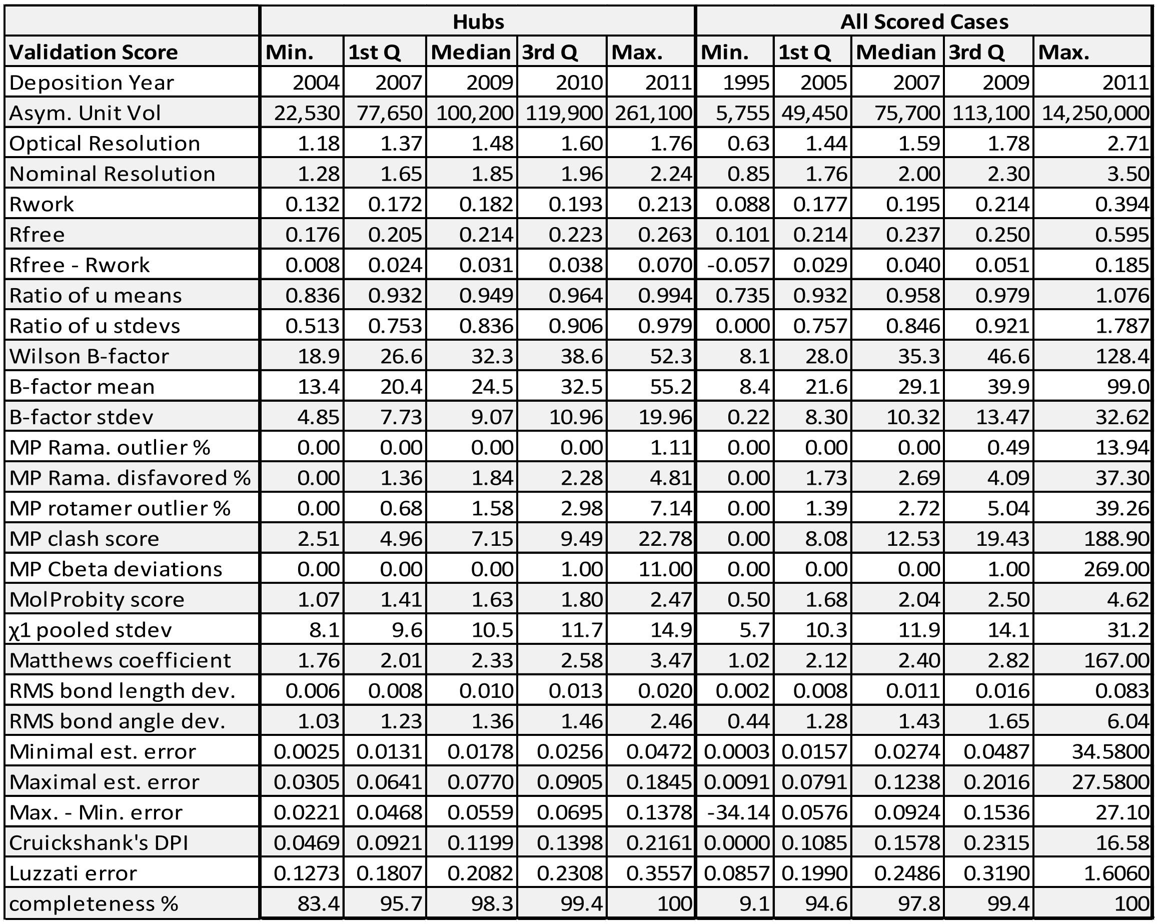
The following table reports the minimum, maximum, and quartile values of select validation scores for the potential hubs and the same for all scored entries. The reported validation scores were chosen primarily from those regarded as "gold standard" by structural biologists. Others were chosen based on the results of other experiments. Score values summaries (for a selection of attributes) pertaining to all models, hubs, and maximal outliers can be found in the supplementary material[l].

The potential hubs have no entries deposited before 2004, but otherwise the distribution of deposition year does not differ drastically from that of the general population. The distribution of asymmetric unit volumes for hubs is narrower than the general population. The attributes directly related to model quality all exhibit trends toward slightly better values and narrower distributions for hubs. These distribution shifts are consistent with the conjecture that potential hubs, defined as entries with LOF_{PDB} less than 1.0, represent entries deposited in the last decade, with average protein size, based on slightly better-than-average diffraction data, and modeled well according to standards established by the domain (both formally and informally). Additional study would be needed to determine whether or not these potential hubs are all prototypes for separate and

[k] http://d-scholarship.pitt.edu/19601/12/SuppH_deep_PDB_inlier_ids.txt

[l] http://d-scholarship.pitt.edu/19601/15/SuppK_select_PDB_attribute_value_summaries.txt

distinct clusters.



	Hubs					All Scored Cases				
Validation Score	**Min.**	**1st Q**	**Median**	**3rd Q**	**Max.**	**Min.**	**1st Q**	**Median**	**3rd Q**	**Max.**
Deposition Year	2004	2007	2009	2010	2011	1995	2005	2007	2009	2011
Asym. Unit Vol	22,530	77,650	100,200	119,900	261,100	5,755	49,450	75,700	113,100	14,250,000
Optical Resolution	1.18	1.37	1.48	1.60	1.76	0.63	1.44	1.59	1.78	2.71
Nominal Resolution	1.28	1.65	1.85	1.96	2.24	0.85	1.76	2.00	2.30	3.50
Rwork	0.132	0.172	0.182	0.193	0.213	0.088	0.177	0.195	0.214	0.394
Rfree	0.176	0.205	0.214	0.223	0.263	0.101	0.214	0.237	0.250	0.595
Rfree - Rwork	0.008	0.024	0.031	0.038	0.070	-0.057	0.029	0.040	0.051	0.185
Ratio of u means	0.836	0.932	0.949	0.964	0.994	0.735	0.932	0.958	0.979	1.076
Ratio of u stdevs	0.513	0.753	0.836	0.906	0.979	0.000	0.757	0.846	0.921	1.787
Wilson B-factor	18.9	26.6	32.3	38.6	52.3	8.1	28.0	35.3	46.6	128.4
B-factor mean	13.4	20.4	24.5	32.5	55.2	8.4	21.6	29.1	39.9	99.0
B-factor stdev	4.85	7.73	9.07	10.96	19.96	0.22	8.30	10.32	13.47	32.62
MP Rama. outlier %	0.00	0.00	0.00	0.00	1.11	0.00	0.00	0.00	0.49	13.94
MP Rama. disfavored %	0.00	1.36	1.84	2.28	4.81	0.00	1.73	2.69	4.09	37.30
MP rotamer outlier %	0.00	0.68	1.58	2.98	7.14	0.00	1.39	2.72	5.04	39.26
MP clash score	2.51	4.96	7.15	9.49	22.78	0.00	8.08	12.53	19.43	188.90
MP Cbeta deviations	0.00	0.00	0.00	1.00	11.00	0.00	0.00	0.00	1.00	269.00
MolProbity score	1.07	1.41	1.63	1.80	2.47	0.50	1.68	2.04	2.50	4.62
χ1 pooled stdev	8.1	9.6	10.5	11.7	14.9	5.7	10.3	11.9	14.1	31.2
Matthews coefficient	1.76	2.01	2.33	2.58	3.47	1.02	2.12	2.40	2.82	167.00
RMS bond length dev.	0.006	0.008	0.010	0.013	0.020	0.002	0.008	0.011	0.016	0.083
RMS bond angle dev.	1.03	1.23	1.36	1.46	2.46	0.44	1.28	1.43	1.65	6.04
Minimal est. error	0.0025	0.0131	0.0178	0.0256	0.0472	0.0003	0.0157	0.0274	0.0487	34.5800
Maximal est. error	0.0305	0.0641	0.0770	0.0905	0.1845	0.0091	0.0791	0.1238	0.2016	27.5800
Max. - Min. error	0.0221	0.0468	0.0559	0.0695	0.1378	-34.14	0.0576	0.0924	0.1536	27.10
Cruickshank's DPI	0.0469	0.0921	0.1199	0.1398	0.2161	0.0000	0.1085	0.1578	0.2315	16.58
Luzzati error	0.1273	0.1807	0.2082	0.2308	0.3557	0.0857	0.1990	0.2486	0.3190	1.6060
completeness %	83.4	95.7	98.3	99.4	100	9.1	94.6	97.8	99.4	100

Table 9 Comparison of score distributions for potential hubs and all scored entries.

With respect to the score distributions, it is gratifying to see evidence of the hypothesized consensus commonalities (section 3.0). This evidence is crucial for establishing the credibility of these methods in structural biology and bioinformatics. Future work will include examination of deep inliers in each contextual subset. We hypothesize that observed validation score distributions will be indicative of “case-controlled”[14] patterns associated with “state of the art” structure determination methods.

6.0 RULE EXTRACTION

6.1 INTRODUCTION

The anomaly detection methods I have employed have two nontrivial weaknesses: opacity and the fate of missing values. Opacity refers to the difficulty in mentally mapping from values in the input feature space to outlier scores. Indeed, because nearest-neighbors methods are "lazy" learners that rely on emergent patterns in data, there is no explicit or general model to consult. Thus, it is often quite difficult to know *why* a case gets a particular outlier score. Missing values are problematic for my methods because ICA requires complete cases for its linear transformations. Consequently, if even just one value is missing, a case cannot receive an outlier score. This resulted in the loss of roughly 50% of the PDB.

I resolved both of these problems by performing rule extraction. Rule extraction uses an association or inductive rule learner to convert opaque relationships between numerous continuous-valued attributes into comprehensible IF-THEN propositions involving discrete ranges of only a handful of attributes at a time. Rules may also be described as hypercubes with axis-parallel surfaces in feature space[223].

Rule extraction has been fruitfully used to enhance the intelligibility of results produced by artificial neural networks[224,225], support vector machines[226,227], and other numerical "black box" machine learning methods[228]. Rules are appealing because they encode knowledge in an intuitively comprehensible form that can be easily verified by domain experts and used to frame new testable hypotheses[224,229,230].

There are two general modes in which rule extraction is performed: decomposition and learning. Decomposition seeks to derive propositional rules from the internal components of a black box tool. For instance, in the case of artificial neural networks, the components are input, hidden, and output nodes, and rules represent the incoming signal strengths that trigger activation of a node. In learning mode, the tool remains a black box. Instead of describing the internal functions, learning mode uses the values predicted by the tool, and the input values associated with them, as training data for an inductive rule learner.

There are no internal nodes amenable to propositional forms in LOF, because it is built on the instance-based learning of k-nearest-neighbors. Hence, I have applied learning mode to my data. I am only aware of one attempt to learn rules from outlier scores generated by an unsupervised detection method (SmartSifter)[231]. The goal of that work was different, with the rule learner used primarily to filter cases for multiple rounds of refined scoring.

6.2 HYPOTHESIS

Hypothesis 9 Rule Extraction

Outlier scoring is sufficiently opaque, having no explanatory power or value, that it would engender healthy skepticism in any domain, including structural biology. Therefore, I hypothesized that disjunctive sets of conjunctive (i.e., disjunctive normal form) IF-THEN rules could be used to post-process outlier scores, associating them with values from small groups of validators, and thereby making them comprehensible and explainable. Support for this hypothesis was established in three ways. The first was to demonstrate better-than-random accuracy for rules predicting outlier scores from benchmark data. The second was to do the same for protein structure model validation and outlier scores. The third was to subjectively assess whether rules generated from validation scores and PDB models were consistent with domain theory.

6.3 GENERAL METHODS

6.3.1 RIPPER

Rule learning was performing JRip, the implementation of RIPPER (Repeated Incremental Pruning to Produce Error Reduction)[232] bundled with WEKA 3.6.9 (64-bit). RIPPER performs a form of incremental reduced error pruning (IREP). In, reduced error pruning (REP), a training set is split into a growing set and a pruning set. Cases are covered by some chosen heuristic (the particulars of which do not matter here). Once all of the cases in the growing set have been covered, the rule set (which likely over-fits the data) is pruned by iteratively removing antecedents or whole rules. Pruning stops when further actions would increase error on the pruning set. Unfortunately, REP is very computationally inefficient. IREP improves upon this method by pruning rules as soon as they are generated. In RIPPER's implementation, cases are sequentially covered and removed, starting with the smallest class, until only the largest class remains. That remaining class is covered by a default rule with no antecedents. RIPPER has been demonstrated to be on par with C4.5[233] in terms of accuracy measures, but significantly faster.

The results presented below indicate that RIPPER was able to learn rules for benchmark and PDB data. However, the error rates are greater than hoped. It is regrettable that I was unable to successfully install and configure the tool I had proposed to use, RL[234-239]. RL has been shown to have error rates competitive with RIPPER and C4.5rules[240]. Other attractive abilities that RL has and RIPPER lacks are covering cases with replacement, setting minimum rule accuracy, setting minimum case coverage per rule, handling hierarchical attributes, and being agnostic.

The first rule to cover a case may not be the best rule, and covering with replacement allows RL to avoid greediness. Setting minimum rule accuracy helps RL modulate error rates per rule, rather than only for whole rule sets. Value hierarchies for attributes allow RL to cover cases at whatever granularity levels are most appropriate, on a per-rule basis. Lastly, and perhaps most importantly, RL is capable of being agnostic, which means there are no default class predictions. This is a desirable trait, because an appropriate response to insufficient evidence is often indecision. Agnostic classification gives a better indication of the actionable evidence available than using a default class prediction. It also allows users to distinguish between confident and deliberate predictions of a class on one hand, and prediction due to lack of contrary evidence on the other. To force RIPPER into learning rules for all cases, dummy cases (labeled “NULL”) were added to data sets. In this way, the default rule predicted the dummy class.

6.3.2 Rule Extraction Evaluation Criteria

- **P**: positives; number of positive cases in a set; TP + FN
- **N**: negatives; number of negative cases in a set; TN + FP
- **TP**: true positives; number of correctly predicted positive cases

- **TN**: true negatives; number of correctly predicted negative cases
- **TP rate**: also called sensitivity or recall; TP/P; TP/(TP + FN)
- **FP rate**: also called false alarm rate; 1 – specificity; FP/(FP + TN)
- **Sensitivity**: see TP rate
- **Specificity**: TN/(TN + FP)
- **Recall**: see TP rate
- **Precision**: also called positive predictive value (PPV); TP/(TP + FP)
- **False alarm rate**: see FP rate
- **PPV**: positive predictive value; see precision
- **F-measure**: harmonic mean of precision and recall; 2*(precision*recall)/(precision + recall)

6.4 OUTLIER SCORE EXPLANATION AND PREDICTION

6.4.1 Benchmark Experiment Methods

Based on the benchmark outlier scoring experiment (sections 5.4 and 5.5), outlier scores from the unreduced ICA experimental condition with `MinPts` = 20 for each benchmark data set were merged with their associated raw input attributes. Equal-width discretization was applied to the target attribute in each condition, but no explicit input attribute discretization was performed. First, ten bins were tried, then three. `RIPPER`, using default settings, was run for each condition data set, with generalizability estimated using ten-fold stratified cross-validation. “Stratified” means that the target class proportions in each fold are roughly the same as those of the whole set.

6.4.2 Benchmark Experiment Results

A complete listing of rules can be found in the supplementary material[m].

The overall accuracies for the rule sets are better than random. However, in some cases they are only just barely so (Table 10, Table 13). Note that all performance statistics are presented with respect to fidelity to assigned outliers scores, not ground truth labels. This is an important distinction, since scientifically interesting data typically do not have labels for inliers and outliers.

Ten-bin discretization of the outlier score attribute was tried first because it seemed intuitive enough to satisfice. The results were mixed. On the positive side, cases with outlier scores in the highest bin were sufficiently different from other cases to be covered by narrow “niche” rules.

[m] http://d-scholarship.pitt.edu/19601/1/SuppA1_RIPPER_winequality_ICA_notrim_k=20_3bins.txt

http://d-scholarship.pitt.edu/19601/2/SuppA2_RIPPER_winequality_ICA_notrim_k=20_10bins.txt

http://d-scholarship.pitt.edu/19601/3/SuppB1_RIPPER_secom_ICA_notrim_k=20_3bins.txt

http://d-scholarship.pitt.edu/19601/4/SuppB2_RIPPER_secom_ICA_notrim_k=20_10bins.txt

http://d-scholarship.pitt.edu/19601/5/SuppC1_RIPPER_totalcrime_ICA_notrim_k=20_3bins.txt

http://d-scholarship.pitt.edu/19601/6/SuppC2_RIPPER_totalcrime_ICA_notrim_k=20_10bins.txt

http://d-scholarship.pitt.edu/19601/7/SuppD1_RIPPER_cardiotocography_ICA_notrim_k=20_3bins.txt

http://d-scholarship.pitt.edu/19601/8/SuppD2_RIPPER_cardiotocography_ICA_notrim_k=20_10bins.txt

On the negative side, almost no predictions were made for cases with scores between the highest and lowest bins (Table 11).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Total Crime	37.3%	3.8%	45.9%	37.3%	41.2%	71.4%
Cardiotocography	35.7%	2.7%	48.6%	35.7%	41.1%	65.1%
SECOM	20.5%	3.5%	35.7%	20.5%	26.0%	57.6%
Wine Quality	25.8%	2.1%	51.6%	25.8%	34.4%	65.5%

Table 10 Benchmark performance statistics for (0.9, 1.0) bin of 10-bin discretization

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
98.9%	88.4%	24.0%	98.9%	38.7%	55.8%	[0.0-0.1]
0.4%	0.2%	13.3%	0.4%	0.7%	54.0%	(0.1-0.2]
0.9%	0.2%	31.3%	0.9%	1.7%	54.4%	(0.2-0.3]
1.1%	0.2%	38.9%	1.1%	2.0%	53.9%	(0.3-0.4]
0.7%	0.1%	50.0%	0.7%	1.4%	53.4%	(0.4-0.5]
0.7%	0.3%	18.2%	0.7%	1.3%	51.2%	(0.5-0.6]
0.7%	0.1%	33.3%	0.7%	1.3%	51.1%	(0.6-0.7]
2.6%	0.8%	22.2%	2.6%	4.6%	51.7%	(0.7-0.8]
15.0%	1.9%	40.4%	15.0%	21.9%	61.3%	(0.8-0.9]
25.8%	2.1%	51.6%	25.8%	34.4%	65.5%	(0.9-1.0)

Table 11 Wine Quality performance statistics for 10-bin discretization, unreduced ICA condition, MinPts = 20

- If
    - volatile acidity >= 0.41 and
    - free sulfur dioxide >= 32 and
    - density >= 0.9982 and
    - free sulfur dioxide >= 48 and
    - residual sugar <= 14.7
- Then
    - Outlier score is in (0.9, 1.0)
    - coverage = 13, FP = 0

Table 12 Simplified example rule predicting 10-bin outlier score for the wine quality data set

Since ten-bin discretization appeared to be too fine, a coarser discretization of three bins was tried. The rationale was to triage cases into three intuitive categories: almost certainly and inlier, almost certainly an outlier, and ambiguous/complicated. With this scheme, overall accuracies were much greater, but at the price of rather broad predictions (Table 14).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Total Crime	57.7%	12.8%	65.5%	57.7%	61.4%	78.8%
Cardiotocography	56.4%	12.1%	64.6%	56.4%	60.2%	80.8%
SECOM	27.7%	5.5%	64.6%	27.7%	38.8%	63.5%
Wine Quality	53.8%	10.4%	65.8%	53.8%	59.2%	74.3%

Table 13 Benchmark performance statistics for (0.67, 1.00) bin of 3-bin discretization

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
83.4%	44.4%	58.0%	83.4%	68.4%	72.6%	[0.00-0.33]
25.9%	12.9%	46.8%	25.9%	33.3%	59.6%	(0.33-0.67]
53.8%	10.4%	65.8%	53.8%	59.2%	74.3%	(0.67-1.00)

Table 14 Wine Quality performance statistics for 3-bin discretization, unreduced ICA condition, MinPts = 20

- If
  - volatile acidity >= 0.485 and
  - color = white and
  - pH <= 3.17
- Then
  - Outlier score is in (0.67, 1.00)
  - coverage = 77, FP = 5

Table 15 Simplified example rule predicting 3-bin outlier score for the wine quality data set

Further research is required to determine why discretization has such a strong influence on rule extraction from outlier scores. A satisficing *via media* of five bins was chosen for extracting rules from PDB data.

Another observation from the learned rules was that RIPPER's discretization leads to multiple conjuncts with the same attribute in a single rule. That is, when a rule is specialized to cover a narrower range of values for a particular attribute, the more general conjunct remains in the rule. This behavior is highly undesirable, particularly in the presence of over 100 attributes, such as my PDB data have. Consequently, PDB input attributes were explicitly discretized prior to rule extraction (as described below).

6.4.3 PDB Experiments Methods

6.4.3.1 Experiment 1

In the first experiment, the effects of ICA transformation and two types of dimensionality reduction on rule induction were examined. Outlier scores from each of the experimental conditions (ICA without reduction, ICA with simple reduction, ICA with semantic reduction, and scaled control) were merged with raw validation score attributes. Five-bin equal-width discretization was applied to the target attribute in each condition. Target attribute discretization was followed by supervised minimum description length (MDL) discretization[241,242] for the input attributes. `WEKA`'s supervised `Discretize` filter was used, with `useBetterEncoding` and `useKononenko` set to `TRUE`. Kononenko's method[241] is an improvement on Fayyad and Irani's method[157] that is less biased toward creating numerous splits. `RIPPER`, using default settings, was run on each condition, with generalizability estimated using ten-fold stratified cross-validation.

Rule extraction from benchmark tests suggested that neither ten equal-width bins nor three was ideal for discretizing outlier scores. With ten bins, only the highest and lowest bins were covered with reasonably high recall and precision; the rest were rarely predicted at all, despite having similar cardinality. In contrast, using three bins produced similar recall and precision for all of them, but the granularity is too coarse for use with PDB data. This is because the proportion and number of models flagged as potential anomalies – 29% of the 29,084 scored training cases – would be intolerably large. There would be far too many cases in need of further inspection. Furthermore, flagging that many cases arguably stretches the notion of "outlier" too far.

Even with five equal-width bins, the proportion of models labeled as potential anomalies is quite large (13.9% of the scored data). Nevertheless, I decided this was an adequate satisficing solution in the near term. Future efforts will include optimization of outlier score discretization. However, I also intend to replace `RIPPER` with `RL`, which I strongly suspect will have unforeseen interactions with different discretization schemes.

6.4.3.2 Experiment 2

In response to the resulting performance statistics of the first experiment, an additional training run was performed. To "trick" RIPPER into making agnostic predictions, 8000 dummy cases were added to the data set, having missing values for all input attributes and a new default class bin. The number 8000 was chosen to be larger than the least populated discretized bin, thus causing `RIPPER` to learn rules for all real classes.

Supervised MDL discretization was used on the input attribute values, as described for previous experiments. Again, default RIPPER settings were used, and ten-fold stratified ross-validation was performed to estimate the generalizability of learned rules.

6.4.3.3 Experiment 3

In the third experiment, the rules learned from the training data with complete outlier scores were then applied to a test set comprised of incomplete sets of input validation score attributes and no outlier scores. Predicted classes were then compared to case labels based on whether cases are in the current or obsoleted PDB, and semi-objective assessments of whether cases are suspicious or not. These assessments were based on documented frauds, known retractions, and my amateur assessments of model validity/quality. The last were based primarily on R-factors and ProCheck stereochemistry reports. Evaluated cases were found in journal articles about prominent retractions or validation tools and practices[8,16,17,29,31,74,81,82,85,86,89,123,148,243-246].

6.4.4 PDB Experiments Results

Table 16 reports F-measure scores for cross-validated class predictions in each experimental condition. Note that entries marked with an asterisk are the default class and therefore represent very high FP rate. Complete rules sets for the unreduced ICA condition can be found in the supplementary material[n].

[n] http://d-scholarship.pitt.edu/19601/16/SuppL_RIPPER_PDB_ICA_notrim_k=20_5bins.txt

http://d-scholarship.pitt.edu/19601/17/SuppM_RIPPER_PDB_ICA_notrim_k=20_5bins_agnostic.txt

Condition	[0.0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1.0]
control (scaled)	.391*	.006	.005	.015	.402
ICA, no reduction	.436	.053	.423*	.089	.463
ICA, simple reduction	.434	.042	.414*	.061	.444
ICA, semantic reduction	.414	.041	.410*	.036	.327
ICA, no reduction (w/ dummy)	.437	.056	.019	.091	.443

Table 16 F-measure scores for cross-validated rule learning for PDB data. Asterisks indicate misleading values due to RIPPER's use of a default class.

Consistent with the results of the benchmark experiments, the unreduced ICA condition produced the best recall, precision, and F-measure for the deep inlier – [0.0, 0.2] –and extreme outlier – (0.8, 1.0] – classes. Performances on the other next lowest and next highest bins were also superior for that condition, though TP rates are too low to be particularly useful.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
35.1%	5.6%	57.4%	35.1%	43.6%	70.1%	**[0.0-0.2]**
2.9%	1.5%	35.2%	2.9%	5.3%	57.2%	**(0.2-0.4]**
88.6%	73.0%	27.8%	88.6%	42.3%	58.1%	**(0.4-0.6]**
4.9%	1.8%	45.1%	4.9%	8.9%	58.7%	**(0.6-0.8]**
36.8%	3.6%	62.5%	36.8%	46.3%	72.1%	**(0.8-1.0]**

Table 17 Performance statistics for rules learned from outlier scores based on the unreduced ICA condition (without dummy class)

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
35.7%	4.4%	56.4%	35.7%	43.7%	70.1%	**[0.0-0.2]**
3.1%	1.3%	33.1%	3.1%	5.6%	55.7%	**(0.2-0.4]**
1.0%	0.6%	28.0%	1.0%	1.9%	54.4%	**(0.4-0.6]**
5.1%	1.4%	43.3%	5.1%	9.1%	57.5%	**(0.6-0.8]**
34.6%	2.6%	61.7%	34.6%	44.3%	70.6%	**(0.8-1.0]**

Table 18 Performance statistics for rules learned from outlier scores based on the unreduced ICA condition (with dummy class)

	Improper Subset Attributes	Rules	Overall Accuracy	Sensitivity	Specificity	Recall	Precision	Area Under ROC Curve
Control (Scaled)	151	27	27.1%	29.8%	96.7%	29.8%	61.5%	64.2%
ICA, unreduced	151	66	34.4%	36.8%	96.4%	36.8%	62.5%	72.1%
ICA, simple reduction	122	97	33.4%	34.4%	96.6%	34.4%	62.3%	71.0%
ICA, semantic reduction	58	79	31.7%	23.0%	97.2%	23.0%	56.2%	65.6%

Table 19 Number of attributes (in the improper subset) used for outlier scoring, number of rules extracted, and performance statistics for predictions. Overall accuracy refers to all predictions. Other statistics given are related specifically to predictions for the (0.8, 1.0) bin.

The performance of rule sets based on outlier scores from each of the experimental conditions seems to indicate that semantic reduction did not work as hoped (Table 19). The extracted rules in the semantic reduction condition were conservative in predicting the highest outlier score bin, resulting in the worst sensitivity and the best specificity of all the conditions. However, while (relatively) high specificity means that negatives are predicted (relatively) well, the poor precision means that many cases predicted to have high outlier scores do not.

The premises behind performing semantic reduction were that attribute redundancy contributed to the curse(s) of dimensionality and reducing or eliminating redundancy would improve outlier detection. Without ground truth labels, this is difficult to prove or disprove. However, I posited (Hypothesis 7) that if semantic reduction facilitated the identification of robust outlier score patterns and trends, they would be reflected in better rule learning performance (with respect to other experimental conditions, especially no reduction). As already indicated, the results do not support this hypothesis. Rather, they appear to agree with studies regarding intrinsic dimensionality and self-similarity, which indicate that far from being problematic, redundancy is beneficial to anomaly detection and irrelevant or noisy attributes are problematic[193-200,202-204,247-250]. In brief, those studies suggest that redundancy produces a beneficial concentration of distances by decreasing intra-cluster distances and increasing inter-cluster distances. Outliers may be regarded as singleton or low-membership clusters.

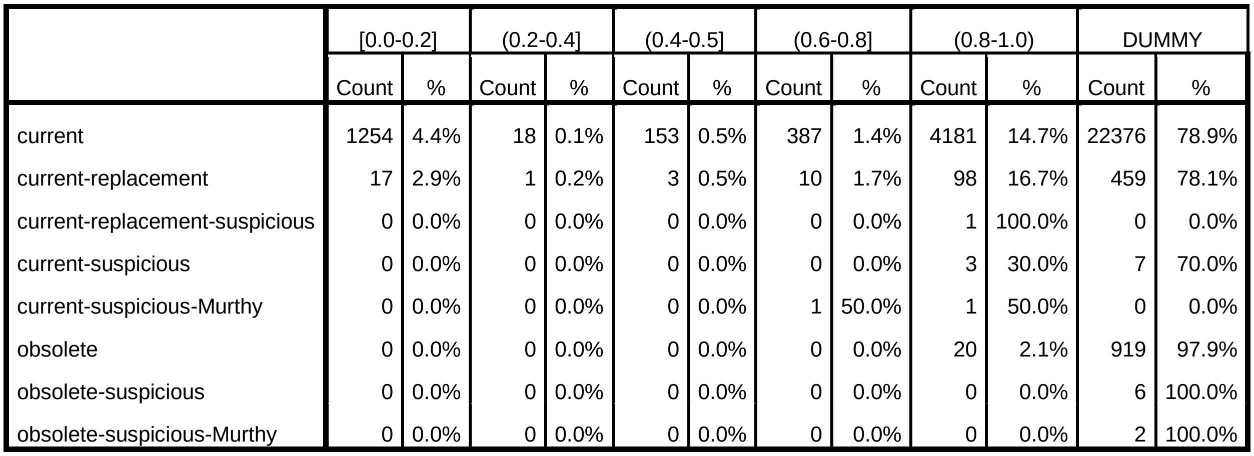
Simple dimensionality reduction appears to have performed as well as no reduction. Sensitivity/recall is slightly lower in the reduced condition, but only slightly. The 29 eliminated components do not seem to have significantly affected RIPPER’s ability to learn rules. Interestingly, simple reduction led to the induction of nearly 50% more rules than the unreduced condition. Further research is needed to determine why that is and how the rule sets differ. Rule learning also about twice as long in the simple reduction condition, suggesting that more rules had to be tried and rejected in order to cover the data. Nevertheless, the learned rules perform almost as well.

- If
  - SFCheck’s 2σ R-factor is in (0.1575, 0.2215] and
  - Asymmetric unit volume is in (77355, 232187] Å^3 and
  - SFCheck’s optical resolution is in (1.25-1.55] Å and
  - σ(ADP) is in (0.074-0.112] Å and
  - The difference between SFCheck’s recomputed R_{work} and reported R_{work} is in (0.004-0.019] and
  - The rms deviation of B-factors for bonded atoms is in (1.15-1.96] Å and
  - MolProbity’s unfavored Ramachandran pair percentage is in (1.10-2.77] and
  - σ(ω torsion angle) is in (4.83-6.49] degrees and
  - SFCheck’s diffraction scale factor is in (0.93-1.03] and
  - ProCheck’s percentage of highlighted Ramachandran pairs is in (0.0773-1.56]
- Then
  - Outlier score is in [0, 0.2], Coverage = 102, FP = 8

Table 20 Example of a (simplified) rule predicting outlier scores from the unreduced condition

Table 21 shows the results of applying extracted rules to model that were not score by LOF. The large number of agnostic non-predictions highlights the need for an inductive rule learner that was properly designed to handle insufficient or contradictory data. The other results are interesting, but further research is required to study the actually predicted cases before any definitive diagnostic statements can be made. However, one can speculate.

The number of *current* PDB entries classified as having a greater than 80% chance of being outliers, 4181, seems rather high. However, it is also less than 15% of all the unscored current cases. Similarly, almost 17% of the *current-replacement* models are predicted to be outliers. It seems likely that many (or perhaps even all) of these cases are outliers because they share common traits that are not well represented in the scored cases. It is encouraging to see that all but one of the cases that I deemed suspicious, and all of the obsolete cases, that did not receive the agnostic class were classified as outliers. The one case that was not a very likely outlier still had a 60-80% chance of being an outlier. Whether these predictions are supportive of my scoring and rule extraction methods, my classifications, or both remains to be seen.



	[0.0-0.2]		(0.2-0.4]		(0.4-0.5]		(0.6-0.8]		(0.8-1.0)		DUMMY	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
current	1254	4.4%	18	0.1%	153	0.5%	387	1.4%	4181	14.7%	22376	78.9%
current-replacement	17	2.9%	1	0.2%	3	0.5%	10	1.7%	98	16.7%	459	78.1%
current-replacement-suspicious	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%
current-suspicious	0	0.0%	0	0.0%	0	0.0%	0	0.0%	3	30.0%	7	70.0%
current-suspicious-Murthy	0	0.0%	0	0.0%	0	0.0%	1	50.0%	1	50.0%	0	0.0%
obsolete	0	0.0%	0	0.0%	0	0.0%	0	0.0%	20	2.1%	919	97.9%
obsolete-suspicious	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	6	100.0%
obsolete-suspicious-Murthy	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%

Table 21 Predicted outlier score ranges for cases that were not scored by LOF

6.5 EXPLANATION OF DEEP INLIERS AND EXTREME OUTLIERS

6.5.1 Introduction

As indicated above, analysis of the confusion matrices from rule learning runs revealed that target classes are not predicted in proportion to their cardinality. Rather, only the deepest inlier and most extreme outlier classes are predicted with high TP and low FP rates. The other classes are only rarely predicted, except for the default class, which is predicted at a high and indiscriminate rate. High sensitivity and specificity across all classes would be ideal, but valuable information can still be learned from extracted rules predicting deep inliers and extreme outliers.

6.5.2 Methods

For this experiment, standardized outlier scores generated from unreduced ICA transformations were used. Equal-width binning was performed for 5, 10, 20, and 40 bins, with each bin corresponding to 20%, 10%, 5%, and 2.5% of the [0,1] interval, respectively. Once the target attribute was discretized, the input attributes were discretized using a supervised minimum description length method[241,242].

RIPPER runs were performed to predict the preceding outlier score ranges of the varying granularity. Default `WEKA JRip` settings were used for learning. Having established the generalizability of rules predicting outlier score ranges in the previous experiments using cross-validation, evaluations for this experiment were performed on just the training set. For each run, rules were extracted if their predicted class had a recall rate of at least 20% and a precision rate of at least 60%. In truth, these criteria are post hoc, since the selection was based on a heuristic visual search for classes with “high” recall and precision. Due to the high contrast between the classification statistics for different bins, this was not a difficult or ambiguous task.

6.5.3 Results

The above criteria resulted in the extraction of rules predicting deep inliers with outlier scores in the intervals [0.0, 0.2], [0.0, 0.1], [0.00, 0.05], and [0.000, 0.025]; and extreme outliers in the intervals (0.8, 1.0], (0.9, 1.0], (0.90, 0.95], (0.95, 1.00], (0.950, 0.975], and (0.975, 1.000].

6.5.3.1 Deep Inliers

There are strong trends among rules predicting deeply inlying outlier scores. Asymmetric unit volume is near average, R-factors are near averages, recomputed R-factors are very close to reported values, main chain B-factors have about half the variance of and slightly lower mean than side chain B-factors, covalent is tightly distributed around reference means, and stereochemistry scores are very good, with very few bumps or clashes. These results would not be surprising to a structural biologist. Deep inliers described by these rules have good agreement between observed and calculated diffraction patterns, “textbook” covalent geometry, and physically reasonable stereochemistry. These are not the best models in the PDB, but they were refined well according to accepted standard practices.

A complete listing of rules can be found in the supplementary material[o]. Simplified examples follow.

The rules for [0.0, 0.2] cover 5141 cases and have recall and precision of 42.4% and 68.9%, respectively. Here is an example:

- If
  - Standard deviation of the gauche$^+$ χ_1 angle is in (8.55-11.45] and
  - Overall standard deviation of u_{atomic} is in (0.0739-0.1138]' and
  - Total WhatCheck bump score is in (1.4205-8.7815] and
  - Fractional difference between R_{free} and R_{work}, recomputed by SFCheck is in (0.0849-0.2625] and
  - Fractional difference between actual and predicted ProSA pairwise z-score is in (-0.1903-0.0514] and
  - Ratio of main chain and side chain u_{atomic} standard deviations is in (0.7383-0.9680] and
  - Volume of the asymmetric unit is in (77,355-232,188] and
  - Solvent correction Ks in in (0.7975-0.8965]
- Then
  - Outlier score is in [0.0, 0.2], Coverage = 136, FP = 60

Table 22 Example of a (simplified) rule predicting outlier scores in the range [0.0, 0.2]

[o] http://d-scholarship.pitt.edu/19601/19/SuppO_deep_PDB_inlier_rules.txt

The rules for [0.0, 0.1] cover 2472 cases and have recall and precision of 23.5% and 76.1%, respectively. Here is an example:

- If
  - Difference between SFCheck's maximal and minimal estimated errors is in (0.0168-0.0821] and
  - MolProbity clash score is in (2.46-8.28] and
  - Fractional difference between actual and predicted mean u_{atomic} is in (-0.1244-0.0706] and
  - Volume of the asymmetric unit is in (234,511-299,010]
- Then
  - Outlier score is in [0.0,0.1], Coverage = 85, FP =23

Table 23 Example of a (simplified) rule predicting outlier scores in the range [0.0, 0.1]

The rules for [0.00, 0.05] cover 1372 cases and have recall and precision of 26.2% and 71.6%, respectively. Here is an example:

- If
  - SFCheck’s Luzzati score in (0.1327-0.2158] and
  - Volume of the asymmetric unit is in (77,633-342,129] and
  - ProCheck’s bad contacts per 100 residues is in (6.25-16.45] and
  - WhatCheck’s B-factor distribution RMS z-score is in (1.446-2.099] and
  - WhatCheck’s covalent bond angle RMS z-score in (0.842-1.362]
- Then
  - Outlier score is in (0.0-0.05], Coverage = 65, FP = 20

Table 24 Example of a (simplified) rule predicting outlier scores in the range [0.00, 0.05]

The rules for [0.000, 0.025] cover 579 cases and have recall and precision of 25.6% and 72.4%, respectively. Here is an example:

- If
    - SFCheck’s recomputed R-factor from all reflections with Signal/Noise >= 2 is in (0.151-0.222] and
    - Volume of the asymmetric unit is in (77,633-300,603] and
    - Best resolution used in refinement is in (1.21-2.00] and
    - WhatCheck’s B-factor distribution RMS z-score is in (1.445-2.141] and
    - Fractional difference between actual and predicted ProSA pairwise z-score is in (0.124278-inf)
- Then
    - Outlier score is in [0.000,0.025], Coverage = 43.0, FP = 14

Table 25 Example of a (simplified) rule predicting outlier scores in the range [0.000, 0.025]

6.5.3.2 Extreme Outliers

As with the deep inliers, there trends in the rules covering extreme outliers. However, due to the sparse and dispersed nature of anomalies, the rules cover a number of small, distinct “niches”, rather than large overlapping regions. There are some general commonalities, though, associated with three basic types of extreme outliers: overwhelmingly poor, highly inconsistent, and simply rare validation scores. Poor validation scores expressed in these rules include (but are not limited to) poor agreement between observed and calculated diffraction patterns, recomputed R-factors that differ greatly from reported values, and poor stereochemistry, with numerous bumps/clashes.

Validation scores can be “highly inconsistent” for at least five reasons. There may have been a processing glitch in one of the validation tools, the model coordinates data or structure factors data file may contain errors, certain data conditions may violate assumptions built into validation programs, there may be unresolved errors in the model (such that attempts to optimize certain validations score cause others to degrade), or fraud may have been committed. Naturally, structural biologists and consumers of protein structure models hope that fraud is a very rare phenomenon. However, as recent scandals have reminded us, we cannot afford to ignore the possibility[13,76,130-132,134,251]. As described in section 2.5.4, a telltale feature of recent frauds has been physically implausible combinations of inconsistent validation scores.

Some extreme outliers may be simply rare, rather than truly anomalous, existing in the tails of inlier distributions. For instance, very small and very large proteins are both rare in the PDB. Certain refinement parameters choices may also be rare, dictated by necessity, and unrelated to model quality. Additionally, the combinatorial explosion associated with high dimensionality suggests that some cases will be outliers in at least one dimension by chance. When a data set contains tens of thousands of cases, such as the PDB does, it stands to reason that some of these anomalies may cluster with sufficient frequency to be explicitly covered by a rule.

Interestingly, learned rules do not seem to extensively cover another known type of extreme outlier: overwhelmingly excellent validation scores. Due to the prerequisite of very high resolution/quality diffraction – and very difficult to obtain – data, I speculate that there are inadequate exemplars for RIPPER to train on. Alternately, these cases may be covered (not necessarily correctly) by other rules, and because `RIPPER` does not cover with replacement, there is no opportunity to cover them separately.

Some of the rules covering extreme outliers bear further scrutiny, such as those that rely on indicators not commonly used in validation. Detailed study of these rules will be among the specific aims of a grant application to be made shortly after the defense of this dissertation.

A complete listing of rules can be found in the supplementary material[p]. Simplified examples follow.

The rules for [0.8, 1.0] cover 4031 cases and have recall and precision of 38.6% and 64%, respectively. Here is an example:

- If
    - WhatCheck covalent bond angle RMS deviation is in (1.736-2.519] and
    - WhatCheck bumps per residue is in (0.2475-0.3175] and
    - WhatCheck bumps in the second mildest bin is in [0.0-6.5] and
    - Ratio of diffraction eigenvalues 2 and 3 is in (0.6337-0.9598]
- Then
    - Outlier score is in (0.8-1.0], Coverage = 34, FP = 11

Table 26 Example of a (simplified) rule predicting outlier scores in the range (0.8, 1.0]

[p] http://d-scholarship.pitt.edu/19601/18/SuppN_extreme_PDB_outlier_rules.txt

The rules for [0.9, 1.0] cover 1568 cases and have recall and precision of 36.5% and 65.5%, respectively. Here is an example:

- If
  - MolProbity’s clash score is in (38.82-58.30] and
  - MolProbity’s Ramachandran outliers are in (3.775-inf) and
  - SFCheck’s observed and calculated diffraction data correlation is in (0.7555-0.8403]
- Then
  - Outlier score is in (0.9-1.0], Coverage = 14, FP = 4

Table 27 Example of a (simplified) rule predicting outlier scores in the range (0.9, 1.0]

The rules for [0.90, 0.95] cover 870 cases and have recall and precision of 22.2% and 74.2%, respectively. Here is an example:

- If
  - Standard deviation of the gauche$^+$ χ_1 angle is in (17.15-19.55] and
  - Volume of the asymmetric unit is in (28,935-48,328] and
  - Fractional difference between SFCheck's recomputed Rfree and Rwork is in (-0.1518-0.0849] and
  - SFCheck's observed and calculated diffraction data correlation is in (0.7555-0.8582]
- Then
  - Outlier score is in (0.90-0.95], Coverage = 9, FP = 2

Table 28 Example of a (simplified) rule predicting outlier scores in the range (0.90, 0.95]

The rules for [0.95, 1.00] cover 698 cases and have recall and precision of 53.3% and 76.7%, respectively. Here is an example:

- If
  - SFCheck's recomputed Rfree is in (0.3285-0.4205] and
  - The standard deviation of the gauche$^+$ χ1 torsion angle is in (19.55-inf) and
  - The difference between SFCheck's recomputed R-factors is in (0.011-0.044] and
  - The MolProbity score is in (3.67-inf)
- Then
  - Outlier score is in (0.95-1.00], Coverage = 22, FP = 6

Table 29 Example of a (simplified) rule predicting outlier scores in the range (0.95, 1.00]

The rules for [0.950, 0.975] cover 331 cases and have recall and precision of 59.2% and 75.1%, respectively. Here is an example:

- If
  - MolProbity clash score is in (38.81-58.30] and
  - Standard deviation of the omega torsion angle is in (7.15-11.35] and
  - Fractional difference between actual and predicted mean u_{atomic} is in (-0.000812-0.000000] and
  - Difference between SFCheck's recomputed and reported Rwork is in (0.000-0.026]
- Then
  - Outlier score is in (0.95-0.975], Coverage = 7, FP = 3

Table 30 Example of a (simplified) rule predicting outlier scores in the range (0.950, 0.975]

The rules for [0.975, 1.000] cover 367 cases and have recall and precision of 76.6% and 78.7%, respectively. Here is an example:

- If
  - Asymmetric unit volume is <= 13,875 and
  - Fractional difference between actual and predicted mean u_{atomic} is in (-0.000812 - 0.000000] and
  - The difference between SFCheck’s maximal and minimal estimated error is in (0.0773-0.1236]
- Then
  - Outlier score is in (0.975-1.000], Coverage = 21, FP = 9

Table 31 Example of a (simplified) rule predicting outlier scores in the range (0.975, 1.000]

7.0 FUTURE WORK

There are a number of potentially fruitful paths of follow-up from this work. Some of them are described below, grouped loosely by common themes. The following list is by no means exhaustive. Rather, it is meant to express some of the unexplored breadth and depth opened by my methods and the techniques they employ.

7.1 LOCAL VALIDATION

All of the validation scores used to detect anomalies among PDB entries were of a global nature, evaluating whole models. Thorough validation requires local validation, too, though. Local validation evaluates models at the residue, bond, or atom level and reports anomalous features *within a model*, as opposed to reporting whole models that are anomalous with respect to other models in a set. Examples include atoms with unusually low or high B-factors, unlikely combinations of torsion angles, violations of van der Waals radii, and unusual covalent bond lengths or angles. Indeed, such measures are aggregated (e.g., as outlier percentages) for global (whole-model) validation.

Large-scale automation of local validation would require an approach to anomaly detection very different from the one applied in this study. Since proteins are of differing sequence lengths, sequence compositions, and three-dimensional structure, direct comparisons of models (for dissimilarity matrix calculations) are usually not feasible. Instead, local patterns much somehow be transformed and/or aggregated to facilitate comparisons. Furthermore, we are not interested (in this context) in how protein models differ sequentially or structurally, but how local regions of models that differ from physical expectations in improbable ways. One possible approach would be to extract sets of features from curves representing per-residue statistics (B-factors, real space correlation, etc.) and use pattern recognition methods to find models with anomalous counts, patterns, or types of troubling or rare features.

7.2 TRANSFORMATION AND REDUCTION

There are options left to be explored in attribute transformation and dimensionality reduction, including different eigenvector retention criteria (aka stopping criteria) for PCA[252,253]. It would also be interesting to explore non-linear dimensionality reduction via manifolds (local linear embedding, diffusion maps, etc.). Non-linear transformations and reductions are particularly appealing for use with validation data, which are known to have non-linear correlations.

7.3 K NEAREST NEIGHBORS

While k nearest neighbors methods have been subjected to considerable study, there may still be aspects that can be "tweaked" to find different kinds of relationships in data. There appear to be several questions that may be interesting to answer.

How big does k have to be to make two given points nearest neighbors? Can that value be used as a sort of geodesic distance? Can anything be learned from examining distributions of values for that new distance? Given some point, A, what fraction of A’s k nearest neighbors also include A as a neighbor? Could that value be used as a local density measure?

7.4 LOF

LOF is a simple algorithm, but it interacts with complex characteristics of the data on which it operates. Only a small fraction of the possible permutations of those characteristics were tested in this study. It would be useful to more exhaustively test combinations of attribute standardizations, attribute transformations, dimensionality reductions, MinPts values, contextualizations, context score merging methods, underlying data distributions, and distance/dissimilarity measures.

A potentially fruitful deviation from the methods in this study would be scaling rows of data in addition to and instead of columns. Calculating the distance between row-standardized points gives a sort of correlation coefficient based on the cosine of the angle between two unit vectors. In such a scheme, points are closer together if their unit vectors points in the same direction. Nearest-neighbor calculations using this metric are likely to differ greatly from those of column-standardized data, and local densities based on them may identify very different types of anomalies.

7.5 CONTEXTS

Some additional work on contextual anomaly detection remains. Cases that receive outlier scores in some contexts, but not all (likely due to missing attributes being treated differently in different contexts), should be investigated. Even among those cases with scores in all contexts, highly discordant scores between contexts should be investigated.

The completion of a task that was part of this study’s proposal, but was abandoned due to insufficient time, may be beneficial. While PDB models are not labeled with outlier status, and several validation scores may be missing for any given model, most contextual attributes are available for all models. I hypothesize that contextual attribute values can be predicted from validation score values, and that outliers may be identified as models differing significantly from others in the same context. The contexts in this domain are well established by physical theory, and strong trends of similarity are expected between members of the same context. Deviant cases in a context are certainly worthy of closer examination.

This method seeks to turn contextual outlier detection in its head. Instead of trying to identify suspicious models based on known labels (of which there are very few), the contextual attributes are used as targets and the validation scores are inputs. This method is promising because it is based on supervised prediction or classification of values that are already known by the relevant domain to be discriminative, and because it is relatively simple to implement.

For instance, a model deposited in 2000 that is classified as having been deposited 1980 was likely not built or refined according to the standards of 2000, and is likely to contain systematic errors. Similarly, a model refined at high resolution that is predicted to be low resolution is likely to be problematic. Conversely, a model predicted to be in a “better” context than it truly belongs to may be anomalously “good” relative to other models.

7.6 HUBS

Exploration of hubs in high-dimensional data sets has only recently begun. While some efforts have been made to use hubs as cluster centroids[222], I am unaware of any attempts to use deep LOF inliers in that way. Even if LOF values less than 1.0 are not found to be proper hubs, there may be value in expanding on my preliminary examination of them as exemplars of deeply inlying cases. LOF hubs may be useful for identifying de facto standards of commonly accepted practices, especially with respect to different contexts related to tools, techniques, protein sequence characteristics, structure characteristics, or crystal characteristics (to name but a few possibilities).

7.7 ULTRAMETRICITY

“Ultrametricity” and fractal dimensionality are promising new areas of research in high-dimensional data mining and anomaly detection[193,194,196,198-200,250]. I suspect that they can be applied to structural biology data in interesting ways. Ultrametricity refers to benefits of one of the supposed curses of dimensionality, namely the concentration of distances. In brief, the theory states that as the number of correlated dimensions increases, intra-cluster points concentrate, and inter-cluster points separate. This property of increasing contrasts between clusters is sometimes referred to as the “blessing of self-similarity”. “Ultrametricity” refers to how distances in very high dimensions approximately obey a stronger form of the triangle inequality – an ultrametric distance function that restricts all triplets of points to form equilateral triangles (within clusters) or small-angled isosceles triangles (between clusters).

It may be possible to exploit ultrametric properties in feature spaces based protein structure models. For instance, using a structural biology database (PDB, CATH[254], SCOP[255], etc.), one could find the triplets of points that are approximately equidistant. If all singletons that form approximately equilateral triangles are clustered, are there any leftover points? Are they anomalies?

Additionally, one could join clusters in triplets that are approximately equidistant by average linkage. This could be continued until no more clusters can be so joined, thereby producing a different kind of hierarchical clustering. What are the properties of clusters identified this way? How might a hierarchical clustering performed in this way differ from classical methods?

Alternately, find all triplets (with replacement) that form approximate equilateral triangles. Count the number of times each pair of points appears in such a triplet. Use that value as a similarity measure. How does the distribution of distances differ from those calculated using classical methods? How would calculating distance in this way affect distributions of LOF scores?

Lastly, count the number of times each point appears in a triplet as just described and explore the distribution of counts and locations of points with high counts. Are points with very low counts outliers? Are points with very high counts hubs?

8.0 SUMMARY AND CONCLUSIONS

JMR: "Well, a nice mess you've got us into, with your nodding head and the deference due to a local outlier factor."

EDW: "Merely corroborative detail, intended to give a statistical verisimilitude to an otherwise bald and unconvincing narrative."

(Adapted from "The Mikado", with apologies to Gilbert and Sullivan[256])

8.1 PRIMARY RESULTS

At the outset of this study, I hypothesized that outlier detection could be used to identify anomalous protein structure models in the PDB. This hypothesis was later supported by the analysis of the models with the most extreme outlier scores in the set, all of which had demonstrably anomalous validation scores. Some could be attributed to reporting errors, and others to possible validator parsing errors. Most encouraging were the two models from a documented fraud, because no attempt was made to deliberately score them highly.

Despite initial optimism that verifiable and defensible anomalies such as those, I knew that I would need a means of explaining and justifying outlier scores would be needed for the structural biology community. The path from validation data to outlier scores, which includes attribute transformations and an instance-based scoring function dependent on emergent phenomena, is rather opaque and unintelligible. This could engender healthy skepticism in any domain, including structural biology. Therefore, I hypothesized that outlier scores could be rendered comprehensible and explainable by extracting disjunctive sets of conjunctive IF-THEN rules that associate scores with values from small groups of well-understood validators. Support for this hypothesis was partially established by demonstrating better-than-random accuracy for rules predicting outlier scores from validation data. Greater statistical confidence (in the form of better sensitivity/recall, specificity, and precision) would have been preferable. However, previous research has indicated that a rule with low statistical confidence may nevertheless be interesting if it advances understanding of the domain. For instance, in one study senior doctors found rules in the low to moderate accuracy range of 40-60% to be novel, interesting, and more accurate than the knowledge of some junior doctors[257]. Low statistical confidence may be a consequence of small sample sizes for sparse outlying clusters, rather than a methodological failure[258]. The use of a suboptimal inductive rule learner may also be at fault, and my future investigations will hopefully utilize `RL` instead of `RIPPER`. On a positive note, the rules extracted for the most inlying and outlying models implicate some validators that are less commonly used or emphasized, which is an interesting development. However, it remains to be seen whether these insights are accepted by the domain, as demonstrated by acceptance for publication.

8.2 SECONDARY RESULTS

Some additional insights were acquired from the benchmark experiments used to calibrate and validate my methods. The apparent superiority of small neighborhood sizes was quite surprising to me, and bears further investigation. In addition to simply observing this phenomenon in a greater number of data sets, greater diversity in data set cardinality, dimensionality, and types of data would be desirable. Furthermore, it may be fruitful to investigate more deeply into the ways neighborhood size affects the distribution of outlier scores in a set.

The superior performance of ICA-transformed data without dimensionality reduction is interesting. As far as I can tell, dimensionality reduction is always performed prior to ICA transformation in other published studies, with the justification being that without reduction there is a risk that the noisiest, rather than the most discriminative, dimension may be pursed as most interesting (least Gaussian). My results appear to contradict this rationale, having shown that dimensionality reduction appears to degrade performance. However, a larger and more diverse sample of data and dimensionality reduction criteria would be needed to properly test such a hypothesis. Likewise, it might be interesting to learn why semantic dimensionality reduction failed so badly. All of these reduction investigations could potentially add substantively to the study of intrinsic dimensionality.

8.3 CONTRIBUTION

A highly cited set of criteria for evaluating an intelligent system that aids scientific discovery states that “the design of the program, or the circumstances of its application, [should] heighten the chances that its use will lead to knowledge that is novel, interesting, plausible, and intelligible”.[259] Based on these criteria, my work is significant to both the intelligent systems community and the structural biology/bioinformatics community, as follows.

8.3.1 Novelty

First, my methods traverse an “interesting distance” toward newly automating (at least partially) aspects of the drudge work that inhibits scientific reasoning, and constitute a valuable contribution by finding a match between scientific problems and existing methods[5]. Second, using rule extraction by induction (a supervised learning method) to improve the intelligibility of outlier score assignments (from an unsupervised method) is new (though it was inspired by similar work on different machine learning tools, namely artificial neural nets and support vector machines). Third, the use of rule learning to estimate outlier scores for unscored data appears to be new. Fourth, outlier detection has not been applied to PDB data before. Fifth and finally, the concurrent combination of so many separate validation scores to study the entire PDB appears to be novel.

8.3.2 Interestingness

First, statistical analysis of validation score trends and patterns in the PDB should be of interest to structural biologists and consumers of PDB structure models. Second, lessons learned with respect to outlier detection optimization should be of interest to the intelligent systems community. Third, statistical relationships between validation scores, expressed in rules, should interest structural biologists. Fourth and finally, anyone employing outlier detection in domains with few labeled cases or many incomplete cases, and those needing to outlier score new cases without recalculating of all scores (such as in real-time systems) should be interested in my use of rule induction predict score ranges.

8.3.3 Plausibility

First, the results from experiments on benchmark data express the plausibility of the outlier scores generated for PDB entries. Second, the content and statistical characteristics of rules learned in benchmark experiments express plausibility of rules learned for PDB entries. Third and finally, rules learned from PDB entries express plausible relationships between validation scores and aspects of structure model quality.

8.3.4 Intelligibility

Machine learning tools can produce highly accurate, precise, and useful scores and predictions, but their results are not typically comprehensible. This is especially true of those that use all attributes in every prediction or opaquely operate directly on continuous-values attributes. Rules, in contrast, are intuitively intelligible, operating on categorical, nominal, or discretized continuous attributes, and typically employing only a handful of attributes in any given rule. These traits, coupled with the discrete nature of IF-THEN predictions, give rules great explanatory potential.

8.3.5 Broader Context and Applicability

The methods presented here are potentially useful beyond the chosen domain. Techniques I used, such as ICA transformation of attributes and probabilistic standardization of outlier scores, are general and not dependent on the source of data. For instance, they could be used to find structural/functional anomalies among sets of the structure models that have been appropriately purged of validation anomalies.

APPENDIX A

BENCHMARK OUTLIER SCORING RESULTS

A.1 WINE QUALITY RESULTS

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 320	k = 640	k = 650
Wine Quality	Outlier MSE	0.14	0.14	0.14	0.15	0.20	0.23	0.22	0.23
35 Outliers	Inlier MSE	0.27	0.27	0.27	0.26	0.25	0.23	0.23	0.23
6462 Inliers	E[cost]	0.14	0.14	0.14	0.15	0.20	0.23	0.22	0.23
Scaled	Rand Out MSE	0.43	0.44	0.45	0.45	0.47	0.49	0.51	0.52
	Rand In MSE	0.28	0.27	0.27	0.27	0.25	0.24	0.23	0.23
	Rand cost	0.43	0.43	0.45	0.45	0.47	0.48	0.51	0.52

Table 32 MSE and expected costs for wine quality data, control (scaled) condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 320	k = 640	k = 650
Wine Quality	Outlier MSE	0.15	0.12	0.12	0.14	0.16	0.18	0.19	0.19
35 Outliers	Inlier MSE	0.28	0.28	0.28	0.27	0.27	0.26	0.25	0.25
6462 Inliers	E[cost]	0.15	0.12	0.12	0.14	0.16	0.18	0.19	0.19
ICA	Rand Out MSE	0.43	0.42	0.43	0.42	0.43	0.45	0.46	0.46
	Rand In MSE	0.28	0.28	0.28	0.28	0.27	0.26	0.25	0.26
	Rand cost	0.43	0.42	0.43	0.42	0.43	0.45	0.46	0.46

Table 33 MSE and expected costs for wine quality data, unreduced ICA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 320	k = 640	k = 650
Wine Quality	Outlier MSE	0.31	0.23	0.22	0.25	0.26	0.26	0.25	0.25
35 Outliers	Inlier MSE	0.24	0.24	0.23	0.22	0.22	0.22	0.22	0.22
6462 Inliers	E[cost]	0.31	0.23	0.22	0.25	0.26	0.26	0.25	0.25
Reduced PCA	Rand Out MSE	0.52	0.51	0.51	0.52	0.54	0.55	0.54	0.55
	Rand In MSE	0.24	0.24	0.23	0.22	0.22	0.22	0.22	0.23
	Rand cost	0.52	0.51	0.51	0.52	0.54	0.54	0.54	0.55

Table 34 MSE and expected costs for wine quality data, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 320	k = 640	k = 650
Wine Quality	Outlier MSE	0.30	0.23	0.21	0.25	0.26	0.26	0.25	0.25
35 Outliers	Inlier MSE	0.24	0.24	0.23	0.22	0.22	0.22	0.23	0.23
6462 Inliers	E[cost]	0.30	0.23	0.21	0.25	0.26	0.26	0.25	0.25
Reduced ICA	Rand Out MSE	0.51	0.51	0.53	0.53	0.55	0.54	0.53	0.51
	Rand In MSE	0.24	0.24	0.23	0.23	0.22	0.22	0.23	0.23
	Rand cost	0.51	0.51	0.53	0.53	0.54	0.53	0.53	0.51

Table 35 MSE and expected costs for wine quality data, reduced ICA condition

A.2 SECOM RESULTS

		k = 10	k = 20	k = 40	k = 80	k = 139
SECOM	Outlier MSE	0.35	0.35	0.35	0.36	0.39
99 Outliers	Inlier MSE	0.27	0.27	0.26	0.25	0.24
1294 Inliers	E[cost]	0.34	0.35	0.34	0.36	0.38
Scale	Rand Out MSE	0.40	0.40	0.41	0.43	0.45
	Rand In MSE	0.27	0.27	0.26	0.26	0.25
	Rand cost	0.39	0.39	0.40	0.42	0.44

Table 36 MSE and expected costs for SECOM data, control (scaled) condition

		k = 10	k = 20	k = 40	k = 80	k = 139
SECOM	Outlier MSE	0.26	0.27	0.29	0.31	0.33
99 Outliers	Inlier MSE	0.29	0.29	0.29	0.28	0.27
ICA	E[cost]	0.26	0.28	0.29	0.31	0.32
	Rand Out MSE	0.35	0.36	0.38	0.41	0.43
	Rand In MSE	0.30	0.30	0.29	0.29	0.28
	Rand cost	0.35	0.35	0.37	0.40	0.42

Table 37 MSE and expected costs for SECOM data, unreduced ICA condition

		k = 10	k = 20	k = 40	k = 80	k = 139
SECOM	Outlier MSE	0.36	0.36	0.34	0.37	0.40
99 Outliers	Inlier MSE	0.26	0.26	0.26	0.25	0.24
Reduced PCA	E[cost]	0.35	0.36	0.34	0.36	0.39
	Rand Out MSE	0.41	0.41	0.42	0.44	0.47
	Rand In MSE	0.27	0.26	0.26	0.25	0.24
	Rand cost	0.40	0.40	0.41	0.43	0.45

Table 38 MSE and expected costs for SECOM data, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 139
SECOM	Outlier MSE	0.33	0.33	0.34	0.36	0.39
99 Outliers	Inlier MSE	0.28	0.27	0.26	0.26	0.24
Reduced ICA	E[cost]	0.32	0.32	0.33	0.35	0.38
	Rand Out MSE	0.37	0.38	0.40	0.42	0.45
	Rand In MSE	0.28	0.27	0.27	0.26	0.25
	Rand cost	0.36	0.38	0.39	0.41	0.44

Table 39 MSE and expected costs for SECOM data, reduced ICA condition

A.3 VIOLENT CRIME RESULTS

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Violent Crime	Outlier MSE	0.29	0.29	0.29	0.31	0.34	0.34
41 Outliers	Inlier MSE	0.28	0.27	0.27	0.26	0.25	0.25
1860 Inliers	E[cost]	0.29	0.28	0.29	0.31	0.34	0.34
Scale	Rand Out MSE	0.41	0.40	0.42	0.43	0.48	0.49
	Rand In MSE	0.28	0.28	0.28	0.27	0.26	0.25
	Rand cost	0.40	0.40	0.41	0.43	0.47	0.48

Table 40 MSE and expected costs for violent crime data, control (scaled) condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Violent Crime	Outlier MSE	0.28	0.29	0.30	0.31	0.33	0.34
41 Outliers	Inlier MSE	0.32	0.31	0.30	0.30	0.29	0.28
1860 Inliers	E[cost]	0.28	0.29	0.30	0.31	0.33	0.34
ICA	Rand Out MSE	0.35	0.36	0.37	0.37	0.40	0.41
	Rand In MSE	0.32	0.31	0.30	0.30	0.29	0.29
	Rand cost	0.35	0.35	0.37	0.37	0.40	0.41

Table 41 MSE and expected costs for violent crime data, unreduced ICA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Violent Crime	Outlier MSE	0.33	0.29	0.32	0.32	0.31	0.31
41 Outliers	Inlier MSE	0.26	0.26	0.26	0.25	0.24	0.24
1860 Inliers	E[cost]	0.33	0.28	0.32	0.32	0.31	0.31
Reduced PCA	Rand Out MSE	0.45	0.45	0.44	0.47	0.49	0.50
	Rand In MSE	0.26	0.26	0.26	0.26	0.25	0.24
	Rand cost	0.45	0.45	0.44	0.46	0.49	0.49

Table 42 MSE and expected costs for violent crime data, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Violent Crime	Outlier MSE	0.34	0.30	0.32	0.33	0.31	0.31
41 Outliers	Inlier MSE	0.26	0.26	0.26	0.26	0.26	0.25
1860 Inliers	E[cost]	0.34	0.30	0.31	0.33	0.31	0.31
Reduced ICA	Rand Out MSE	0.46	0.46	0.45	0.46	0.47	0.48
	Rand In MSE	0.26	0.26	0.26	0.26	0.26	0.25
	Rand cost	0.46	0.45	0.45	0.46	0.47	0.48

Table 43 MSE and expected costs for violent crime data, reduced ICA condition

A.4 NONVIOLENT CRIME

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Nonviolent Crime	Outlier MSE	0.25	0.24	0.25	0.27	0.32	0.32
48 Outliers	Inlier MSE	0.28	0.27	0.27	0.26	0.25	0.25
Scale	E[cost]	0.25	0.25	0.25	0.27	0.31	0.32
	Rand Out MSE	0.41	0.42	0.42	0.44	0.48	0.49
	Rand In MSE	0.28	0.28	0.28	0.27	0.26	0.25
	Rand cost	0.41	0.41	0.42	0.44	0.47	0.48

Table 44 MSE and expected costs for nonviolent crime data, control (scaled) data

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Nonviolent Crime	Outlier MSE	0.20	0.21	0.22	0.23	0.25	0.25
48 Outliers	Inlier MSE	0.31	0.31	0.30	0.30	0.28	0.28
ICA	E[cost]	0.20	0.22	0.23	0.23	0.25	0.25
	Rand Out MSE	0.34	0.36	0.38	0.38	0.41	0.41
	Rand In MSE	0.32	0.31	0.31	0.30	0.29	0.29
	Rand cost	0.34	0.36	0.38	0.38	0.40	0.40

Table 45 MSE and expected costs for nonviolent crime data, undreduced ICA condiotion

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Nonviolent Crime	Outlier MSE	0.24	0.24	0.27	0.32	0.32	0.33
48 Outliers	Inlier MSE	0.26	0.26	0.26	0.25	0.24	0.24
Reduced PCA	E[cost]	0.24	0.24	0.27	0.31	0.32	0.33
	Rand Out MSE	0.45	0.45	0.46	0.46	0.49	0.50
	Rand In MSE	0.26	0.26	0.26	0.26	0.25	0.24
	Rand cost	0.45	0.44	0.45	0.46	0.48	0.50

Table 46 MSE and expected costs for nonviolent crime data, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Nonviolent Crime	Outlier MSE	0.22	0.23	0.27	0.30	0.32	0.33
48 Outliers	Inlier MSE	0.26	0.26	0.26	0.26	0.26	0.25
Reduced ICA	E[cost]	0.22	0.23	0.27	0.30	0.32	0.32
	Rand Out MSE	0.45	0.46	0.45	0.45	0.47	0.49
	Rand In MSE	0.26	0.26	0.26	0.26	0.26	0.25
	Rand cost	0.44	0.45	0.44	0.45	0.47	0.48

Table 47 MSE and expected costs for nonviolent crime data, reduced ICA condition

A.5 TOTAL CRIME

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Total Crime	Outlier MSE	0.28	0.27	0.27	0.28	0.30	0.31
53 Outliers	Inlier MSE	0.28	0.27	0.27	0.26	0.25	0.25
Scale	E[cost]	0.28	0.27	0.27	0.27	0.30	0.31
	Rand Out MSE	0.42	0.42	0.41	0.44	0.48	0.48
	Rand In MSE	0.28	0.28	0.28	0.27	0.26	0.25
	Rand cost	0.41	0.42	0.41	0.44	0.47	0.47

Table 48 MSE and expected costs for total crime data, control (scaled) condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Total Crime	Outlier MSE	0.20	0.21	0.22	0.23	0.25	0.25
53 Outliers	Inlier MSE	0.31	0.31	0.30	0.30	0.28	0.28
ICA	E[cost]	0.21	0.22	0.23	0.23	0.25	0.25
	Rand Out MSE	0.34	0.36	0.38	0.38	0.41	0.41
	Rand In MSE	0.32	0.31	0.31	0.30	0.29	0.29
	Rand cost	0.34	0.36	0.38	0.38	0.40	0.40

Table 49 MSE and expected costs for total crime data, unreduced ICA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Total Crime	Outlier MSE	0.26	0.25	0.27	0.30	0.28	0.28
53 Outliers	Inlier MSE	0.26	0.26	0.26	0.25	0.24	0.24
Reduced PCA	E[cost]	0.26	0.25	0.27	0.30	0.28	0.28
	Rand Out MSE	0.45	0.45	0.45	0.47	0.49	0.50
	Rand In MSE	0.26	0.26	0.26	0.26	0.25	0.24
	Rand cost	0.44	0.45	0.44	0.47	0.49	0.50

Table 50 MSE and expected costs for total crime, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 190
Total Crime	Outlier MSE	0.27	0.25	0.28	0.29	0.28	0.29
53 Outliers	Inlier MSE	0.26	0.26	0.26	0.26	0.26	0.25
Reduced ICA	E[cost]	0.27	0.25	0.28	0.29	0.28	0.29
	Rand Out MSE	0.46	0.45	0.45	0.46	0.47	0.48
	Rand In MSE	0.26	0.26	0.26	0.26	0.26	0.25
	Rand cost	0.45	0.44	0.45	0.46	0.47	0.48

Table 51 MSE and expected costs for total crime, reduced ICA condition

A.6 CARDIOTOCOGRAPHY RESULTS

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 213
Cardiotocography	Outlier MSE	0.41	0.32	0.25	0.18	0.13	0.11
176 Outliers	Inlier MSE	0.27	0.27	0.27	0.25	0.23	0.22
1950 Inliers	E[cost]	0.40	0.32	0.25	0.19	0.14	0.12
Scale	Rand Out MSE	0.44	0.43	0.42	0.42	0.45	0.45
	Rand In MSE	0.27	0.28	0.28	0.27	0.26	0.25
	Rand cost	0.43	0.42	0.41	0.41	0.43	0.44

Table 52 MSE and expected costs for cardiotocography data, control (scaled) condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 213
Cardiotocography	Outlier MSE	0.33	0.27	0.23	0.22	0.18	0.16
176 Outliers	Inlier MSE	0.29	0.28	0.28	0.26	0.25	0.24
1950 Inliers	E[cost]	0.33	0.27	0.24	0.22	0.18	0.17
ICA	Rand Out MSE	0.40	0.38	0.37	0.40	0.41	0.42
	Rand In MSE	0.29	0.29	0.30	0.29	0.28	0.27
	Rand cost	0.39	0.37	0.37	0.39	0.40	0.41

Table 53 MSE and expected costs for cardiotocography data, unreduced ICA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 213
Cardiotocography	Outlier MSE	0.42	0.31	0.20	0.16	0.14	0.14
176 Outliers	Inlier MSE	0.21	0.20	0.19	0.18	0.16	0.15
1950 Inliers	E[cost]	0.41	0.30	0.20	0.16	0.14	0.14
Reduced PCA	Rand Out MSE	0.54	0.54	0.55	0.56	0.60	0.62
	Rand In MSE	0.22	0.22	0.22	0.21	0.20	0.19
	Rand cost	0.51	0.52	0.52	0.53	0.56	0.58

Table 54 MSE and expected costs for cardiotocography data, reduced PCA condition

		k = 10	k = 20	k = 40	k = 80	k = 160	k = 213
Cardiotocography	Outlier MSE	0.42	0.28	0.19	0.14	0.13	0.14
176 Outliers	Inlier MSE	0.21	0.20	0.18	0.18	0.16	0.15
1950 Inliers	E[cost]	0.41	0.27	0.19	0.15	0.14	0.14
Reduced ICA	Rand Out MSE	0.54	0.53	0.56	0.56	0.60	0.62
	Rand In MSE	0.22	0.22	0.21	0.21	0.20	0.19
	Rand cost	0.51	0.51	0.53	0.53	0.57	0.58

Table 55 MSE and expected costs for cardiotocography data, reduced ICA condition

BIBLIOGRAPHY

1. Laskowski RA. Structural Quality Assurance. In: Bourne PE, Weissig H, editors. Structural Bioinformatics. New Jersey: John Wiley; 2003. p 273-303.

2. Džeroski S, Langley P, Todorovski L. Computational discovery of scientific knowledge. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Volume 4660 LNAI; 2007. p 1-14.

3. Langley P. Lessons for the computational discovery of scientific knowledge. Proceedings of First International Workshop on Data Mining Lessons Learned 2002:9-12.

4. Langley P. The Computational Support of Scientic Discovery. Machine Learning and Its Applications; 2001. p 230-248.

5. Valdés-Pérez RE. Computer science research on scientific discovery. Knowledge Engineering Review 1996;11(1):57-66.

6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Databank. Nucleic Acids Research 2000;28:235-242.

7. Rupp B. Scientific inquiry, inference and critical reasoning in the macromolecular crystallography curriculum. Journal of Applied Crystallography 2010;43(5 PART 2):1242-1249.

8. Brändén CI, Jones TA. Between objectivity and subjectivity. Nature 1990;343:687-689.

9. Wimsatt WC. False Models as Means to Truer Theories. In: Nitecki MH, Hoffman A, editors. Neutral Models in Biology. New York: Oxford University Press; 1987. p 23-52.

10. Kleywegt GJ. Practical "Model Validation". 2006.

11. Kleywegt GJ. Validation of protein crystal structures. Acta Crystallographica 2000;D56:249-265.

12. Read Randy J, Adams Paul D, Arendall WB, Brunger Axel T, Emsley P, Joosten Robbie P, Kleywegt Gerard J, Krissinel Eugene B, Lütteke T, Otwinowski Z, Perrakis A, Richardson Jane S, Sheffler William H, Smith Janet L, Tickle Ian J, Vriend G, Zwart Peter H. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure (London, England : 1993) 2011;19(10):1395-1412.

13. UAB Statement on Protein Data Bank Issues. UAB Reporter 2009 12/08/2009.

14. Read RJ, Kleywegt GJ. Case-controlled structure validation. Acta Crystallographica Section D: Biological Crystallography 2009;65(2):140-147.

15. Kleywegt GJ. On vital aid: The why, what and how of validation. Acta Crystallographica Section D: Biological Crystallography 2009;65(2):134-139.

16. Hawkins PCD, Warren GL, Skillman AG, Nicholls A. How to do an evaluation: Pitfalls and traps. Journal of Computer-Aided Molecular Design 2008;22(3-4):179-190.

17. Kleywegt GJ, Jones TA. Homo Crystallographicus - Quo Vadis? Structure 2002;10:465-472.

18. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996;4:1395-1400.

19. Kleywegt GJ, Brünger AT. Checking your imagination: applications of the free R value. Structure 1996;4:897-904.

20. Kleywegt GJ, Jones TA. Where freedom is given, liberties are taken. Structure 1995;3:535-540.

21. Kleywegt GJ, Jones TA. Braille for Pugilists. In: Hunter WN, Thornton JM, Bailey S, editors. Making the Most of your Model. Daresbury, U.K.: SERC Daresbury Laboratory; 1995. p 11-24.

22. Dauter Z, Baker EN. Editorial: Black sheep among the flock of protein structures. Acta Crystallographica Section D: Biological Crystallography 2010;66(1):1.

23. Miller G. Scientific publishing. A scientist's nightmare: software problem leads to five retractions. Science 2006;314(5807):1856-1857.

24. Tickle IJ. Statistical quality indicators for electron-density maps. Acta Crystallographica Section D: Biological Crystallography 2012;68(4):454-467.

25. Jaskolski M. From Atomic Resolution to Molecular Giants: An Overview of Crystallographic Studies of Biological Macro-molecules with Synchrotron Radiation. Acta Physica Polonica-Series A General Physics 2010;117(2):257.

26. Dauter Z, Lamzin VS, Wilson KS. The benefits of atomic resolution. Current Opinion in Structural Biology 1997;7(5):681-688.

27. Weiss MS. Global indicators of X-ray data quality. Journal of Applied Crystallography 2001;34(2):130-135.

28. Evans P. Scaling and assessment of data quality. Acta Crystallographica Section D: Biological Crystallography 2006;62(1):72-82.

29. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Science 1993;2:1511-1519.

30. Headd JJ, Immormino RM, Keedy DA, Emsley P, Richardson DC, Richardson JS. Autofix for backward-fit sidechains: Using MolProbity and real-space refinement to put misfits in their place. Journal of Structural and Functional Genomics 2009;10(1):83-93.

31. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355-362.

32. Joosten RP, Joosten K, Murshudov GN, Perrakis A. PDB-REDO: Constructive validation, more than just looking for errors. Acta Crystallographica Section D: Biological Crystallography 2012;68(4):484-496.

33. Wlodawer A, Lubkowski J, Minor W, Jaskolski M. Is too creative language acceptable in crystallography? Acta Crystallographica Section D: Biological Crystallography 2010;66(9):1041-1042.

34. Ginzinger SW, Gruber M, Brandstetter H, Sippl MJ. Real space refinement of crystal structures with canonical distributions of electrons. Structure 2011;19(12):1739-1743.

35. Jeffrey P. X-ray Data Collection Course. 2010.

36. Schmidt A, Teeter M, Weckert E, Lamzin VS. Crystal structure of small protein crambin at 0.48 A resolution. Acta Crystallographica Section F 2011;67(4):424-428.

37. Sheldrick G. Phase annealing in SHELX-90: direct methods for larger structures. Acta Crystallographica Section A 1990;46(6):467-473.

38. Morris RJ, Bricogne G. Sheldrick's 1.2 A rule and beyond. Acta Crystallographica Section D 2003;59(3):615-617.

39. Urzhumtsev A, Afonine PV, Adams PD. On the use of logarithmic scales for analysis of diffraction data. Acta Crystallographica Section D: Biological Crystallography 2009;65(12):1283-1291.

40. Kantardjieff KA, Rupp B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. Protein Science 2003;12(9):1865-1871.

41. Carugo O, Bordo D. How many water molecules can be detected by protein crystallography? Acta Crystallographica Section D 1999;55(2):479-483.

42. Bacchi A, Lamzin VS, Wilson KS. A Self-Validation Technique for Protein Structure Refinement: the Extended Hamilton Test. Acta Crystallographica Section D 1996;52(4):641-646.

43. Hamilton W. Significance tests on the crystallographic R factor. Acta Crystallographica 1965;18(3):502-510.

44. Eastman P, Pellegrini M, Doniach S. Protein flexibility in solution and in crystals. Journal of Chemical Physics 1999;110(20):10141-10152.

45. Carugo O, Argos P. Reliability of atomic displacement parameters in protein crystal structures. Acta Crystallographica Section D: Biological Crystallography 1999;55(2):473-478.

46. Parthasarathy S, Murthy MRN. On the correlation between the main-chain and side-chain atomic displacement parameters (B values) in high-resolution protein structures. Acta Crystallographica Section D: Biological Crystallography 1999;55(1):173-180.

47. Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Engineering 1997;10(7):777-787.

48. Li DW, Brüschweiler R. All-atom contact model for understanding protein dynamics from crystallographic B-factors. Biophysical journal 2009;96(8):3074-3081.

49. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins: Structure, Function and Genetics 2005;61(1):115-126.

50. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins: Structure, Function and Bioformatics 2009;76(3):617-636.

51. Weiss MS. On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. Acta Crystallographica Section D: Biological Crystallography 2007;63(12):1235-1242.

52. Halle B. Flexibility and packing in proteins. Proceedings of the National Academy of Sciences of the United States of America 2002;99(3):1274-1279.

53. DePristo MA, De Bakker PIW, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 2004;12(5):831-838.

54. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallographica 1991;A47:392-400.

55. Brünger AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 1992;355(6359):472-475.

56. Tickle IJ, Laskowski RA, Moss DS. Rfree and the Rfree ratio II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. Acta Crystallographica 2000;D56:442-450.

57. Tickle IJ, Laskowski RA, Moss DS. Rfree and the Rfree ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Crystallographica 1998;D54:547-557.

58. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Advanced Protein Chemistry 1968;23:283-438.

59. Walther D, Cohen FE. Conformational attractors on the Ramachandran map. Acta Crystallographica Section D: Biological Crystallography 1999;55(2):506-517.

60. Anderson RJ, Weng Z, Campbell RK, Jiang X. Main-chain conformational tendencies of amino acids. Proteins: Structure, Function and Genetics 2005;60(4):679-689.

61. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys 2009;41(3).

62. Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics 1969;11(1):1-21.

63. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. Anomaly Detection: A Survey 2007.

64. Edgeworth FY. On discordant observations. Philosoph Mag 1887;23(5):364-375.

65. Wikipedia. There are known knowns. 2012.

66. Keyes R. The quote verifier: who said what, where, and when: Macmillan; 2007.

67. Jaskolski M, Gilski M, Dauter Z, Wlodawer A. Stereochemical restraints revisited: How accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallographica Section D: Biological Crystallography 2007;63(5):611-620.

68. MacArthur MW, Thornton JM. Deviations from planarity of the peptide bond in peptides and proteins. Journal of Molecular Biology 1996;264(5):1180-1195.

69. Zimek A, Schubert E, Kriegel H-P. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining 2012;5(5):363-387.

70. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Research 2000;28(1):254-256.

71. Chandonia JM, Hon G, Walker NS, Conte LL, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucleic Acids Research 2004;32(suppl 1):D189-D192.

72. Brown EN, Ramaswamy S. Quality of protein crystal structures. Acta Crystallographica Section D: Biological Crystallography 2007;63(9):941-950.

73. Wlodawer A, Lubkowski J, Minor W. Is too 'creative' language acceptable in crystallography? Acta Crystallographica Section D: Biological Crystallography 2010;66(9):1041-1042.

74. Wlodawer A, Minor W, Dauter Z, Jaskolski M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS Journal 2008;275(1):1-21.

75. Kleywegt GJ. Quality control and validation. In: Doublie, editor. Methods in Molecular Biology. Volume 364; 2007. p 255-272.

76. Borrell B. Fraud rocks protein community. Nature 2009;462(7276):970.

77. Baker EN, Dauter Z, Einspahr H, Weiss MS. In defence of our science - Validation now! Acta Crystallographica Section F: Structural Biology and Crystallization Communications 2010;66(2):112.

78. Baker EN, Dauter Z, Einspahr H, Weiss MS. In defence of our science - Validation now! Acta Crystallographica Section D: Biological Crystallography 2010;66(2):115.

79. Jones TA, Kleywegt GJ. Storing diffraction data [5]. Nature 1996;383(6595):18-19.

80. Vriend G. WHAT IF: A molecular modeling and drug design program. Journal of Molecular Graphics 1990;8:52-56.

81. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356(6364):83-85.

82. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992;12:345-364.

83. Brünger AT. Assessment of phase accuracy by cross validation: The free R value. Methods and Applications. Acta Crystallographica 1993;D49:24-36.

84. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 1993;26:283-291.

85. Laskowski RA, Moss DS, Thornton JM. Main-chain bond lengths and bond angles in protein structures. Journal of Molecular Biology 1993;231:1049-1067.

86. Vriend G, Sander C. Quality control of protein models: directional atomic contact analysis. Journal of Applied Crystallography 1993;26:47-60.

87. Dodson EJ, Kleywegt GJ. Report of a workshop on the use of statistical validators in protein X-ray crystallography. Acta Crystallographica 1996;D52:228-234.

88. Hooft RWW, Sander C, Vriend G. Verification of protein structures: side-chain planarity. Journal of Applied Crystallography 1996;29:714-716.

89. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. Journal of Molecular Biology 1996;264:121-136.

90. Brünger AT. The free R value: a more objective statistic for crystallography. Methods in Enzymology 1997;277:366-396.

91. Hooft RWW, Sander C, Vriend G. Objectively judging the quality of protein structure from a Ramachandran plot. Computer Applications in the Biosciences 1997;13:425-430.

92. Dodson E. The role of validation in macromolecular crystallography. Acta Crystallographica Section D: Biological Crystallography 1998;54(6):1109-1118.

93. Dodson EJ, Davies GJ, Lamzin VS, Murshudov GN, Wilson KS. Validation tools: Can they indicate the information content of macromolecular crystal structures? Structure 1998;6(6):685-690.

94. Laskowski RA, MacArthur MW, Thornton JM. Validation of protein models derived from experiment. Current Opinions in Structural Biology 1998;8(5):631-639.

95. Network EUDV. Who checks the checkers? four validation tools applied to eight atomic resolution structures. Journal of Molecular Biology 1998;276:417-436.

96. Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund AC, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa AL, Deleage G, Diarena M, Fabbretti R, Fettahi G, Flegel V, Gisel A, Kasam V, Kervinen T, Korpelainen E, Mattila K, Pagni M, Reichstadt M, Breton V, Tickle IJ, Vriend G. PDB-REDO: Automated re-refinement of X-ray structure models in the PDB. Journal of Applied Crystallography 2009;42(3):376-384.

97. Joosten RP, Joosten K, Cohen SX, Vriend G, Perrakis A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. Bioinformatics 2011;27(24):3392-3398.

98. Joosten RP, Womack T, Vriend G, Bricogne G. Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. Acta Crystallographica Section D: Biological Crystallography 2009;65(2):176-185.

99. Powell HR. Molecular structure from X-ray diffraction. Annual Reports on the Progress of Chemistry - Section C 2010;106:192-210.

100. Wang J, Wlodawer A, Dauter Z. What happens when the signs of anomalous differences or the handedness of substructure are inverted? Acta Crystallographica Section D 2007;63(7):751-758.

101. Matthews BW. Five retracted structure reports: Inverted or incorrect? Protein Sci 2007;16(6):1013-1016.

102. Dawson RJP, Locher KP. Structure of a bacterial multidrug ABC transporter. Nature 2006;443(7108):180-185.

103. Chang G, Roth CB. Structure of MsbA from E. coli: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters. Science 2001;293(5536):1793-1800.

104. Ward A, Reyes CL, Yu J, Roth CB, Chang G. Flexibility in the ABC transporter MsbA: Alternating access with a twist. Proceedings of the National Academy of Sciences of the United States of America 2007;104(48):19005-19010.

105. Tate CG. Comparison of three structures of the multidrug transporter EmrE. Current Opinion in Structural Biology 2006;16(4):457-464.

106. Reyes CL, Chang G. Structure of the ABC transporter MsbA in complex with ADP·vanadate and lipopolysaccharide. Science 2005;308(5724):1028-1031.

107. Pornillos O, Chen YJ, Chen AP, Chang G. Structural biology: X-ray structure of the EmrE multidrug transporter in complex with a substrate. Science 2005;310(5756):1950-1953.

108. Petsko GA. And the second shall be first. Genome Biology 2007;8(2).

109. Miller C. Pretty structures, but what about the data? [1]. Science 2007;315(5811):459.

110. Ma C, Chang G. Erratum: Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli (Proceedings of the National Academy of Sciences of the United States of America (March 2, 2004) 101, 9 (2852-2857) DOI: 10.1073/pnas. 0400137101)). Proceedings of the National Academy of Sciences of the United States of America 2007;104(9):3668.

111. Ma C, Chang G. Structure of the multidrug efflux transporter EmrE from Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 2004;101(9):2852-2857.

112. Joosten RP, Vriend G. PDB improvement starts with data deposition [3]. Science 2007;317(5835):195-196.

113. Jones TA, Kleywegt GJ. Experimental data for structure papers [2]. Science 2007;317(5835):194-195.

114. Chen YJ, Pornillos O, Lieu S, Ma C, Chen AP, Chang G. X-ray structure of EmrE supports dual topology model. Proceedings of the National Academy of Sciences of the United States of America 2007;104(48):18999-19004.

115. Chang G, Roth CB. Erratum (Retracted article): Structure of MsbA from E. coli: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters (Science (2001) 293 (1793)). Science 2006;314(5807):1875.

116. Chang G. Retraction of "Structure of MsbA from Vibrio cholera: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation" [J. Mol. Biol. (2003) 330 419-430] (DOI:10.1016/S0022-2836(03)00587-4). Journal of Molecular Biology 2007;369(2):596.

117. Chang G. Structure of MsbA from Vibrio cholera: A multidrug resistance ABC transporter homolog in a closed conformation. Journal of Molecular Biology 2003;330(2):419-430.

118. Jeffrey PD. Analysis of errors in the structure determination of MsbA. Acta Crystallographica Section D: Biological Crystallography 2009;65(2):193-199.

119. Reyes CL, Chang G. Structure of the ABC Transporter MsbA in Complex with ADP•Vanadate and Lipopolysaccharide. Science 2005;308(5724):1028-1031.

120. Pornillos O, Chen Y-J, Chen AP, Chang G. X-ray Structure of the EmrE Multidrug Transporter in Complex with a Substrate. Science 2005;310(5756):1950-1953.

121. Ma C, Chang G. Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli. PNAS 2004;101(9):2852-2857.

122. Chen Z, Blanc E, Chapman MS. Improved free R factors for cross-validation of macromolecular structure - importance for real-space refinement. Acta Crystallographica 1999;D55:219-224.

123. Chang G, Roth CB, Reyes CL, Pornillos O, Chen Y-J, Chen AP. Retraction. Science 2006;314(5807):1875b-.

124. Pellegrini M, Grønbech-Jensen N, Kelly JA, Pfluegl GMU, Yeates TO. Highly constrained multiple-copy refinement of protein crystal structures. Proteins: Structure, Function, and Bioinformatics 1997;29(4):426-432.

125. Chen Z, Chapman MS. Conformational Disorder of Proteins Assessed by Real-Space Molecular Dynamics Refinement. Biophysical journal 2001;80(3):1466-1472.

126. Fabiola F, Korostelev A, Chapman MS. Bias in cross-validated free R factors: Mitigation of the effects of non-crystallographic symmetry. Acta Crystallographica Section D: Biological Crystallography 2006;62(3):227-238.

127. Hanson MA, Stevens RC. Retraction: Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 A resolution. Nat Struct Mol Biol 2009;16(7):795-795.

128. Hanson MA, Stevens RC. Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 A resolution. Nat Struct Mol Biol 2000;7(8):687-692.

129. Janssen BJC, Read RJ, Brunger AT, Gros P. Crystallography: Crystallographic evidence for deviating C3b structure. Nature 2007;448(7154):E1-E2.

130. wwPDB Statement on Retraction of PDB Entries. 2009.

131. Weiss MS, Einspahr H, Baker T, Dauter Z. Another case of fraud in structural biology. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 2012;68(4):365.

132. Zaborsky N, Brunner M, Wallner M, Himly M, Karl T, Schwarzenbacher R, Ferreira F, Achatz G. Response to Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 2012;68(4):377.

133. Zaborsky N, Brunner M, Wallner M, Himly M, Karl T, Schwarzenbacher R, Ferreira F, Achatz G. Antigen aggregation decides the fate of the allergic immune response. Journal of Immunology 2010;184(2):725-735.

134. Rupp B. Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 2012;68(4):366-376.

135. Joosten RP, Te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G. A series of PDB related databases for everyday needs. Nucleic Acids Research 2011;39(SUPPL. 1):D411-D419.

136. Spek A. Structure validation in chemical crystallography. Acta Crystallographica Section D 2009;65(2):148-155.

137. Simon HA. Rational choice and the structure of the environment. Psychological review 1956;63(2):129.

138. Frank A, Asuncion A. UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences; 2010.

139. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems 2009;47(4):547-553.

140. U.S. Department of Justice FBoI. Crime in the United States. In: U.S. Department of Justice FBoI, editor: U.S. Department of Justice, Federal Bureau of Investigation; 1995.

141. U.S. Department of Justice BoJS. Law Enforcement Management And Administrative Statistics. In: U.S. Department Of Commerce BOTCP, editor. Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan: U.S. Department of Justice, Bureau of Justice Statistics; 1992.

142. U. S. Department of Commerce BotC. Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a. In: U. S. Department of Commerce BotCP, editor. Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan: U. S. Department of Commerce, Bureau of the Census; 1992.

143. Ayres-de-Campos D, Bernardes J, Garrido A, Marques-de-Sa J, Pereira-Leite L. SisPorto 2.0: a program for automated analysis of cardiotocograms. Journal of Maternal-Fetal and Neonatal Medicine 2000;9(5):311-318.

144. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Research 2007;35(suppl 1):D301-D303.

145. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nature Structural Biology 2003;10(12).

146. Tridgell A, Mackerras P. The rsync algorithm. 1996.

147. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Research 2005;33(suppl 1):D233-D237.

148. Eisenberg D, Lüthy R, Bowie J. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in Enzymology 1997;277:396.

149. Bowie J, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253(5016):164-170.

150. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 2007;35(Web Server issue).

151. Tosatto SCE, Battistutta R. TAP score: Torsion angle propensity normalization applied to local protein structure evaluation. BMC Bioinformatics 2007;8.

152. Chen VB, Arendall III WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: All-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D: Biological Crystallography 2010;66(1):12-21.

153. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D 2010;66(2):213-221.

154. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallographica 1999;D55:191-205.

155. Hooft RWW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381(6580):272-272.

156. Egli M. Diffraction techniques in structural biology. Current Protocols in Nucleic Acid Chemistry 2010(SUPPL. 41).

157. Brunger AT, Adams PD. Molecular dynamics applied to X-ray structure refinement. Accounts of Chemical Research 2002;35(6):404-412.

158. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D: Biological Crystallography 1997;53(3):240-255.

159. Adams PD, Pannu NS, Read RJ, Brünger AT. Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proceedings of the National Academy of Sciences of the United States of America 1997;94(10):5018-5023.

160. Breuniq MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. SIGMOD Record (ACM Special Interest Group on Management of Data) 2000;29(2):93-104.

161. Torgo L. Data Mining with R: Learning with Case Studies. 2010.

162. Kriegel H-P, Kröger P, Schubert E, Zimek A. Interpreting and Unifying Outlier Scores. 2011 April 28-30, 2011; Mesa, Arizona. SIAM / Omnipress.

163. Delignette-Muller ML, Pouillot R, Denis J-B, Dutang C. Fitdistrplus: help to fit of a parametric distribution to non-censored or censored data. R package version 0.1-3, URL http://CRAN. R-project. org/package= fitdistrplus; 2010.

164. Crooks GE. The Amoroso Distribution. arXiv preprint arXiv:10053274 2010.

165. Stacy E. A generalization of the gamma distribution. The Annals of Mathematical Statistics 1962;33(3):1187-1192.

166. Amoroso L. Ricerche intorno alla curva dei redditi. Annali di matematica pura ed applicata 1925;2(1):123-159.

167. Cullen AC, Frey HC. Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs: Springer; 1999.

168. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine 1996;17(3):37-53.

169. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis for the analysis of data. Biometrika 1968;55(1):1-17.

170. Jones MC. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. Statistical Methodology 2009;6(1):70-81.

171. Kumaraswamy P. A generalized probability density function for double-bounded random processes. Journal of Hydrology 1980;46(1–2):79-88.

172. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Machine Learning 2004;31:1-38.

173. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence: Springer; 2006. p 1015-1021.

174. Landgrebe TCW, Paclik P, Duin RPW, Bradley AP. Precision-recall operating characteristic (P-ROC) curves in imprecise environments. 2006 0-0 0. p 123-127.

175. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997;30(7):1145-1159.

176. Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. 2006.

177. Schubert E, Wojdanowski R, Zimek A, Kriegel HP. On evaluation of outlier rankings and outlier scores. 2012. p 1047-1058.

178. North DW. A Tutorial Introduction to Decision Theory. Systems Science and Cybernetics, IEEE Transactions on 1968;4(3):200-210.

179. Keller F, Muller E, Bohm K. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. 2012 1-5 April 2012. p 1037-1048.

180. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural networks 2000;13(4):411-430.

181. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Springer; 2001.

182. Kabán A. On the distance concentration awareness of certain data reduction techniques. Pattern Recognition 2011;44(2):265-277.

183. Team RC. R: A language and environment for statistical computing. R Foundation Statistical Computing 2013.

184. Jackson DA. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 1993:2204-2214.

185. Peres-Neto PR, Jackson DA, Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis 2005;49(4):974-997.

186. MacArthur RH, MacArthur JW. On bird species diversity. Ecology 1961;42(3):594-598.

187. MacArthur RH. On the relative abundance of bird species. Proceedings of the National Academy of Sciences of the United States of America 1957;43(3):293.

188. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biology Direct 2007;2.

189. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on 1999;10(3):626-634.

190. Marchini J, Heaton C, Ripley B. fastICA: FastICA algorithms to perform ICA and Projection Pursuit. R package 2013.

191. Chavez E, Navarro G, Baeza-Yates R, Marroquin JL. Searching in metric spaces. ACM Comput Surv 2001;33(3):273-321.

192. Pagel B-U, Korn F, Faloutsos C. Deflating the Dimensionality Curse Using Multiple Fractal Dimensions. Proceedings of the 16th International Conference on Data Engineering: IEEE Computer Society; 2000. p 589.

193. Murtagh F. On ultrametricity, data coding, and computation. Journal of Classification 2004;21(2):167-184.

194. Murtagh F, Downs G, Contreras P. Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding. SIAM Journal on Scientific Computing 2008;30(2):707-730.

195. Murtagh F. The remarkable simplicity of very high dimensional data: application of model-based clustering. Journal of classification 2009;26(3):249-277.

196. Murtagh F. Thinking ultrametrically. Classification, Clustering, and Data Mining Applications 2004:3-14.

197. Murtagh F. Symmetry in data mining and analysis: a unifying view based on hierarchy. Proceedings of the Steklov Institute of Mathematics 2009;265(1):177-198.

198. Murtagh F. From data to the p-adic or ultrametric model. P-Adic Numbers, Ultrametric Analysis, and Applications 2009;1(1):58-68.

199. Murtagh F. Identifying and exploiting ultrametricity. Advances in Data Analysis 2007:263-272.

200. Murtagh F. Quantifying ultrametricity. 2004. p 1561-1568.

201. Pestov V. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? arXiv preprint arXiv:11104347 2011.

202. Pestov V. Intrinsic dimensionality. arXiv preprint arXiv:10075318 2010.

203. Pestov V. An axiomatic approach to intrinsic dimension of a dataset. Neural Networks 2008;21(2):204-213.

204. Pestov V. Intrinsic dimension of a dataset: what properties does one expect? ; 2007. IEEE. p 2959-2964.

205. Kriegel HP, Kröger P, Renz M, Schubert M. Metric spaces in data mining: applications to clustering. SIGSPATIAL Special 2010;2(2):36-39.

206. Wong W, Moore A, Cooper G, Wagner M. Bayesian network anomaly pattern detection for disease outbreaks. 2003. p 808.

207. Song X, Wu M, Jermaine C, Ranka S. Conditional anomaly detection. Knowledge and Data Engineering, IEEE Transactions on 2007;19(5):631-645.

208. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Šali A, Studier FW, Swaminathan S. Structural genomics: Beyond the Human Genome Project. Nature Genetics 1999;23(2):151-157.

209. Terwilliger TC, Berendzen J. Exploring structure space: A protein structure initiative. Genetica 1999;106(1-2):141-147.

210. Domingues FS, Koppensteiner WA, Sippl MJ. The role of protein structure in genomics. FEBS Letters 2000;476(1-2):98-102.

211. Smaglik P. Genomics initiative to decipher 10,000 protein structures. Nature 2000;407(6804):549.

212. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. Journal of molecular biology 2005;348(5):1235-1260.

213. Petsko GA. An idea whose time has gone. Genome Biology 2007;8(6).

214. Banci L, Baumeister W, Heinemann U, Schneider G, Silman I, Stuart DI, Sussman JL. An idea whose time has come. Genome Biology 2007;8(11).

215. Moore PB. Let's Call the Whole Thing Off: Some Thoughts on the Protein Structure Initiative. Structure 2007;15(11):1350-1352.

216. Harrison SC. Comments on the NIGMS PSI. Structure 2007;15(11):1344-1346.

217. NIGMS Invites Biologists to Join High-Throughput Structure Initiative. National Institute of General Medical Sciences; 2009.

218. Gao J, Tan P-N. Converting output scores from outlier detection algorithms into probability estimates. 2006. IEEE. p 212-221.

219. Radovanović M, Nanopoulos A, Ivanović M. Hubs in space: Popular nearest neighbors in high-dimensional data. The Journal of Machine Learning Research 2010;9999:2487-2531.

220. Milo, #353, Radovanovi, #263, Nanopoulos A, Ivanovi M, #263. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: ACM; 2009. p 865-872.

221. François D, Wertz V, Verieysen M. The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering 2007;19(7):873-886.

222. Tomašev N, Radovanović M, Mladenić D, Ivanović M. The Role of Hubness in Clustering High-Dimensional Data. In: Huang J, Cao L, Srivastava J, editors. Advances in Knowledge Discovery and Data Mining. Volume 6634, Lecture Notes in Computer Science: Springer Berlin Heidelberg; 2011. p 183-195.

223. Fung G, Sandilya S, Rao RB. Rule extraction from linear support vector machines. 2005. ACM. p 32-40.

224. Setiono R, Liu H. Understanding neural networks via rule extraction. 1995. Citeseer. p 480-487.

225. Andrews R, Diederich J, Tickle AB. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems 1995;8(6):373-389.

226. Barakat N, Diederich J. Learning-based rule-extraction from support vector machines. 2004. not found.

227. Barakat N, Diederich J. Eclectic rule-extraction from support vector machines. International Journal of Computational Intelligence 2005;2(1):59-62.

228. Duch W, Setiono R, Zurada JM. Computational intelligence methods for rule-based data understanding. Proceedings of the IEEE 2004;92(5):771-805.

229. Davis R, Buchanan B, Shortliffe E. Production rules as a representation for a knowledge-based consultation program. Artificial intelligence 1977;8(1):15-45.

230. Freitas AA, Wieser DC, Apweiler R. On the importance of comprehensible classification models for protein function prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2010;7(1):172-182.

231. Yamanishi K, Takeuchi J-i. Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. 2001. ACM. p 389-394.

232. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11(1):10-18.

233. Quinlan JR. C4. 5: programs for machine learning: Morgan kaufmann; 1993.

234. Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. Bioinformatics 2010;26(5):668-675.

235. Provost F, Aronis J, Buchanan BG. Rule-space search for knowledge-based discovery. New York: Stern School of Business, New York University; 1999. Report nr IS 99-012.

236. Aronis JM, Provost FJ, Buchanan BG. Exploiting Background Knowledge in Automated Discovery. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining 1999.

237. Lee Y, Buchanan BG, Aronis JM. Knowledge-Based Learning in Exploratory Science: Learning Rules to Predict Rodent Carcinogenicity. Machine Learning 1998;30(2):217-240.

238. Hennessy DN, Gopalakrishnan V, Buchanan BG, Rosenberg JM, Subramanian D. Induction of rules for biological macromolecule crystallization. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 1994;2:179-187.

239. Clearwater SH, Provost FJ. RL4: A tool for knowledge-based induction. 1990. p 24-30.

240. Lustgarten J, Visweswaran S, Grover H, Gopalakrishnan V. An evaluation of discretization methods for learning rules from biomedical datasets. 2008.

241. Kononenko I. On biases in estimating multi-valued attributes. 1995. LAWRENCE ERLBAUM ASSOCIATES LTD. p 1034-1040.

242. Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

243. Eisenberg D, Bowie JU, Luthy R, Choe S. Three-dimensional profiles for analysing protein sequence-structure relationships. Faraday Discussions 1992;93:25-34.

244. Ishida T, Nakamura S, Shimizu K. Potential for assessing quality of protein structure based on contact number prediction. Proteins: Structure, Function and Genetics 2006;64(4):940-947.

245. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins: Structure, Function and Genetics 2000;40(3):389-408.

246. Will Sheffler DB. RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Science 2009;18(1):229-239.

247. Flip K. On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'. IEEE Transactions on Knowledge and Data Engineering 2001;13(1):96-111.

248. de Sousa EPM, Traina C, Traina A, Faloutsos C. How to use fractal dimension to find correlations between attributes. 2002.

249. Malcok M, Aslandogan YA, Yesildirek A. Fractal dimension and similarity search in high-dimensional spatial databases. 2006. IEEE. p 380-384.

250. Murtagh F. From data to the physics using ultrametrics: new results in high dimensional data analysis. 2006. p 151.

251. Berman HM, Kleywegt GJ, Nakamura H, Markley JL, Burley SK. Safeguarding the integrity of protein archive. Nature 2010;463(7280):425-425.

252. Scholz M, Gibon Y, Stitt M, Selbig J. Independent component analysis of starch deficient pgm mutants. 2004. GI, Bielefeld, Germany. p 95-104.

253. Aggarwal CC. On the effects of dimensionality reduction on high dimensional similarity search. 2001. ACM. p 256-266.

254. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.

255. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 1995;247(4):536-540.

256. Gilbert WS, Sullivan A. The Mikado; or, The Town ofTitipu. The Complete Plays of Gilbert and Sullivan 1885:345-400.

257. Wong ML, Leung KS. Data Mining Using Grammar-Based Genetic Programming and Applications: Kluwer Academic Publishers; 2000. 213 p.

258. Clare A, King RD. Machine learning of functional class from phenotype data. Bioinformatics 2002;18(1):160-166.

259. Valdés-Pérez RE. Principles of human-computer collaboration for knowledge discovery in science. Artif Intell 1999;107(2):335-346.