

# The Geometry and Physicochemistry of Protein Binding Sites and Ligands and their Detection in Electron Density Maps



Abdullah Kahraman

European Bioinformatics Institute

Clare College

and

Faculty of Biology

University of Cambridge

A dissertation submitted to the Faculty of Biology at the University of  
Cambridge for the degree of Doctor of Philosophy

6<sup>th</sup> February 2009

# Preface

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the 300 page limit for the Degree Committee of the Faculty of Biology.

# Abstract

Most protein function prediction methods identify small ligand molecules for protein structures by applying the principles of molecular complementarity. They assume that the ligand has complementary geometrical and physicochemical properties to the binding site and that similar binding sites bind similar ligands. Here, I present a systematic analysis and comparison on the degree of complementarity between protein binding sites and their ligands. For this purpose, a new data set was compiled comprising various sets of non-homologous binding sites that each binds the same ligand. Using the data set it was discovered that binding sites have a greater variation in their shapes than can be accounted for by the conformational variability of their ligand. Separating shape from size information revealed that a significant proportion of the recognition power of a binding site for its ligand resides in its shape. It could be shown that the large variation in size and shape was caused by a “buffer zone”, which is a region of free space between the protein and the ligand. The buffer zone causes binding sites to be two to three times larger in volume than the ligand that they bind. A similar analysis on the physicochemical properties demonstrated an even larger variation for the physicochemical properties within binding sites that bind the same ligand. The variation was often to such an extent that only a qualitative similarity remained. Nevertheless, the comparison between the hydrophobicity and the electrostatic potential in binding sites showed that the former varies less and that the latter is highly influenced by neighbouring chemical compounds and the dielectric constant. An attempt to correlate the computed properties with experimental observations gave only modest results. Overall, the results in this thesis are suggesting that geometrical complementarity is in general not sufficient to drive molecular recognition. The protein rather engages in a subtle balancing act between electrostatic and hydrophobic interactions in order to not bind the ligand too strong and disrupt its own biochemical function. In some protein-ligand complexes hydrophobic interactions were

observed to override repulsive electrostatic interactions, while in others repulsive interactions were disclosed as prerequisites for the biochemical function of the protein. Finally, as a proof of principle, I present a method to automatically predict ligands for protein structures not based on binding site characteristics but rather on the electron density data produced in X-ray crystallography experiments. The method although performing similar to the well-established ligand-fitting module in the protein modelling software ARP/warp, is superior in terms of speed and accuracy allowing it to be effortlessly integrated in various electron density screening scenarios. In summary, this thesis highlights the complexities of molecular recognition and underlines the challenges in computational structural biology to develop methods for the identification of intermolecular interactions for drugs design and *in silico* simulation of living cells.

# Acknowledgement

This work would have not been possible without the support, help and advice of so many people with whom I had the pleasure to work, collaborate and form friendships. First and foremost, I would like to thank my primary supervisor Janet Thornton for accepting me as her student and guiding me through the last four years of my life. Her constant support and encouragement kept me going at times where the burden of a PhD seemed unbearable. I also would like to thank my secondary 'unofficial' supervisor and desk mate Roman Laskowski, who took up the father role in my PhD and lead me with his torch of knowledge through the darkness of science. Furthermore, thanks to my TAC members Nick Luscombe and Robert Glen for advising me on the various stages of my PhD and my second roommate James Watson in particular for his valuable comments on so many drafts of manuscripts, abstracts and this thesis.

Also thanks to Rafi Najmanovich, Angelo Favia and Asad Rahman for scientific discussions on various topics and Fabian Gerick for testing CleftXplorer. Victor Lamzin, Gerrit Langer and Helene Doerksen deserve a special acknowledgement for collaborating on Chapter 5 and offering me their hospitality for three weeks in Hamburg, as do Kazuto Yamazaki, Thomas Funkhouser, Carla Mattos and Jim Warwicker for their collaborations.

And I shall not forget the past and present group members of the Thornton (spice) group, each of them making the last four years unforgettable. In alphabetical order: Dan Andrews, Jonathan Barker, Matthew Bashton, Eric Blanc, Victor Chiskoff, Julia Fischer, Shiri Freilich, Nick Furnham, Fabian Glaser, Alex Gutteridge, Gemma Holliday, Nicola Kerrison, Tim Massingham, Pilar Miguel-Ortega, Irilenia Nobeli, Marialuisa Pellegrini-Calace, Hannes Ponstingl, Gabby Reeves, Eugene Schuster, Hugh Shanahan, Mike Stevens, Gareth

Stockwell, David Talavera, James Torrance, Jonathan Ward, Daniela Wieser and our secretaries Gillian Adams, Helen Barker and Stacy Schab. Thanks also to the trainees and visitor students for their friendships: Claudia Andreini, Lorenzo Baldacci, Gabriele Cavallaro, Martin Grana, Tim Maiwald, Marian Novotny, Romina Oliva, Ferdinando Spagnolo and Noriko Hiroi.

A special recognition goes to Richard Morris. The extent to how much I am grateful to him cannot be expressed in few lines. You were the source of my inspirations. Most of what I know and what I have achieved, I owe to you.

It is self-evident that without the everlasting support of my parents and my family, I would not be here, where I am now. Thanks and a big hug mum and dad. And finally, but most importantly a huge thanks from deep of my heart to my wonderful wife, who has sacrificed so much of her life to make my dream come true.

For my wife and our sunshine Ammar.

# Contents

<b>Preface</b> .....	<b>i</b>
<b>Abstract</b> .....	<b>ii</b>
<b>Acknowledgement</b> .....	<b>iv</b>
<b>Contents</b> .....	<b>vi</b>
<b>List of Figures</b> .....	<b>xi</b>
<b>List of Tables</b> .....	<b>xiii</b>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
<b>1.1 Overview of thesis</b> .....	<b>3</b>
<b>Chapter 2 Background</b> .....	<b>6</b>
<b>2.1 Protein structures</b> .....	<b>6</b>
2.1.1 The Protein Data Bank.....	6
2.1.1.1 The worldwide Protein Data Bank .....	7
2.1.2 Biological relevant protein conformation .....	8
2.1.3 Cognate ligands .....	9
<b>2.2 Protein structure determination</b> .....	<b>10</b>
2.2.1 Macromolecular X-ray crystallography .....	11
2.2.1.1 Interpretation of diffraction data .....	13
2.2.1.1.1 Bragg's law.....	13
2.2.1.1.2 Resolution .....	14
2.2.1.1.3 Structure factors .....	14
2.2.1.1.4 Phase problem .....	16
2.2.1.1.5 Obtaining initial phases .....	16
2.2.1.2 Refinement of protein model and phases .....	18

2.2.1.2.1	Atomic disorder in refinement .....	19
2.2.1.2.2	Calculating structure factors.....	20
2.2.1.2.3	Judging the protein model.....	20
2.2.1.2.4	Refinement procedures.....	21
2.2.1.3	Automated protein modelling.....	23
<b>2.3</b>	<b>Enzymes .....</b>	<b>25</b>
2.3.1	Enzyme binding sites .....	26
2.3.1.1	Active site .....	27
2.3.1.2	Cofactor/coenzyme binding site.....	28
2.3.1.3	Allosteric sites.....	28
2.3.2	Characteristics of enzyme binding sites.....	29
2.3.2.1	Volume.....	31
2.3.2.2	Depth .....	32
2.3.2.3	Shape .....	33
2.3.2.4	Flexibility .....	33
2.3.2.5	Conservation.....	35
2.3.2.6	Interaction energy.....	36
<b>2.4</b>	<b>Intermolecular forces .....</b>	<b>37</b>
2.4.1	Electrostatics <i>in vacuo</i> .....	38
2.4.1.1	Electrostatic potential.....	39
2.4.1.2	Poisson equation .....	40
2.4.2	Electrostatics in a dielectric medium .....	40
2.4.2.1	Protein electrostatics in aqueous media .....	41
2.4.2.2	Protein electrostatics in ionized aqueous media.....	42
2.4.2.2.1	Poisson-Boltzmann equation .....	42
2.4.3	Water and the hydrophobic effect .....	43
2.4.3.1	Hydrogen bond .....	44
2.4.3.2	Hydrophobic effect.....	45
2.4.3.3	Partition coefficient $\log P$ .....	45



2.4.3.4	Hydrophobic interaction .....	46
<b>2.5</b>	<b>Surface and shape of molecules.....</b>	<b>46</b>
2.5.1	Molecular surface description.....	47
2.5.2	Molecular surface comparison .....	49
2.5.2.1	Graph matching .....	50
2.5.2.2	Geometric hashing.....	50
2.5.2.3	Spherical harmonics .....	51
2.5.2.3.1	Associated Legendre polynomials .....	54
2.5.2.3.2	Spherical harmonics expansion .....	55
2.5.2.3.3	Spherical <i>t</i> -design .....	57
<b>Chapter 3</b>	<b>Shape Variation in Protein Binding Pockets and their Ligands.....</b>	<b>58</b>
<b>3.1</b>	<b>Introduction.....</b>	<b>58</b>
<b>3.2</b>	<b>Methods .....</b>	<b>62</b>
3.2.1	Cleft reduction .....	67
3.2.1.1	Conserved cleft model .....	67
3.2.1.2	Interact cleft model .....	68
3.2.1.3	Ligand cleft model.....	68
3.2.2	Classification and data analysis .....	69
<b>3.3</b>	<b>Data set.....</b>	<b>70</b>
<b>3.4</b>	<b>Results and discussion.....</b>	<b>72</b>
3.4.1	Shape reproduction quality and comparison metric .....	72
3.4.1.1	Reconstruction error .....	72
3.4.1.2	Comparison to surface RMSD .....	75
3.4.2	Shape Variation.....	76
3.4.2.1	Ligand conformations .....	76
3.4.2.2	Binding pocket shape diversity in ligand sets .....	80
3.4.2.3	Binding pocket shape vs. ligand shape .....	82
3.4.2.4	Shape vs. size .....	85
3.4.2.5	Performance of cleft models .....	86

3.4.3	Limitations and problems .....	88	
3.4.3.1	Binding pocket prediction.....	88	
3.4.3.2	Partially bound ligands.....	89	
3.4.3.3	Star-like shapes and rotational variance.....	90	
3.4.3.4	Single property descriptor.....	90	
<b>3.5</b>	<b>Conclusions .....</b>	<b>90</b>	
<b>Chapter 4 On the Diversity of Physicochemical Environments Experienced by</b>			
<b>Identical Ligands in Binding Pockets of Unrelated Proteins .....</b>			<b>93</b>
<b>4.1</b>	<b>Introduction.....</b>	<b>93</b>	
<b>4.2</b>	<b>Methods .....</b>	<b>96</b>	
4.2.1	Calculating protein physicochemical properties on ligand molecules .....	96	
4.2.2	Electrostatic potential .....	97	
4.2.3	Scoring the hydrophobic environment.....	98	
4.2.4	Other physicochemical properties .....	100	
4.2.5	Average properties and their variation .....	101	
4.2.6	Data set.....	102	
<b>4.3</b>	<b>Results .....</b>	<b>102</b>	
4.3.1	Factors affecting the electrostatic potential .....	102	
4.3.2	Physicochemical properties of proteins in ligand binding sites .....	106	
4.3.2.1	Electrostatic potential on ligands .....	106	
4.3.2.2	Adenosine-5'-triphosphate (ATP) .....	106	
4.3.2.3	Nicotinamide adenine dinucleotide (NAD) .....	109	
4.3.2.4	Heme type B.....	114	
4.3.2.5	Remaining ligand sets .....	116	
4.3.3	Non-electrostatic interactions between protein and ligand.....	118	
4.3.4	Average and variation of physicochemical properties .....	121	
4.3.5	Comparison of properties between ligand sets .....	125	
<b>4.4</b>	<b>Discussion.....</b>	<b>130</b>	
<b>4.5</b>	<b>Conclusion .....</b>	<b>135</b>	

<b>Chapter 5 Automated Ligand Recognition in Electron Density Maps .....</b>	<b>136</b>
<b>5.1 Introduction .....</b>	<b>136</b>
<b>5.2 Methods .....</b>	<b>141</b>
5.2.1 Algorithm summary .....	141
5.2.2 Performance measure .....	142
<b>5.3 Data set .....</b>	<b>142</b>
<b>5.4 Results .....</b>	<b>144</b>
5.4.1 Parameter assessment .....	144
5.4.2 Ligand recognition with spherical harmonics .....	146
5.4.3 Spherical harmonics vs. geometric features .....	149
5.4.3.1 Performance comparison .....	150
5.4.3.2 Coefficient distance vs. geometric feature similarity .....	154
<b>5.5 Discussion .....</b>	<b>155</b>
<b>5.6 Conclusion .....</b>	<b>159</b>
<b>Chapter 6 Final Remarks .....</b>	<b>160</b>
6.1.1 Caveats .....	160
6.1.2 Future developments .....	162
6.1.3 Function prediction .....	163
6.1.4 Final conclusion .....	164
<b>Appendix A .....</b>	<b>167</b>
<b>Data set I .....</b>	<b>167</b>
<b>Appendix B .....</b>	<b>174</b>
<b>Data set II .....</b>	<b>174</b>
<b>Data set III .....</b>	<b>175</b>
<b>References .....</b>	<b>188</b>

# List of Figures

Figure 2.1: ASU vs. biologically relevant conformation.....	9
Figure 2.2: Binding Site of <i>Escherichia coli</i> asparagine synthetase. ....	27
Figure 2.3: Characteristics of enzyme binding sites. ....	30
Figure 2.4: SURFNET spheres filling a cleft on the protein surface. ....	31
Figure 2.5: Depth of a protein surface calculated by Travel Depth.....	32
Figure 2.6: Conservation scores mapped on a protein structure by ConSurf. ....	35
Figure 3.1: CleftXplorer algorithm for binding pocket shape description. ....	63
Figure 3.2: Reconstructed shape of cleft models from different binding sites. ....	66
Figure 3.3: Shape reconstruction with spherical harmonics. ....	73
Figure 3.4: Error while shape reconstruction with spherical harmonics.....	74
Figure 3.5: Coefficient distance correlation to surface RMSD. ....	75
Figure 3.6: All-against-all coefficient distance matrices for Data set I. ....	78
Figure 3.7: Shape variation of protein binding pockets and ligands. ....	81
Figure 3.8: Histograms comparing binding site and ligand coefficient distances. ....	83
Figure 3.9: Histogram of first true hits in coefficient distance calculations.....	84
Figure 3.10: Buffer-zone and water molecules in binding sites. ....	85
Figure 3.11: Partially occupied binding pockets.....	89
Figure 4.1: Scatter plot of $\Delta G^{\text{logP}}$ versus $\Delta G^{\text{obs}}$ .....	99
Figure 4.2: Influence of NCCs on protein's electrostatic potentials .....	103
Figure 4.3: Influence of dielectric constant on protein's electrostatic potentials. ....	105
Figure 4.4: Electrostatic potentials experienced by ATP molecules. ....	107
Figure 4.5: Influence of metal ions on ATP's experienced electrostatic potential.....	108
Figure 4.6: Functional necessary repulsive forces in cytochrome b5 reductase 1ib0.....	110
Figure 4.7: Electrostatic potentials experienced by NAD molecules.....	112

Figure 4.8: Aromatic interactions compensate repulsive electrostatic interactions.....	113
Figure 4.9: Electrostatic potentials experienced by heme molecules. ....	115
Figure 4.10: Electrostatic potentials experienced by remaining ligand sets. ....	117
Figure 4.11: Hydrophobicity experienced by ATP, NAD, heme molecules. ....	120
Figure 4.12: Hydrophobicity experienced by remaining ligand sets.....	121
Figure 4.13: Average electrostatic potentials for Data set I. ....	122
Figure 4.14: Average hydrophobicity scores for Data set I.....	123
Figure 4.15: Variation of physicochemical properties in protein binding pockets. ....	124
Figure 4.16: Spherical harmonics reconstruction of electrostatic potential distribution. ....	127
Figure 5.1: Fragmentation tree filtering of electron density blobs .....	138
Figure 5.2: Electron density of an ATP. ....	140
Figure 5.3: Electron density at various contour thresholds. ....	144
Figure 5.4: Spherical harmonics performance at changing parameters. ....	145
Figure 5.5: Spherical harmonics performance on Data set III. ....	146
Figure 5.6: Limitations of spherical harmonics: ligand flexibility. ....	147
Figure 5.7: Limitations of spherical harmonics: poor phases.....	147
Figure 5.8: Limitations of spherical harmonics: non star-like shapes. ....	148
Figure 5.9: Limitations of spherical harmonics: false ligand conformation.....	148
Figure 5.10: Scatter plot of spherical harmonics vs. geometric features performances. ....	149
Figure 5.11: Good spherical harmonics vs. good geometric features performance. ....	150
Figure 5.12: Bad spherical harmonics vs. good geometric features performance.....	151
Figure 5.13: Bad spherical harmonics vs. bad geometric features performance.....	152
Figure 5.14: Good spherical harmonics vs. bad geometric features performance. ....	153
Figure 5.15: Scatter plot of spherical harmonics vs. geometric feature scores. ....	154
Figure 5.16: Histogram on complementary ranks between both shape representations.....	155

# List of Tables

Table 2.1: The extent of enzyme data in some structural databases. ....	26
Table 3.1: Statistics on coefficient distances. ....	77
Table 3.2: AUCs for various classification approaches on Data set I. ....	79
Table 3.3: Statistics on the volume of Interact cleft models. ....	85
Table 4.1: Average and standard deviation of physicochemical properties. ....	106
Table 4.2: AUCs for geometrical and physicochemical properties. ....	129
Table 4.3: Binding free energies calculated for Data set I. ....	134
Table A.1: Data set of 100 binding sites being non-homologous in 9 ligand sets. ....	167
Table B.1: Small data set of 12 ligand molecules within difference electron density maps. .	174
Table B.2: Large data set of 536 ligand molecules within difference electron density maps.	175

# Chapter 1

## Introduction

*Proteins* are of utmost importance to living cells. More than half of the dry weight of human cells is made up of proteins. They participate in almost all cellular processes, to enumerate a few: as enzymes proteins drive catalytic reactions, whilst as cell membrane bound receptors they are involved in signal transduction. Transmembrane proteins transport actively or passively molecules through the membrane barrier and immunoglobins recognize infectious organisms and induce an immune response. Proteins also make up the DNA transcription machines that are directly involved in gene expression (Alberts, *et al.*, 1994b). Without any doubt, life without proteins would not be possible or at least not in its current form.

Proteins do not exist on their own. Due to the crowded nature of the interior of a cell (Ellis, 2001), they are constantly in contact with other molecules. However, to perform any of the functions listed above, they must specifically recognize their interaction partners, in general called *ligands*, and form a molecular complex. The recognition process is mediated by a distinct region on the surface of the protein, which is referred to as the *binding site*, and which forms the trigger to set the protein into action (Bergner and Günther, 2004). Hence, there is an intimate link between the structure of a protein and its function.

In 1960, the Austrian molecular biologist Max Ferdinand Perutz published the first high resolution structure of a protein, namely horse haemoglobin (Perutz, *et al.*, 1960) using X-ray crystallography here at the University of Cambridge. This publication launched a new era for the investigation of protein structures, molecular complexes and their functions (Fersht, 1984). For the first time it was possible to observe the interaction between proteins and ligands in atomic detail, analyse the identity and the location of ligands in a protein structure

and assess the intermolecular forces and features that drive molecular recognition. However, the high-resolution structure determination technique of X-ray crystallography has limitations for the analysis of protein function. For example, time-varying molecular processes such as enzymatic reactions or the entire process of recognising, binding and releasing ligand molecules cannot be traced in detail. X-ray crystallography can only provide a single snapshot of the protein's life and of the molecular processes that take place in proteins. The catalytic machinery of enzymes, for example, must be blocked with substrate/cofactor analogues or inhibitors, if substrate or other functionally important ligands are to be found in the protein structure. This results in many structures, which do not have the biologically or functionally relevant ligands bound. Nevertheless X-ray crystallography has helped scientists to understand how proteins recognize other molecules, how they perform specific catalytic reactions or whether proteins act alone or in concert with other proteins or other molecules (Laskowski, 2003).

With the launch of the worldwide *structural genomic initiative* at the start of this century (Blundell and Mizuguchi, 2000), new problems arose from the functional assessment of proteins. Traditional protein experiments seek to solve the structure of a protein in order to understand the molecular mechanism of the protein's function. In contrast, structural genomic projects (Baker and Sali, 2001; Brenner, 2001; Chandonia and Brenner, 2006) aim to obtain a structural model to all genomic protein sequences of an organism, while reducing the average cost and time of the structure determination process through the development and improvement of automation and high-throughput technology (Chandonia and Brenner, 2006; Laskowski, 2006). The functional assessment of proteins is thereby of secondary interest and is usually determined after the release of the protein structure or in conjunction with other labs. Moreover, in the race to be first to publish a structural model of a protein, unexpected interactions between protein and ligand can remain unidentified or unnoticed. Consequently, as of today (25/01/2009) 2290 protein structures deposited in the Protein Data Bank from structural genomics projects are classified as hypothetical proteins for which neither all functionally relevant ligands nor the function is known.



Computational tools for the prediction of protein ligand interactions can help X-ray crystallography and structural genomic projects to overcome their limitations and provide experimentalists with clues for the identification or verification of the protein's potential function. Crucial to the success of ligand prediction software is a complete understanding of key features in molecular recognition. However, the inadequate performance of most ligand predictions programs makes it clear that molecular recognition is far more complex than generally appreciated. Currently, computational methods still remain incapable of calculating accurate absolute binding affinities or distinguishing true-positive from true-negative ligand predictions (Gilson and Zhou, 2007). These persisting problems highlight misconceptions in the current state of our knowledge and necessitate further analysis and investigations into the nature of molecular recognition. As the actions of drugs are determined by the same forces that drive natural protein-ligand interactions, a complete knowledge of molecular recognition will also greatly benefit the development of cheaper, more effective drugs with fewer undesirable side effects.

## 1.1 Overview of thesis

The work presented in this thesis describes an investigation into the nature of molecular recognition. During the course of this thesis, I will address questions about the degree of shape complementarity between binding sites and their associated ligands as well as the extent to which their physicochemical properties vary. The results of these investigations should guide the future development of function prediction methods that rely on only the three-dimensional coordinates of the protein structure. Furthermore, analyses will be presented on the use of a new shape recognition methodology in the detection of ligand electron densities in X-ray diffraction experiments, which could help to locate and identify automatically small molecules in diffraction data for structural genomic projects. In this context the thesis is structured as following:

Chapter 2 introduces a number of topics relevant throughout this thesis, providing background information on protein structures, X-ray crystallography, enzymes binding sites and their characteristics. Furthermore, intermolecular forces with an emphasis on electrostatics and hydrophobic interactions are introduced together with a section on molecular surfaces and shapes, their visualisation, description and comparison. Some sections in this chapter have already been published in (Kahraman and Thornton, 2008).

The next three chapters detail the results of three different projects undertaken during the course of my graduate education. Each chapter is written in a journal publication manner starting with a brief introduction and literature review followed by sections on methods, results, discussion and conclusion.

Chapter 3 presents published results (Kahraman, et al., 2007a) on the prediction of ligands for protein structures lacking ligand coordinates from X-ray diffraction data. This is achieved by analysing the extent to which shape complementarity between binding site and ligand can guide function prediction efforts. For the comparison, a new shape description is introduced that is based on the Fourier analysis of molecular shapes with spherical harmonic functions.

Chapter 4 demonstrates the difficulty of predicting ligands for protein structures from a physicochemical perspective. The physicochemical complementarity in terms of electrostatic potential and hydrophobicity is tested between protein binding sites and ligands. The results have been partially published in (Kahraman, et al., 2007b) and submitted to a peer-reviewed journal (Kahraman, et al., 2009).

Chapter 5 addresses the recognition of ligands bound to protein structures from an experimental perspective through detection of ligands in X-ray diffraction data of the protein crystal.

Chapter 6 will conclude the presented work in this thesis, list details about potential future developments to incorporate this work into existing function prediction methods and give an outlook to main challenges of ligand recognition and function prediction that researchers will face in the coming years.

# Chapter 2

## Background

### 2.1 Protein structures

The atomic structure of a protein provides a great wealth of information. It reveals details on the final three-dimensional conformation of the protein sequence, its secondary structure elements ( $\alpha$ -helices,  $\beta$ -sheets, loop regions), how it functions at molecular level (catalysis, binding specificity, substrate promiscuity) or which forces act between protein and ligand in molecular recognition processes (electrostatics, hydrogen bonds, hydrophobic interaction). Key to the analysis of all these aspects in protein science is the determination and availability of protein structures.

#### 2.1.1 The Protein Data Bank

The three-dimensional coordinates of protein structures are usually deposited in the *Protein Data Bank* (PDB) (Berman, et al., 2000), the internationally recognized primary depository for all published three-dimensional biological macromolecules. The PDB was founded in 1971 at the Brookhaven National Laboratories containing an initial set of seven protein structures. In 1998 the PDB was put under the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB) at the Rutgers University, New Jersey (Berman, et al., 2000).

Since the beginning of the 1990's, the number of deposited structures in the PDB has been increasing exponentially. As of today (25/01/2009) the PDB holds 55,419 models of

macromolecule structures of which more than 92% are proteins, 4% are protein-nucleic acid complexes, 2% are DNA, 1% are RNA and few are carbohydrates and antibiotics. 99% of the protein structures were determined by X-ray crystallography (86%) and Nuclear Magnetic Resonance Spectroscopy (14%). 212 structures were solved with electron microscopy. Initially users were allowed to deposit their theoretical models from *ab initio* or homology modelling calculations, but this practice was stopped in 2006 (Berman, et al., 2000). For the latest statistics on the PDB's content and growth, see the PDB statistics on the homepage of the RCSB (<http://www.rcsb.org/pdb/>).

### **2.1.1.1 The worldwide Protein Data Bank**

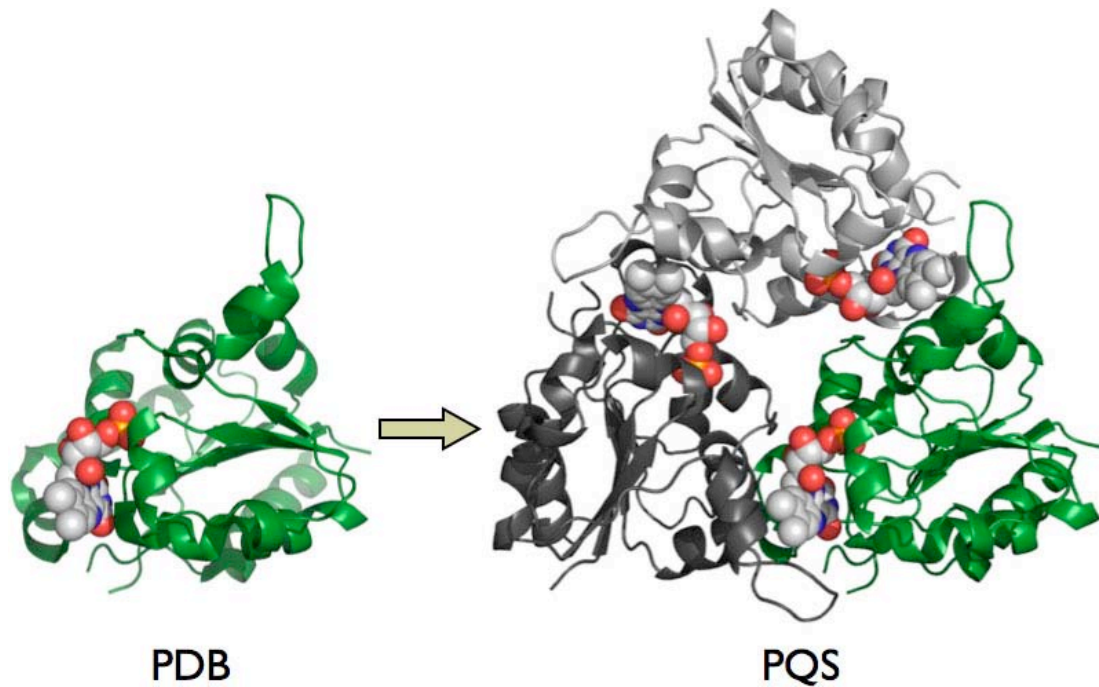
In 2003, the worldwide Protein Data Bank (wwPDB) (Berman, et al., 2003) was announced by its three founding members, namely RSCB, European Macromolecular Structure Database (MSD-EBI) and the Japanese PDB depository (PDBj) with the aim to sustain the PDB as the single non-profit and worldwide accessible depository for structural models of biological macromolecules. In the years since the PDB was founded, new experimental techniques have been developed to determine the structure of macromolecules. Automated scripts and applications have replaced most of the manual curation of deposited data and the Internet has evolved changing the way users submit, access and receive data from the PDB. All these innovations have required at certain times in the past adjustments to the data content and data format that, along with disagreements between curators and depositors, introduced inconsistencies in the PDB archive. It was recognised that a global effort on international level would be necessary to unify the PDB's data content and data format, which led to the founding of the wwPDB. The remediation project (completed at the end of 2007) within the wwPDB addressed the inconsistencies mentioned above. The project standardized chemical nomenclatures and labelling of amino acids, nucleic acids and small molecules, removed differences between sequences in the different depositories of the founding organizations,

updated citations to primary references, databases and taxonomies and improved the representation of large assemblies and viruses (Berman, *et al.*, 2007; Henrick, *et al.*, 2008).

## 2.1.2 Biological relevant protein conformation

Protein structures deposited in the PDB do not necessarily reflect the quaternary structure of the protein, i.e. the biologically relevant conformation, but generally show the *asymmetric unit* (ASU) of the protein's crystal. The ASU corresponds to the smallest fraction of the crystal that can construct the entire protein crystal with crystal symmetry operations on ASU duplicates. Structures deposited in the PDB as single chains are often actually dimers or tetramers or sometimes vice versa. This can lead to problems when analysing protein binding sites. Particularly, allosteric binding sites are often found at the interfaces of subunits and protein chains (Traut, 1994) (see Figure 2.1). In those cases, protein models as deposited in the PDB provide only a partial picture of the binding site and in the worst-case result in misleading conclusions drawn from incomplete data. It is therefore of utmost importance to use the biologically relevant conformation for any structural analyses on proteins especially those involving binding sites.

The Protein Quaternary Structure (PQS) (Henrick and Thornton, 1998) file server is a depository of the most likely quaternary conformations of all PDB structures. The quaternary structures in PQS were computationally calculated by applying crystal symmetry operations on ASU duplicates and selecting from the resultant assemblies those that have a certain loss in solvent accessible area, a particular difference in solvation energy and a minimum number of salt and interchain disulphide bridges. Recently an advanced version of the method was introduced called Protein Interfaces, Surfaces and Assemblies (PISA) (Krissinel and Henrick, 2007) that selects the most likely conformation on the basis of thermodynamic stability calculations.



**Figure 2.1: ASU vs. biologically relevant conformation.**

The asymmetric unit (ASU) in the PDB structure of the decarboxylase 1mvl shows only a monomer with the FMN being exposed to the solvent. However, the biological relevant conformation is a trimer as calculated by PQS, with a FMN binding site at the interface of two subunits.

### 2.1.3 Cognate ligands

Ligands bound to enzymes in crystal structures may not always be the native substrate or cofactor etc. Many such molecules found in the active site are enzyme inhibitors or ligand analogues, similar in structure to the native ligand molecule but prohibit the enzyme from completing its chemical reaction. Only when the enzyme is locked at a certain reaction step, is it possible to determine the structure of the protein-ligand complex. In addition, some ligands can be artefacts of the crystallization buffer, which contains different solvents promoting the crystallization process of a protein. In general, all ligands not required for the enzyme function can be classified as *non-cognates*, whereas endogenous ligands that are functionally related to an enzyme can be designated as *cognate*. For the protein-ligand interactions studied in this thesis, an attempt was made to use only binding sites with cognate ligands bound. This was important, as only cognate ligand binding sites allow conclusions to

be drawn on the convergent evolution of protein binding sites. Information about cognate ligands in enzymes can nowadays be retrieved from the PROCOGNATE (Bashton, et al., 2006) database that provides structural similarity scores between cognate and non-cognate ligands.

## 2.2 Protein structure determination

Different experimental techniques have been developed to determine the three dimensional structure of molecules. The current method of choice for determining atomic-resolution structures of proteins and small molecules is *X-ray crystallography* (Laskowski, 1992; Rhodes, 2000) followed by *Nuclear Magnetic Resonance* (NMR) spectroscopy (Wuthrich, 1990) (see statistics in section 2.1.1). Both techniques are fundamentally different in their approach, but produce comparable models of protein structures (Fan and Mark, 2003). In contrast to X-ray crystallography, protein structures analysed by NMR spectroscopy are studied in solution and thus are not affected by crystal packing (Jacobson, et al., 2002), but the size of the protein is generally limited to no more than 150 amino acids (<15 to 20 kDa) (Yee, et al., 2006). Furthermore, NMR spectroscopy provides an ensemble of conformations that give a better insight into the dynamics of the protein structure and reports the positions of hydrogen atoms, which are missing in most protein crystal structures.

The third most common technique to obtain a model of a protein structure is that of *electron microscopy* (EM) or its derivative cryo-EM, which are in particular powerful to solve the structure of large protein complexes like virus particles, ribosomes or spliceosomes (Auer, 2000; Frank, 2002). Protein structures determined by cryo-EM are to date far below high-atomic resolution (in general of 10-20 Å). However among the various techniques cryo-EM is the only one that allows single molecule imaging (Frank, 2002). Combining images from low-resolution cryo-EM protein complexes with high-resolution X-ray protein structures is



becoming a powerful tool to investigate the molecular mechanism of 'massive' complexes in atomic detail (Mitra and Frank, 2006).

A knowledge-based theoretical model can be constructed for proteins using a homologous protein with an experimentally determined structure; a process that is generally referred to as *homology modelling* (Blundell, et al., 1987). At the heart of homology modelling is a sequence alignment that determines potential insertions, deletions and replacements in the known protein structure, and an energy minimization procedure that removes all steric clashes between protein atoms in the modelled structure eliminating energetically unfavourable protein and amino acid conformations.

Here, I will solely focus on X-ray crystallography, as all protein structures and experimental data analysed in this thesis have been obtained using that technique.

## 2.2.1 Macromolecular X-ray crystallography

The atomic structure of molecules cannot be observed with conventional light-microscopes. The wavelength of visual light, which lies around 400–700 nm, is too large for such observations. To observe atomic details in molecules, wavelengths at atomic scale are required, i.e. in the range of 0.1–0.2 nm (1–2 Å). The light spectra that correspond to such ultra-short wavelengths are *X-rays*. Irradiating a molecule with X-rays produces a diffraction image that holds the information of the spatial distribution of the molecule's electron clouds. The diffraction from a single molecule however is too weak to be detected, although current developments in X-ray lasers are expected to allow single particle diffraction (Bogan, et al., 2008; Neutze, et al., 2000). Most radiation passes through the molecule without diffraction. However, in a crystal, myriads of the same proteins arrange into *unit cells*, which are the smallest building block of a crystal that can generate the entire crystal with just translation operations. Upon exposing the crystal to an X-ray beam, all unit cells scatter in the same

manner and together their single diffractions add up to produce crystal diffraction and detectable reflections on the detector.

An X-ray crystallography experiment could be divided into three stages of which the first is to crystallise a sample of a protein, followed by the second stage at which the protein crystal is irradiated with an X-ray beam and the X-ray diffraction images are collected on a detector. The final stage is concerned with producing a protein model from the measured diffraction images. It incorporates the estimation of initial phases for each reflection and the calculation of an initial electron density map. This is followed by an iterative process of model building and phase refinement that repeatedly improves the electron density map as well as the model of the protein structure. The refinement process continues until the protein model best agrees with the experimentally observed reflection data.

Throughout the various stages in the experiment, some obstacles have to be overcome. One of these is to bring a protein in solution to crystallise, i.e. form a three-dimensional lattice of well-ordered protein molecules that the experiment relies on. For an overview on the crystallisation of proteins, see (McPherson, June 2000). Protein crystals are held together by weak non-covalent interactions of electrostatic and hydrophobic character that can break under little stress. Membrane proteins present a particular challenge as their highly hydrophobic transmembrane helices cause membrane proteins to aggregate irregularly rather than form well-ordered crystals (Lacapere, et al., 2007). Even for soluble proteins, it might take some time to find with trial and error experiments appropriate physical and chemical conditions to induce crystallisation. Once a crystal has been obtained however, the next challenge is the data collection where the crystal is exposed to certain amount of X-rays. This exposure causes the crystal to heat up from the energy absorbed by the proteins. *Cryo-techniques* have been introduced to reduce this inevitable *radiation damage* and expand the life span of crystals in X-ray beams by rapidly cooling the crystal to 100K (Hope, 1990; Ravelli and Garman, 2006). However, new generation of *synchrotrons* produce X-ray beams of such high-intensity that radiation damage may remain problematic even at 100K (Garman, 1999).

The final major obstacle in determining the structure of a protein by X-ray crystallography is the phase problem, which will be illustrated later in section 2.2.1.1.4.

## 2.2.1.1 Interpretation of diffraction data

Mathematically crystal diffraction corresponds to the convolution of two functions, namely a continuous function that describes the electron density in a single unit cell and a discrete function that describes the lattice of unit cells in the crystal. The result of the convolution is a diffraction pattern that consists of discrete spots at which the continuous function of a single unit cell is sampled discretely by the discrete function of the crystal lattice.

### 2.2.1.1.1 Bragg's law

A protein crystal and the unit cell within it can be sliced into various sets of regular spaced parallel planes. Each set of parallel planes is labelled by *Miller indices*  $hkl$  where the numbers denote the number of times the planes intercepts the unit cell edges  $a$ ,  $b$ ,  $c$ . When the crystal is irradiated each member in the  $(hkl)$  planes function as a mirror, reflecting the incident X-ray waves, whereby the angle of incidents  $\theta_I$  equals the angle of reflection  $\theta_R$ . Reflected waves from successive planes interfere with each other, but the interference is constructive only if the waves are in phase, in which case the wave amplitudes add up to give a stronger reflection on the detector. If the waves are out of phase, they cancel out each other to give no reflection. According to *Bragg's law* constructive interference of monochromatic X-ray waves can only occur at those set of parallel planes  $(hkl)$  for which

$$2d_{hkl} \sin\theta = n\lambda \quad , \quad (2.1)$$

where  $d_{hkl}$  is the interplanar distance of the  $(hkl)$  diffracting planes,  $\theta$  is the scattering angle,  $\lambda$  is the wavelength of the X-ray beam and  $n$  is the order of reflection, an arbitrary positive integer.

### 2.2.1.1.2 Resolution

All  $(hkl)$  planes that satisfy Bragg's law produce a single distinct reflection on the detector with a likewise associated Miller index  $hkl$ . The distance between the reflections on the detector are inversely related to the distances in the unit cell, as according to Bragg's law smaller interplanar distances  $d$  diffract only at larger scattering angles  $\theta$ . The inverse relationship between the crystal lattice, the *real lattice*, and the reflections, the *reciprocal lattice*, has the consequence that all reflections that are close to the origin of the diffraction pattern encode low-resolution information, whereas reflections farther away hold higher resolution information. Using the outermost reflections in a diffraction pattern and their scattering angles  $\theta_{\max}$ , one can calculate the minimum interplanar distance  $d_{\min}$ , i.e. the *resolution* of the experiment:

$$d_{\min} = \frac{\lambda}{2 \sin \theta_{\max}} \quad . \quad (2.2)$$

Resolution in X-ray crystallography and optics is not equivalent. The concept of resolution in optics gives information about the minimum distance that can be resolved in a specimen (Garini, et al., 2005). In X-ray crystallography, however, structural constraints known *a priori* (e.g. atom identity, bond length and bond angle) can improve the precision of atom positions in the protein model, such that protein structures with  $d_{\min} = 2.0 \text{ \AA}$  resolution can show structural details that are below  $d_{\min}$ .

### 2.2.1.1.3 Structure factors

Each X-ray that impinges on the detector and manifests itself as a reflection  $hkl$ , is the sum of superimposed X-ray waves diffracted from all  $(hkl)$  planes sampling the molecular electron density within all unit cells. Thus, every single protein atom with its electrons contributes to

each reflection on the detector. If each atom is considered as a point scatterer, then the impinging X-rays can be written as a sum of  $N$  scatterers:

$$F(hkl) = \sum_{j=1}^N f_j e^{2\pi i(hx_j + hy_j + hz_j)} \quad , \quad (2.3)$$

where  $f_j$  is the scattering factor of atom  $j$  (see section 2.2.1.2.2),  $hkl$  are the Miller indices and the coordinates  $x, y, z$  are the fractional coordinates (e.g.  $x/a$ ) in the unit cell. The quantity  $F(hkl)$  is designated as *structure factor* with an amplitude  $|F(hkl)|$ , the phase angle  $\alpha_{hkl}$  and the wavelength, which is the wavelength of the X-ray source:

$$F(hkl) = |F(hkl)| e^{i\alpha_{hkl}} \quad , \quad (2.4)$$

A different expression for structure factors and the above equation in diffraction experiments is related to the fact that structure factors are the *Fourier Transforms*  $\mathcal{F}$  of the electron density  $\bar{\rho}$  within the average unit cell in the crystal:

$$F(hkl) = \mathcal{F}(hkl) = \int_V \bar{\rho}(x, y, z) e^{2\pi i(hx + ky + lz)} dV \quad , \quad (2.5)$$

where the integral is over the entire average unit cell volume  $V$  and the coordinates are again to be taken as fractionals.

Given the diffraction data on the other hand, the associated electron density can be calculated with the *inverse Fourier transform* of the structure factors  $F(hkl)$ . Due to the discrete set of reflections on the detector the integral in equation ( 2.5 ) becomes a triple sum over the Miller indices:

$$\bar{\rho}(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F(hkl) e^{-2\pi i(hx + ky + lz)} \quad . \quad (2.6)$$

Note the different sign of the argument in the exponential function as compared to the Fourier transform in equation ( 2.5 ).

#### 2.2.1.1.4 Phase problem

According to equation ( 2.4 ) the complete description of the structure factor  $F(hkl)$  as a complex function requires not only the knowledge of its amplitude  $|F(hkl)|$ , but also its phase angle  $\alpha_{hkl}$ . The amplitude corresponds to the square root of the measured reflection intensity on the detector:

$$|F(hkl)| = \sqrt{I_{hkl}} \quad . \quad (2.7)$$

The phase angle however cannot be measured directly on the detector and must be inferred by other means, hence the term *phase problem*.

#### 2.2.1.1.5 Obtaining initial phases

Various experimental techniques have been developed over the last decades to obtain phase information from diffraction experiments. See *Volume F* in the *International Tables for Crystallography* for a comprehensive list of these techniques. The three techniques most commonly used are *Multiple Isomorphous Replacement* (MIR) (Mathews, 1966), *Multiwavelength Anomalous diffraction* (MAD) (Walsh, et al., 1999) and *Molecular Replacement* (MR) (Rossmann, 1990). Among those is MIR the oldest technique and was introduced by Max F. Perutz into protein crystallography in his attempt to solve the X-ray structure of haemoglobin (see Chapter 1). MIR works on the idea of comparing the diffraction pattern of a native crystal with the diffraction pattern of a crystal that was soaked or co-crystallized in a solution of heavy metal ions like e.g. mercury (Hg), platinum (Pt) or gold (Au). Some of these metal ions bind at specific binding sites on the protein such that they are found in the same positions in each unit cell. In the case that these metal ions do not significantly

distort the protein conformation or the unit cell dimensions, the protein crystal remains isomorphous and produces the same diffraction pattern. However, the heavy metal ions, being strong diffractors due to their large number of electrons, alter the intensities of all reflections relative to the intensities observed for the native structure and thus serve as reference scatterers. The difference in the amplitudes of the measured reflections is exploited to obtain the initial set of phases with the Patterson map. Information on the preparation and the characterisation of heavy atom derivatives can be found in the Heavy-Atom Databank (<http://www.sbg.bio.ic.ac.uk/had/>) (Carvin, et al., 2006).

Similar changes in the reflection intensities can be produced with the MAD. MAD exploits the electronic transition phenomenon of electrons at certain characteristic energy levels called the *absorption edges*. Atoms that are exposed to X-rays with a wavelength close to one of their absorption edge produce anomalous diffraction i.e. the diffracted X-ray has a disturbed amplitude and phase as compared to the diffraction far away from the absorption edge of the atom. The difference in the reflection intensities can be exploited similarly to MIR to estimate the set of initial phase angles for the structure factor calculation. Most of the success of MAD can be ascribed to its application with modified amino acids like selenomethionine (Se-Met) replacing methionine in protein structures. Selenium being a strong anomalous scatterer allows the computation of initial phases on a single protein crystal (Carvin, et al., 2006).

In those cases where the protein structure of a close homolog has been solved, which is likely to have a similar fold (Barton, 1992; Chothia and Lesk, 1986), MR promotes the usage of initial phase angles derived from the known atomic positions of the homologous structure. Despite an arguable similarity in their fold, both proteins might have different arrangements in the unit cell, which demands exhaustive translation and rotation operations on the homologous structure to find the location and orientation of the target protein in the unit cell.

### 2.2.1.2 Refinement of protein model and phases

Once initial phase estimates for the structure factors are obtained, a first *electron density map* for the protein can be calculated. An electron density map is a three dimensional scalar field of electron density  $\rho(x,y,z)$  values reflecting the content of electron charges per cubic Ångström of the average unit cell in a crystal and is usually visualised as a contour plot of isosurfaces at a constant  $\rho(x,y,z)$  value. Experimental inaccuracies e.g. in reflection measurements, or approximation errors in estimating the phase angles, or systematic errors in working with non-isomorphous derivatives in isomorphous replacement, prevent the initial electron density map to show atomic details of the protein structure. A generally iterative process helps crystallographers to reduce these errors and refine the protein model at the same time. This refinement process involves a back and forth transition between the electron density and the protein model in real space and the calculated and measured structure factor amplitudes in reciprocal space with constant adjustments of parameters (see below). Depending on whether parameters are adjusted in real space or reciprocal space, the refinement process is referred to as *real-space* or *reciprocal-space refinement*. The latter has at least two advantages over the former. Firstly, reciprocal-space refinement can be performed either only on the structure factor amplitudes or on the whole structure factors. Secondly, in reciprocal-space refinement each of the structure factors can be weighted according to their reliability, whereas the electron density in the real-space refinement is treated everywhere as equally reliable (Blundell and Johnson, 1976).

The refinement process will be crucial in Chapter 5, where insufficient refinement leads to an inability to automatically recognise bound ligand molecules in electron density maps. Without appropriate refinement, the shape of the electron density of a ligand molecule often deviates from the shape of the ligand molecule to such an extent that the similarity between both shapes is lost.



### 2.2.1.2.1 Atomic disorder in refinement

Protein atom coordinates are far from static and often vibrate around an equilibrium position within the protein structure. These vibrations are induced by thermal motion and are subject to environmental constraints such as covalent bonds or hydrogen bonds (dynamic disorder). As a consequence, side chain atoms or solvent exposed amino acids vibrate more than main chain or amino acids in the protein interior. The crystallographic experiment presents a time-averaged snapshot of a structure presenting all different states of atomic motion at the same time. During refinement, it is possible to determine the distribution of positions for each individual atom. This information is absorbed into and described by the *temperature factor* or *B-factor*. For isotropic distributions, i.e. vibrations with equal magnitude in different directions, the temperature factor  $B_j$  of an atom  $j$  is given by

$$B_j = 8\pi^2 \overline{\mu}_j^2 = 79 \overline{\mu}_j^2 \quad , \quad (2.8)$$

where  $\overline{\mu}_j^2$  is the squared average displacement of atom  $j$  around its mean position in the crystal.

High temperature factors in parts of the protein will induce dynamic/static disorder in the protein crystal affecting the diffraction of incident X-ray beams and reducing the resolution of the crystal. Thus, it is essential to include the temperature factor into the refinement process.

Furthermore, parts of the protein or even single amino acid side chains can have alternative locations and occupy different conformations in different unit cells (static disorder), in which case the *occupancy value* informs about the relative frequency of each conformation as a percentage. Again, alternative locations within the protein crystal will affect the X-ray diffraction and as a result must be included in the refinement process in order to obtain a best-fit model to the experimental observed data.

### 2.2.1.2.2 Calculating structure factors

The necessity of including the temperature factor  $B_j$  and occupancy value  $n_j$  for a better fit to the observed experimental data, requires adjustments to the structure factors equation ( 2.5 ):

$$F(hkl) = G \cdot \sum_{j=1}^{N_{\text{atoms}}} f_j S_j n_j e^{2\pi i(hx_j + ky_j + lz_j)} \quad . \quad (2.9)$$

$G$  scales the computed structure factor amplitudes such that they are comparable to the experimentally observed amplitudes. The intensity of the reflections is related to the *atomic scattering factors*  $f_j$ , which correlates with the number of electrons in an element and falls off with increasing scattering angle  $\sin\theta$ . Tabulated values of  $f_j$  for each element can be found in *Volume C* of the *International Tables of Crystallography*. The amount of diffracted X-rays at larger scattering angle  $\sin\theta$  decreases further with increasing temperature factor  $B_j$ , which is described by the temperature factor correction term  $S_j$ :

$$S_j = e^{-B_j(\sin\theta/\lambda)^2} \quad . \quad (2.10)$$

### 2.2.1.2.3 Judging the protein model

When judging how good a model fits the underlying data, it is convenient to have a single statistical quantity whose value is indicative of the quality of the fit. In protein crystallography the *Reliability-factor* or *R-factor* is the most common used statistical measure to assess the fit of the protein model to the experimental data (Laskowski, 2003). The *R-factor* is defined as

$$R = \frac{\sum_{hkl} \left| |F_o(hkl)| - |F_c(hkl)| \right|}{\sum_{hkl} |F_o(hkl)|} \quad , \quad (2.11)$$

where  $|F_o(hkl)|$  is the experimental observed structure factor amplitude of reflection  $hkl$  and  $|F_c(hkl)|$  is its associated computed amplitude from equation ( 2.9 ).  $R$ -factors display a dependence on the resolution of the X-ray data but tend to be around 0.20 for correct protein and nucleic acid structures. Random structures typically have a  $R$ -factor in the range of 0.40 to 0.60 (Laskowski, 2003). At the beginning of the refinement process, the agreement between experimental and computed structure factor amplitudes must be better than random and should improve during the refinement process to a value below 0.2.

A general problem of refinement in X-ray crystallography is the risk of overfitting the relatively small amount of experimental data (number of reflections) that is available in order to infer the large number of parameters (including  $x,y,z$  coordinates, temperature factor, occupancy value for each atom) in a protein structure. Overfitting can produce a model of a protein structure that is incorrect despite a low  $R$ -factor. To avoid overfitting, the *free R-factor* or  $R_{\text{free}}$  was introduced (Brunger, 1992), which is generally considered more reliable than the ordinary  $R$ -factor. The  $R_{\text{free}}$  calculation is based on a cross-validation of the refinement process with a small test set of reflections randomly chosen from all observed reflections on the detector and put aside at the beginning of the refinement process. When the refinement process initiates, the refinement is performed only on the remaining reflections, the working set, and not on the reflections in the test set. The test set is used only to judge how well the model fits the data that have not been used for its refinement.

#### **2.2.1.2.4 Refinement procedures**

A number of methods have been developed that all aim to deliver an improved agreement between the experimentally observed and calculated structure factors. The most common techniques are the least-squares difference refinement, energy minimisation refinement, maximum-likelihood refinement and molecular dynamics/simulated annealing refinement.

The oldest of the refinement techniques, the *reciprocal-space least-squares difference refinement*, aims to minimize the sum  $D$  of squared differences between the observed  $F_o(hkl)$  and calculated  $F_c(hkl)$  structure factor amplitudes:

$$D = \sum_{hkl} w_{hkl} \left( |F_o(hkl)| - |F_c(hkl)| \right)^2, \quad (2.12)$$

with  $w_{hkl}$  being a factor to weight the reliability of the observed structure factor amplitude. However, the least-squares technique coupled with a plain gradient search will only find the nearest local minimum to a starting point. This can be problematic at the beginning stages of the refinement process. *Energy minimization refinement* (Levitt and Lifson, 1969) adds to the least-square equation ( 2.12 ) further terms that minimize the energies of the bond distances, bond angles, torsion angles and van der Waals potentials.

Rather than optimising the mean squared difference between structure factor amplitudes, *maximum-likelihood refinement* (Murshudov, *et al.*, 1997; Ten Eyck and Watenpaugh, 2006) explores the probability that the calculated structure factor amplitudes correspond to the observed data and is thus independent of the size of deviations between both set of structure factor amplitudes. Under the assumption that all observed reflections are independent from each other, the probability of estimating calculated structure factor amplitudes from given observed structure factors is given by the likelihood function  $L$ :

$$L = \prod_{hkl} P\left(|F_o(hkl)|, |F_c(hkl)|\right), \quad (2.13)$$

where  $P$  is the conditional probability distribution of  $F_o$  when  $F_c$  is known. Similarly, the likelihood function can be expressed as a sum simplifying its calculation:

$$\log L = \sum_{hkl} \log P\left(|F_o(hkl)|, |F_c(hkl)|\right). \quad (2.14)$$

In the course of the refinement process, the logarithmic likelihood function  $\log L$  becomes minimized.

The most recent refinement technique uses *molecular dynamics* in combination with *simulated annealing* (Brunger, et al., 1987). The idea behind this technique is to virtually heat up the protein structure to high temperatures (2,000-4,000K) inducing large atomic motions in the protein model, followed by a stepwise cooling process to room temperature (300K) at which the protein atoms come to rest at an energy minimum (Laskowski, 1992). The motions of the atoms follow Newton's laws, while the energy of the protein model is evaluated with the same energy terms that were indicated above. Whenever a protein atom in the model changes its location by some distance (usually 0.4 Å), the energy of the protein model is re-evaluated and the model is accepted only if its energy has decreased or its Boltzmann probability  $\exp(-\Delta E/kT)$  is lower than a random number. The simulated annealing technique allows the exploration of large conformational space and in contrast to conventional refinement techniques overcomes local minima of erroneous models without manual intervention.

### 2.2.1.3 Automated protein modelling

Just 10 years ago, the entire process of protein structure determination as described above required several months and often years of work. However, structural genomic approaches have helped to reduce the time span needed to only a couple of weeks if not days. This achievement has been realized mainly by the introduction of automation protocols that have significantly reduced the demand for human intervention at various levels of the structure determination process.

One process that has greatly benefited from automation is that of *protein modelling* where a molecular model of a protein is produced after an initial electron density map is calculated. Traditionally this process was carried out manually on a graphics terminal using specialized software such as O (Jones, et al., 1991) and XtalView (McRee, 1992) and required an expert in both fields of structural biology and chemistry. However with the implementation of various

algorithms to support the manual process of protein modelling it became clear that the whole process of protein modelling could be fully automated leading to software like ARP (Automated Refinement Protocol) (Lamzin and Wilson, 1993), QUANTA (Oldfield and Hubbard, 1994), RESOLVE (Terwilliger, 2003) or Buccaneer (Cowtan, 2006).

All these software have in common that they gradually build the protein structure into the electron density. The main chain is often the start point to the building process, as it tends to have continuous and relatively strong density. This is followed by side chains that are less rigid with weaker and more smeared electron density. In the final stages of the building process solvent molecules that have usually the weakest density are fitted into the map (Palmer and Niwa, 2003). QUANTA, for example, runs a skeletonisation process to reduce the electron density to a set of lines that captures the connectivity of the protein structure. RESOLVE on the other hand compares each point in the electron density to a set of common density templates, whereas Buccaneer uses a density likelihood function to approximate the Ca atom positions in an electron density map.

In ARP and its successor ARP/wARP (Cohen, et al., 2008) most steps of model building are combined with structure refinement in an iterative manner (Morris, et al., 2007). The conception behind ARP/wARP is to scatter free dummy atoms at certain distances from each other at locations close to high-density peaks. These atoms are treated as potential Ca atoms and connected to peptides if they show resemblance to a library of known peptide structures. A hybrid model is built with fragments of protein structure and free atoms, followed by an iterative cycle where free atoms are step-wise connected to fragments and are assigned their chemical identity. Once the main chain is determined, a rotamer library is employed to fit the side chain into the density. Non-rigid loop regions are gradually modelled by predicting the residues in the loop from the conformation of the preceding residues. Repeating these predictions produces a set of possible loop conformations from which the one with the best fit to the density is selected. The overall iteration process terminates as soon as the best possible phases and a complete protein model is obtained. Continuous development efforts

have been invested into ARP/wARP over the last decade. Given data with sufficient resolution, in general 2.7 Å or better, and a source of reasonable phase estimates, ARP/wARP reliably models the protein structure automatically from an electron density map and a given protein sequence (Cohen, et al., 2008; Langer, et al., 2008).

## 2.3 Enzymes

*Enzymes* belong to one of the most important protein classes. As highly specific macromolecular catalysts, they induce chemical reactions at physiologically mild conditions. Most enzymes are proteins, although some RNA molecules have also been found to catalyse chemical reactions. Like all catalysts, enzyme speed up their chemical reaction on its target molecule by lowering the reaction's activation energy. Often the reaction is thereby speeded up by a factor of several magnitudes. For example, carbonic anhydrase hydrolyses the conversion of carbon dioxide to carbonic acid  $10^6$  times faster than a similar non-catalysed reaction (Sly and Hu, 1995). Like all other catalysts, an enzyme is not consumed, i.e. not changed, after participating in a reaction and therefore can immediately take part in further reactions (Alberts, *et al.*, 1994c). Table 2.1 lists some statistics on enzymes from publicly accessible databases.

Enzymes are classified according to the chemical reaction they catalyse and the substrate they act upon. The Enzyme Commission (EC) assigns each enzyme an *EC-number* that consists of four digits. The first number (class) indicates the reaction type; the second number (subclass) together with the third number (sub-subclass) represents the occurring chemistry and the last number gives the substrate specificity (Barrett, 1997).

**Table 2.1: The extent of enzyme data in some structural databases.**

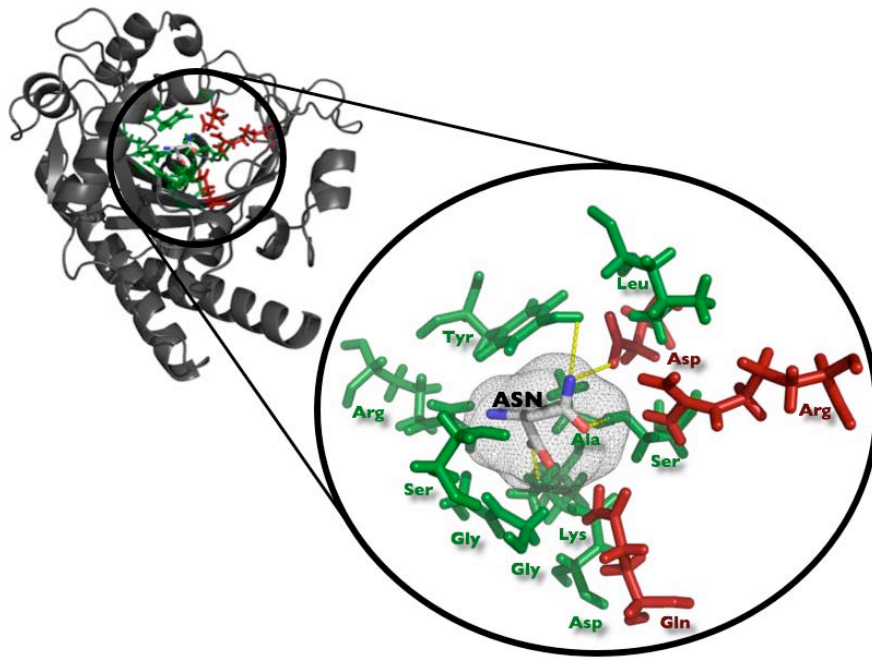
Number of	Quantity
Known enzyme reactions (unique EC numbers)	4,115
Enzymes in UniProtKB/Swiss-Prot (Apweiler, et al., 2004)	175,974
PDB files of enzymes	22,935
EC reactions in PDB	1,681
Enzymes with catalytic residues in CSA	968
Enzymes with catalytic mechanisms in MACiE (Holliday, et al., 2005)	223
<i>Enzymes as specified by EC number in PDB with the largest number of structures</i>	
1. DNA-directed DNA polymerase, EC 2.7.7.7	1,229
2. Lysozyme, EC 3.2.1.17	1,173
3. Ribonuclease H, EC 3.1.26.4	855
<i>Most enzymes in PDB originate from</i>	
1. Human	5,521
2. <i>Escherichia coli</i>	2,481
3. Cow	958
4. Baker's yeast	725
5. House mouse	453
No of organisms that have one or more enzyme structures in PDB	834
The data was collected on the 25/01/2009.	

### 2.3.1 Enzyme binding sites

All result chapters in this thesis analyse principles of molecular recognition between proteins and small molecules. Therefore, the focus of the following two sections will be on binding sites of small molecules within enzyme structures. All facts and concepts introduced in both sections can be regarded universal to all protein binding sites that bind small molecules.

Enzyme *binding sites* are regions on the surface of an enzyme specially evolved to interact with other molecules (see Figure 2.2). These binding sites can differ in their functions and the





**Figure 2.2: Binding Site of *Escherichia coli* asparagine synthetase.**

The structure of the *Escherichia coli* asparagine synthetase. (PDB-Id: 12as) is shown with a zoom-in into the binding site of the substrate asparagine. Binding site residues as determined by HBPLUS are coloured in green, catalytic active residues were extracted from Catalytic Site Atlas are coloured in red and the substrate is varicoloured. Hydrogen bonds between binding site residues and substrate are indicated by yellow dashed lines. The binding site shape is shown as a grey mesh as approximated with spherical harmonic functions (see section 2.5.2.3).

molecules they bind. In this thesis, a binding site is defined as the cluster of protein atoms on the protein surface, which interacts with the binding partner via hydrogen-bonding and other non-covalent bonds. In contrast, a *binding pocket* is the negative picture of the binding site i.e. the voluminous imprint of the binding site in space, which bears the ligand.

### 2.3.1.1 Active site

Amongst the most important binding sites for the function of an enzyme is the *active site*, which consists of at least two parts (see Figure 2.3a). The first part is the *catalytic site*, which contains the catalytic machinery of the enzyme in the form of usually two to six amino acids that perform the catalytic reaction. The second part is the *substrate binding site*, which specifically recognizes the molecule upon which the enzyme acts. Beside the specificity, the

substrate binding site also provides binding energy to keep the substrate bound on the active site while the catalytic reaction progresses. Enzymes can act on a huge variety of substrates, from small molecules like hormones and sugar, moderately sized molecules like polypeptides and oligosaccharide, to macromolecules like DNA and even other proteins. Figure 2.2 shows an example of a substrate binding-site for an asparagine amino acid in the structure of the *Escherichia coli* asparagine synthetase.

### 2.3.1.2 Cofactor/coenzyme binding site

As enzymes are proteins, they usually consist of a set of 20 *amino acids*. Each amino acid is distinct in its chemical characteristics with either a hydrophobic, polar or charged side chain. For some catalytic reactions, the chemical properties of these amino acids may be sufficient, but for the majority of reactions such as redox reactions or chemical group transfers, enzymes require the assistance of additional molecules that bind to the third part of the active site. These molecules are defined as either *cofactors*, which are tightly bound to the enzyme throughout the catalytic reaction, or *coenzymes*, which are released during the reaction. Cofactors distinguish themselves from coenzymes by being not consumed in the catalytic reaction. Though they are altered while the catalysis takes place, they are recovered again in the same catalytic process. In contrast, coenzymes support the enzyme reaction by providing chemical groups to the substrate and subsequently detach from the enzyme to start a recovery process outside the enzyme. Typical cofactors are the inorganic metals and sulphate ions or the organic flavin and heme groups. Examples of coenzymes are vitamins or the cellular energy carrier ATP.

### 2.3.1.3 Allosteric sites

Some enzymes, especially those composed of several domains or several chains, can have *allosteric sites* in addition to the substrate and cofactor/coenzyme binding sites (see Figure

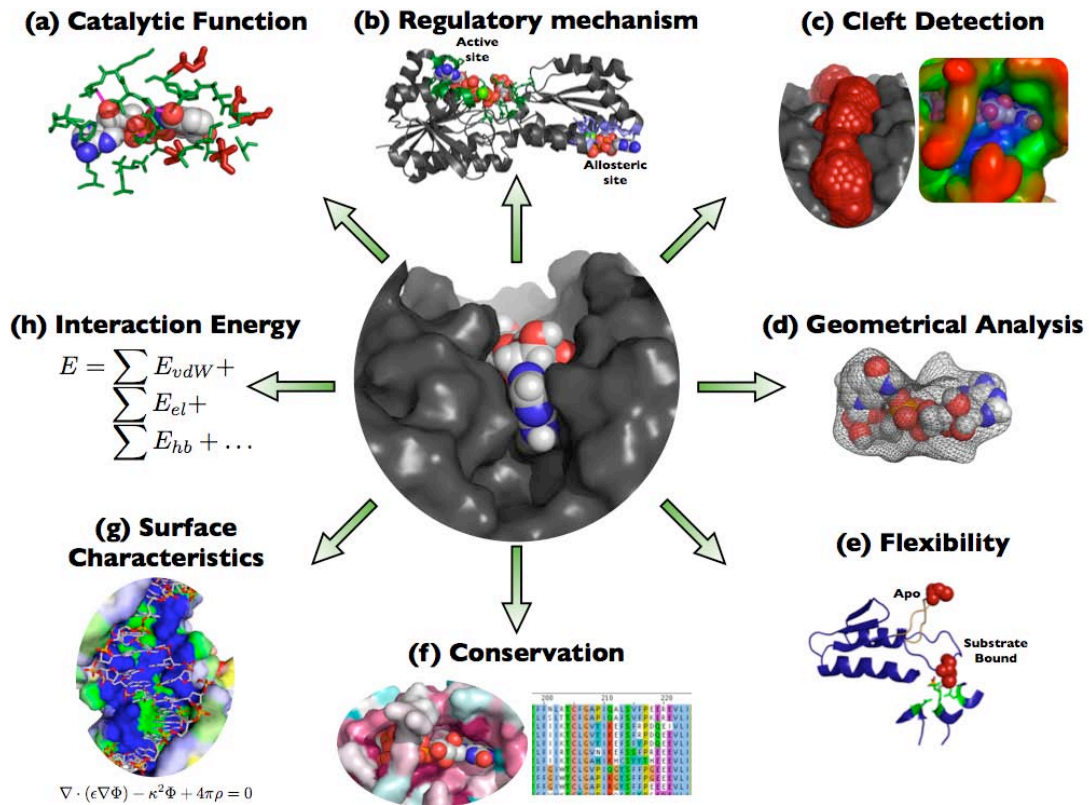
2.3b). These allosteric sites play an important role in the regulation of an enzyme's activity as they induce conformational changes on the whole enzyme structure upon binding of a regulator molecule, which can affect the conformation of the active site. Depending on whether the regulator molecule is an *effector* or an *inhibitor*, the changes on the active site can either enhance or reduce the enzymatic catalysis.

## 2.3.2 Characteristics of enzyme binding sites

From investigation on the three-dimensional structures of enzymes, it became evident that substrates and secondary molecules like cofactors and coenzymes do not bind randomly on the enzyme surface. The same molecule always binds at the same site within the same enzyme structure. This has led to the assumption that binding sites must have unique features that distinguish them from other areas on the enzyme surface (Ringe, 1995) and in addition allow the binding site to distinguish its associated molecule from the thousands that exist in a living cell.

Two models were suggested to explain the particular specificity of active sites. Firstly, the *Lock and Key model* by Emil Fischer (Fischer, 1894) and secondly the *Induced Fit model* by Daniel Koshland (Koshland, 1958). The Lock and Key model assumes that a ligand is geometrically complementary to its active site and that both shapes fit exactly into one another. The more recent model of Induced Fit is a modification to the Lock and Key model and incorporates the flexibility of enzymes and substrates. The model suggests an 'open' state for an enzyme when the substrate binds, followed by a 'closed' state where the enzyme encloses the bound substrate and performs its catalysis (Gutteridge and Thornton, 2005). In the process of converting from the open state to the closed state the active site adjusts its shape to the transition state that is the conformation of the ligand at the highest reaction energy. It is therefore generally assumed that the transition state is more complementary to the enzyme binding site than the substrate molecule. (Branden and Tooze, 1999).

The next subsections introduce the difference characteristics of protein binding sites that are depicted in Figure 2.3. Each subsection will provide some background information to each feature and describe one exemplary methodology to calculate it.



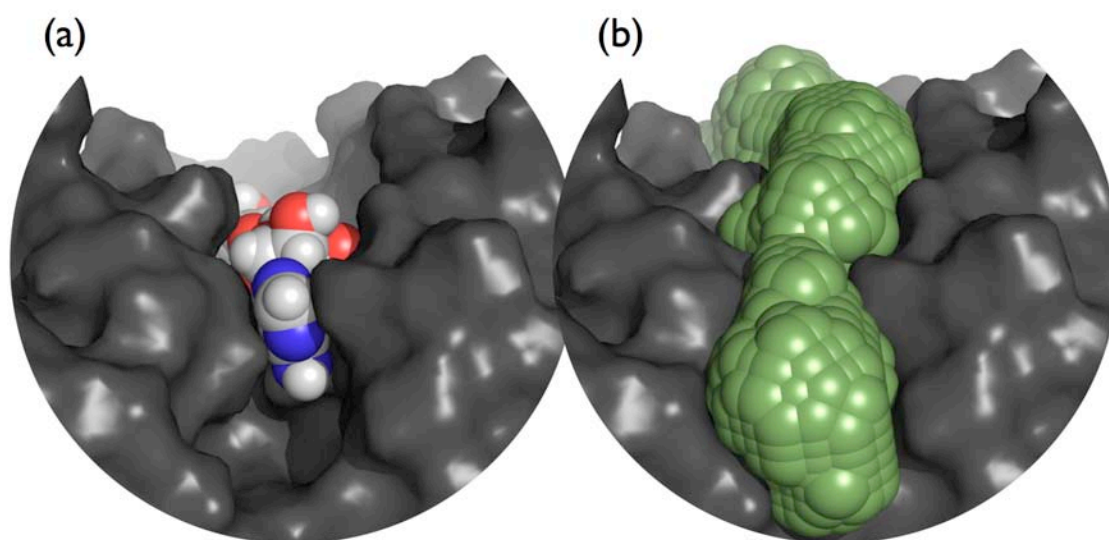
**Figure 2.3: Characteristics of enzyme binding sites.**

(a) The active site is a specific binding site in an enzyme that contains catalytic residues to perform the enzymatic reaction on a substrate. (b) The activity of an enzyme can be regulated for example by allosteric regulator molecules that bind to a remote binding site. (c) In most enzymes the active site is found in the largest or deepest cleft of the enzyme, (d) and encloses at least partially the ligand with amino acids, resulting in similar geometrical shapes for binding site and ligand. (e) Binding sites can undergo major conformational changes upon substrate binding, especially when some parts of the site are located in flexible loops. (f) As binding sites are essential for the function of a protein, their residues are often amongst the most highly conserved residues. (g) The binding affinity of a ligand is influenced by the physicochemical properties on the binding site surface like complementary electrostatic potentials or perturbed  $pK_a$  values (h) which can be exploited to calculate estimated binding energies between ligand and binding sites.

### 2.3.2.1 Volume

Enzyme active sites tend to be within sizeable depressions on the protein's surface, which are known as *clefts* or *pockets* (see Figure 2.3c). In 70-85% of enzymes the largest of these clefts is where the substrate and relevant cofactors or coenzymes bind (Laskowski, et al., 1996). The average volume of a binding site depends on the ligand it binds and ranges mostly between 400 to 2000 Å<sup>3</sup> (see Table 3.3).

SURFNET (Laskowski, 1995) is a simple approach to identify and visualize clefts in proteins. It detects gap regions within the protein by fitting spheres of a certain range of sizes between protein atoms. The spheres are not allowed to clash with any neighbouring protein atoms. Overlapping SURFNET spheres are clustered and regarded as protein clefts (see Figure 2.4). Placing a grid on the cleft and determining the number of grid cells enclosed by any sphere, enables the calculation of the volume for each cleft.



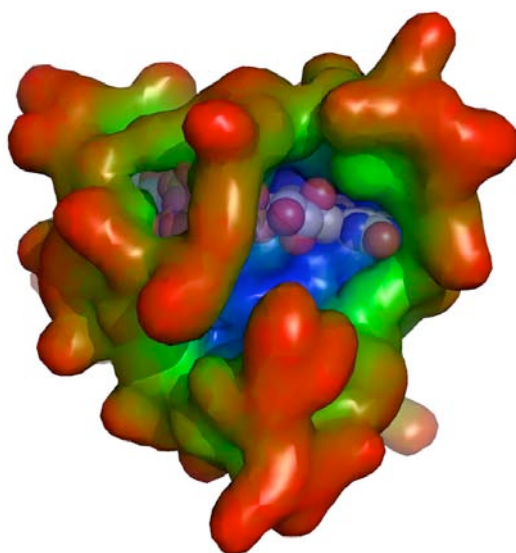
**Figure 2.4: SURFNET spheres filling a cleft on the protein surface.**

a) Spherical section of the protein structure of ribosyltransferase (PDB-Id: 1og3) coloured in dark grey, with bound coenzyme NAD in the active site. b) Largest cleft as determined by SURFNET contains the active site. SURFNET spheres are represented by light green spheres.

### 2.3.2.2 Depth

Enzymes maximise the number of interactions with their ligands by surrounding the ligand using large and deep clefts (see Figure 2.3c) (Kraut, et al., 2006). In particular, active sites are often found in the deepest cleft of an enzyme. The average depth of a cleft that contains a binding site depends on the protein size and can be as deep as 30 Å (Coleman and Sharp, 2006).

The algorithm of travel depth (Coleman and Sharp, 2006) is an elegant way to visualize and measure the depth of clefts relative to the *convex hull* (Barber, et al., 1996) of the enzyme's molecular surface (see Figure 2.5). The convex hull is defined for a simplified two-dimensional molecule as the region that is enclosed by an imaginary rubber band that is stretched around the whole molecule. The travel depth algorithm finds for a probe sphere on the protein surface the minimum distance to reach the convex hull. It works by placing the protein into a grid and assigning to all grid cells outside the convex hull a depth of zero. For grid cells inside the convex hull the algorithm scans recursively through the grid and adds to the size of each grid cell the minimum depth of its neighbouring cells.



**Figure 2.5: Depth of a protein surface calculated by Travel Depth.**

Travel Depth algorithm applied to PDB structure 1p4m. Molecular surface is coloured according to depth, starting from red for 0 Å depth, to green for 7.5 Å depth to blue for 15 Å depth. The deepest surface patch corresponds to the cleft that contains both binding sites. Ligands ADP and FMN are partially visible as atomic spheres in the deepest cleft.

### 2.3.2.3 Shape

It is a common assumption that shapes of protein binding pockets are complementary to the shapes of the ligands they bind (see Figure 2.3d). This assumption is manifested by the Lock and Key model and Induced Fit model for molecular binding (see section 2.3.2).

For the analysis and visualization of binding site and ligand shapes, an elegant approach will be introduced in Chapter 3. The method is based on the Fourier analysis of a radial function that describes the surface of a molecule.

### 2.3.2.4 Flexibility

The Induced Fit model for molecular binding states that enzymes undergo conformational changes upon substrate binding. For a small fraction of enzymes, these changes are large, particularly if they include a flexible loop region that closes/opens the entrance to the active site, thus preventing/allowing the binding of a ligand (see Figure 2.3e). However, for the majority of enzymes the changes are small. The average RMSD (see below) upon ligand binding between C $\alpha$  atoms of binding sites and catalytic residues is less than 1 Å (Gutteridge and Thornton, 2005). Similar values are observed for the side chain atoms. It is interesting to note that residues in active sites are on average more flexible than other residues in the protein structure due to geometrical adjustments of the active site residues to the transition state of the ligand. Binding sites that undergo large conformational changes upon binding a ligand were often found to have a large number of hydrophobic residues, including the large aromatic amino acid tryptophan and form interactions that do not require a directional constellation of residues such as hydrophobic-hydrophobic, aromatic-aromatic and hydrophobic-polar residue pair interactions. The lack of directionality allows non-polar residues to maintain the interaction network during altering conformations of the binding site (Gunasekaran and Nussinov, 2007). But there are also enzymes, like prothymosin- $\alpha$ , that are intrinsically disordered in their native state (Uversky, et al., 2000). Neither the Lock and Key

nor the Induced Fit model can describe their functionality. A third model, the 'New View' model has recently been suggested and states that a protein exists in an ensemble of pre-existing conformations with discrete and similar free energies. Amongst them is the structure of the bound conformation. The actual binding of the ligand induces a shift in the equilibrium of existing conformations towards the bound conformation and causes the protein to appear well structured in a X-ray crystal (James and Tawfik, 2003).

The standard method for measuring the flexibility of enzyme binding sites is to calculate the *Root Mean Square Deviation* (RMSD) between different conformations of the binding site. The RMSD is calculated between the Cartesian coordinates of all atom pairs between both proteins *a* and *b* using the following formula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^N [(x_{ia} - x_{ib})^2 + (y_{ia} - y_{ib})^2 + (z_{ia} - z_{ib})^2]}{N}}, \quad (2.15)$$

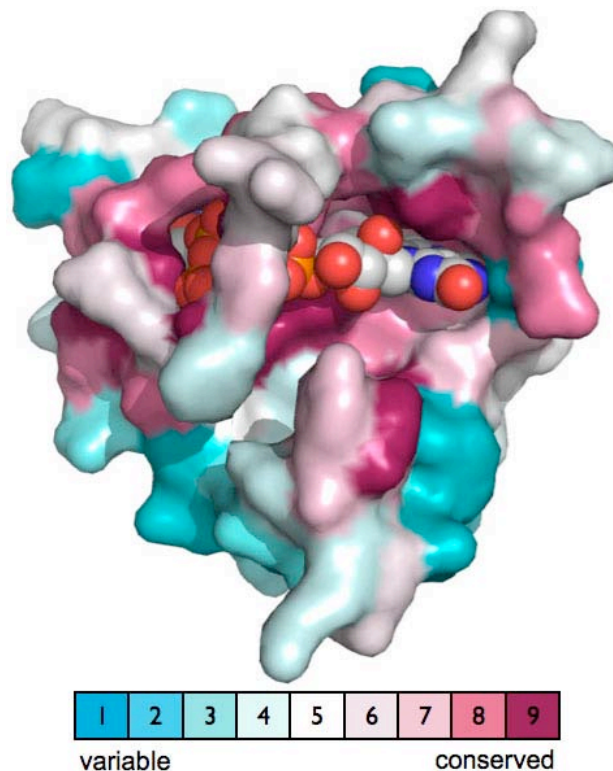
where *x*, *y*, *z* are the Cartesian coordinates of the protein atoms and *N* is the number of compared atoms. Depending on the scientific question or the available data one can calculate the RMSD of all atoms, of all residue side chain atoms, or of only the C $\alpha$  atoms between two structures. STRuster (Domingues, et al., 2004) is a web service for qualitative measurement of protein flexibility. Its algorithm analyzes an ensemble of different conformations of a protein by calculating the Euclidean distances between all residues in each conformation. The Euclidean distances are summed up for each conformation and plotted in an 'all-conformation vs. all-conformation' distance matrix. The distance matrix is used to cluster each conformation according to its level of flexibility and group similar conformations.



### 2.3.2.5 Conservation

Another characteristic of enzyme binding sites is that the residues forming the sites tend to be strongly conserved within protein families (see Figure 2.3f). Residues forming binding sites and especially catalytic residues in active sites are amongst the most important residues in an enzyme (Chelliah, *et al.*, 2004; Lichtarge, *et al.*, 1996). Any mutations to these residues could prohibit the enzyme from recognizing its ligand or catalysing its chemical reaction and thus lead to a loss of the protein's function. Most often binding site residues are either polar or charged (up to 70% of residues are Arg, Asp, Cys, Glu, His and Lys) (Bartlett, *et al.*, 2002).

ConSurf (Glaser, *et al.*, 2003) calculates the conservation of each amino acid in a protein sequence using the evolutionary trace method (Lichtarge, *et al.*, 1996). This method first retrieves homologous sequences for the sequence of the protein structure from a protein sequence database and runs a multiple sequence alignment on this set of homologous



**Figure 2.6: Conservation scores mapped on a protein structure by ConSurf.**

The protein is that of PDB entry 1p4m. Note the higher conservation in and around the binding sites.

sequences. In the second step, the method uses the alignment to compute a phylogenetic tree, which represents the evolutionary relationship of the homologous sequences. Next, the homologous sequences are divided into groups and subgroups based on the branches of the tree followed by an analysis on the frequency of residue changes in each subgroup. In the analysis, each residue is assigned a rank that reflects the position in the tree at which the residue becomes invariant in the succeeding subgroups. In the final step the residues of the protein structure are divided into 9 classes according to their rank with '1' being the least conserved and '9' being the most conserved residue and colour coded on the structure (see Figure 2.6). A visual inspection of the colour coded protein structure can help to identify clusters of highly conserved residues on the protein surface that might indicate the location of binding sites.

### **2.3.2.6 Interaction energy**

The process of molecular binding requires in the first instance shape complementarity to allow ligand atoms to approach binding site atoms. The proximity between both binding partners is important as their binding energy depends very much on the distances between their atoms. One theory about electrostatic complementarity between binding sites and ligands suggests that electrostatic potentials at binding sites are strong enough to attract the ligand from the solvent into the active site. This assumption has been derived from enzymes that have catalysis rates approaching the diffusion limit, like the copper-zinc-superoxide-dismutase protein family. This protein family exerts a positive electric field over the active site, which attracts negatively charged oxygen radicals towards the active site copper ion (Livesay, et al., 2003). The visualization of the electrostatic potentials mapped on the structure surface is particular useful for identifying DNA binding sites (see Figure 2.3g). Many DNA binding proteins possess a large patch of positively charged amino acids on their surface to electrostatically attract their negatively charged binding partner (Tsuchiya, et al., 2004). In a study where organic solvent molecules were computationally mapped on the protein surface

to predict potential binding sites of ligands, it was found that hydrophobic patches are also important within binding sites, inducing organic solvents to cluster therein (Silberstein, et al., 2003). The results are in agreement with earlier experiments which showed that binding affinities of ligands can be increased by promoting hydrophobic interactions between binding sites and ligands (Davis and Teague, 1999). Since ligand molecules do not bind at random sites on a protein structure, their binding sites should feature particular high binding energies towards their cognate ligand.

Q-SiteFinder (Laurie and Jackson, 2005) calculates the potential binding energies on a protein surface and detects energetically favourable surface patches that may present ligand binding sites. The favourable patches are found by placing the protein in a grid and rolling a probe sphere along the grid points over the molecular surface. At each grid point an energy function (see Figure 2.3h), which incorporates van der Waals potential, electrostatic potential and hydrogen bond potential, is applied to the probe sphere. Grid points that exceed a predefined energy threshold are clustered if they are below a certain separation. For each cluster, the single interaction energies of the grid points are summed up and ranked according to their total interaction energy. The cluster with the most favourable interaction energy is then identified and considered as a potential binding site.

## 2.4 Intermolecular forces

Intermolecular forces (Israelachvili, 1991a) are attractive or repulsive forces acting upon distinct molecules. They are the result of valence electrons that redistribute as molecules approach each other. Compared to covalent bonds, intermolecular forces are of long range with energies between 0.5-20 kcal/mol.

According to the *Hellman-Feynman theorem*, all intermolecular forces are essentially electrostatic in origin (Israelachvili, 1991b). However it is common to approximate the intermolecular forces using three seemingly different categories of interaction types (Pliska, 2001) that vary in their range of distance (Israelachvili, 1991d). *Purely electrostatic forces* are long-range forces that act between charged or permanently polarized atoms/molecules and follow Coulombs law. The interaction energies of electrostatic forces fall off with  $r^{-2}$ , where  $r$  is the distance between two atoms/molecules. Of medium range ( $r^{-4}$ ) are *induction* or *polarization forces* that arise between charged or permanently polarized atoms and nearby non-polar atoms/molecules, the former inducing with its electric field a dipole moment on the latter. The third class of forces, also called *London dispersion forces*, are found uniformly between all atoms and molecules. Although their interaction energies can only be calculated accurately with quantum mechanics, they are often approximated with a power function of  $r^{-6}$ . Due to their low interaction energies, London dispersion forces prevail only in interactions between non-polar atoms/molecules. The physical origin of London dispersion forces lies in the appearances of *instantaneous dipoles* formed by the constant motion of electrons around their atom nuclei. The instantaneous dipoles induce in turn a dipole on the neighbouring atom/molecule and produce an attractive induction force (Leach, 2001a). Induction and dispersion force combined with the *repulsion force* that hinders the overlap of electron orbits of different atoms is also known as *van der Waals force*. Hydrogen bonds and hydrophobic interactions do not conform to the above classification and will be treated separately in section 2.4.3.

## 2.4.1 Electrostatics *in vacuo*

Electrostatics is a specific branch in physics that deals with electric phenomena of systems with resting/static charges (Feynman, *et al.*, 1989b).

### 2.4.1.1 Electrostatic potential

All physical quantities in electrostatics can be calculated once the electrostatic potential at every point in space is known. The *electrostatic potential*  $\phi(r)$  is a function of the distance  $r$  and given as a scalar field. It is defined as the *electrostatic potential energy*  $U(r)$  per unit test charge  $q_e$ :

$$\phi(r) = \frac{U(r)}{q_e} = \frac{Q}{4\pi\epsilon_0} \frac{1}{r} = - \int_{\infty}^r \mathbf{E} \cdot d\mathbf{s} \quad , \quad (2.16)$$

where  $\epsilon_0$  is the *dielectric constant* of empty space, measuring the polarisability of real vacuum in an electric field.  $U(r)$  is the work or the energy to carry a test charge  $q_e$  against the electrostatic field  $\mathbf{E}$  of the point charge  $Q$  along a path  $s$  to a location  $r$  distance away from  $Q$ . As a reference, the starting point of the path is set to infinity. The electrostatic field, a vector field, can be calculated from the partial derivatives of the electrostatic potential  $\phi$  in  $x, y, z$  direction:

$$\mathbf{E} = - \left( \frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z} \right) = -\nabla\phi \quad , \quad (2.17)$$

where the nabla symbol  $\nabla$  symbolises the vector differential operator *del*:

$$\nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (2.18)$$

that measures how fast the potential varies as the coordinates in three dimensions change.

### 2.4.1.2 Poisson equation

The scalar product of  $\nabla$  with the electrostatic field  $\mathbf{E}$  in equation ( 2.17 ) results in a scalar field (Wong, 1991) that describes the *charge density*  $\rho$  at position  $x,y,z$  (Feynman, *et al.*, 1989a):

$$\nabla \cdot \nabla \phi = \nabla^2 \phi = -\frac{\rho}{\epsilon_0} \quad , \quad (2.19)$$

where  $\nabla^2$  is the differential operator *Laplacian* and the charge density  $\rho$  is given by

$$\rho = \frac{\partial^3 q}{\partial x \partial y \partial z} \quad . \quad (2.20)$$

Equation ( 2.19 ) is a partial differential equation of second order and generally known as the *Poisson equation*. The entire subject of electrostatics can be reduced to find solutions to Poisson's equation ( 2.19 ). The electrostatic field  $\mathbf{E}$  for example, can be obtained by differentiating the solution to Poisson's equation according to equation ( 2.17 ).

## 2.4.2 Electrostatics in a dielectric medium

The interaction between molecules in aqueous solution differs from those in free space, where the total interaction energy is governed only by the mutual interaction between the molecules. In aqueous solution however, each molecule also interacts with the surrounding water molecules (Israelachvili, 1991c). Water molecules are distinct from organic molecules in their electric properties; they have a high dipole moment and are relatively unconstrained in their motion allowing the molecules to align instantaneously to an electric field and shield/weaken charge-charge interaction (Gilson, 2000). As a result, the dielectric constant of water is relatively high and generally assessed at  $\epsilon = 78$  (Brucoleri, *et al.*, 1997). Proteins on

the other hand usually have a low dielectric constant between 1 and 4. Although a polypeptide chain also consist of a sequence of polar amide groups, their dipolar re-orientation is restricted by their immobilisation within the protein structure. On the exposure to an electric field, the polypeptide chain only responds with an electronic polarization (Gilson, et al., 1985) that was measured for condensed media to be around  $\epsilon = 2$  (Gilson, 2000).

### 2.4.2.1 Protein electrostatics in aqueous media

The calculation of the electrostatic potential in and around a protein requires a simplistic view of the protein-solvent system. An explicit model of the system, where the protein and each surrounding water molecule are simulated as single molecular entities, is still computationally expensive. As an alternative, implicit models have been implemented that treat water as a continuum, offering computational speed for the price of reduced computational accuracy. The *continuum electrostatic model* describes a protein as an arbitrarily shaped complex that consists of a set of spherical atoms with partial charges defined by a force field. The interior of the protein is treated as a homogenous low dielectric medium, which is surrounded by a continuum of a high dielectric aqueous solvent.

The electrostatic potential in a continuum electrostatic model can be calculated by solving Poisson's equation ( 2.19 ) for the model. The charge distribution  $\rho(r)$  is given by the atomic coordinates of the protein. However, in contrast to Poisson's equation *in vacuo*, the dielectric constant of free space  $\epsilon_0$  has to be scaled with the dielectric constant function  $\epsilon(r)$ , that depending on  $r$ , adopts the dielectric constant either of water or of the protein:

$$\nabla^2 \phi(r) = -\frac{\rho(r)}{\epsilon(r)\epsilon_0} \quad , \quad (2.21)$$

$$\nabla \cdot [\epsilon(r)\nabla\phi(r)] = -\frac{\rho(r)}{\epsilon_0} \quad . \quad (2.22)$$

## 2.4.2.2 Protein electrostatics in ionized aqueous media

Molecules under physiological conditions interact not just with water and other molecules. The cytoplasm surrounding the proteins in the cell is enriched with dissolved *electrolytes* such as sodium ions ( $\text{Na}^+$ ), chloride ions ( $\text{Cl}^-$ ), magnesium ions ( $\text{Mg}^{2+}$ ) etc. (Alberts, *et al.*, 1994a). Depending on the charge of the electrolyte and the charge on the protein surface, electrolytes will either be attracted to or repelled from the protein. In the former case, the concentration of electrolytes surrounding the charge on the protein will be higher, whilst in the latter case the concentration will be lower as compared to the electrolytes' bulk concentration far away from the protein. The charge distribution of electrolytes  $\rho_e(r)$  in space follows the Boltzmann distribution and depends on the charge magnitude of the electrolyte  $q_i$ , its bulk concentration  $\rho_{\text{bulk}}$  and the electrostatic potential  $\phi(r)$  of the protein at the point  $r$ :

$$\rho_e(r) = q_i \cdot \rho_{\text{bulk}} \cdot e^{-\frac{\phi(r) \cdot q_i}{kT}}, \quad (2.23)$$

where  $k = 1.38 \times 10^{-23}$  J/K is the Boltzmann factor and  $T$  is the temperature of the medium given in Kelvin (K).

### 2.4.2.2.1 Poisson-Boltzmann equation

Electrolytes that are positively or negatively charged create their own electrostatic field. The same field influences the electrostatic field of the protein, when electrolytes are accumulating around the protein. For an accurate model of the electrostatics in and around a protein, the charge distribution of the electrolytes must be included at the right hand side of Poisson's equation ( 2.22 ). The resulting equation is the *Poisson-Boltzmann equation* (Gilson, 2000; Honig and Nicholls, 1995; Leach, 2001c):

$$-\varepsilon_0 \nabla \cdot [\varepsilon(r) \nabla \phi(r)] = \rho(r) + \sum_{i=1}^n q_i \rho_i^{\text{bulk}} e^{-\frac{\phi(r) \cdot q_i}{kT}}, \quad (2.24)$$



where the sum in the second term on the RHS is over the various types of electrolytes that surround the protein structure. The equation can be rewritten using the relation between hyperbolic and exponential functions (Kreyszig, 1999c)

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad , \quad (2.25)$$

which for binary electrolytes such as NaCl with a single positive and negative charge at 1:1 concentration becomes the *Nonlinear-Poisson-Boltzmann equation*

$$-\varepsilon_0 \nabla \cdot [\varepsilon(r) \nabla \phi(r)] - \rho_{\text{bulk}} 2 \sinh\left(\frac{\phi(r)}{kT}\right) = \rho(r) \quad . \quad (2.26)$$

For small  $x$  values, the hyperbolic sine function can be approximated with  $\sinh \approx x$  allowing for relatively small electrostatic potentials  $\phi(r)$  to express equation ( 2.26 ) as the *linear-Poisson-Boltzmann equation*:

$$-\varepsilon_0 \nabla \cdot [\varepsilon(r) \nabla \phi(r)] - \rho_{\text{bulk}} 2 \frac{\phi(r)}{kT} = \rho(r) \quad . \quad (2.27)$$

This linear approximation leads to significant errors for systems with high electrostatic potentials such as polynucleotic DNA and RNA or polylysine molecules, and requires the application of the computationally more expensive nonlinear-Poisson-Boltzmann equation ( 2.26 ).

### 2.4.3 Water and the hydrophobic effect

Water as a liquid has several unique features that have been essential for the evolution of life on earth. Probably, the two most important properties of water are first its high melting and boiling temperature despite its tiny size and second the location of its highest density at 4°C

causing ice layers on lakes to float on the water surface and preserve life underneath. The next sections will introduce the reasons for these properties and the importance of the effect they have on the interaction between molecules in aqueous solution (Israelachvili, 1991e).

### 2.4.3.1 Hydrogen bond

A single water molecule consists of two hydrogen atoms attached to a divalent oxygen atom. The high electronegativity of the oxygen atom pulls electrons from each hydrogen atom towards the oxygen nucleus inducing a high dipole moment with partial negative charge  $\delta^-$  on the oxygen and a partial positive charge  $\delta^+$  on each hydrogen atom. In bulk solution, two water molecules form a *hydrogen bond*, i.e. two partial negative charge oxygen atoms from two water molecules interact with each other via a positive charged hydrogen atom. The intermolecular distance between hydrogen and oxygen atom is around 1.8 Å, which is shorter than the sum over the oxygen and hydrogen van der Waals radius (1.5 Å + 1.0 Å = 2.5 Å). This in addition to the high directional dependence causes hydrogen bonds to have weak covalent bond characteristics. However, by nature a hydrogen bond is considered as an electrostatic interaction with binding energies of around 0.25 to 1 kcal/mol. A single water molecule can form up to four hydrogen bonds in a tetrahedral arrangement, where the two lone pairs of its oxygen each accepts a single hydrogen atom from two neighbouring molecules and where its two hydrogen atoms are donated to two other neighbouring water oxygen. The result is a dense mesh or network of water molecules that are heavily interconnected. The high melting and boiling temperature of water, which was mentioned above, is due to the total hydrogen bond energy that is stored in the tight hydrogen bond network of bulk water. The highest density of water at 4°C on the other hand is caused by the increase of the number of hydrogen bonds per water molecule from four in the solid state to an average of five in the liquid state (Israelachvili, 1991e).

### 2.4.3.2 Hydrophobic effect

A *hydrophobic*, i.e. non-polar, molecule lacks polar or charged groups to form hydrogen bonds with the water molecules when it is solvated. Without the potential to form hydrogen bonds, the surrounding water molecules form a clathrate cage-like structure around the molecule. In the cage-like conformation, the water molecules preserve most of their hydrogen bonds in the water network. However, the cage formation demands water molecules to reorientate and take on a well-ordered structure, both of which are entropically unfavourable actions that result in the insolubility of hydrophobic molecules in water (Israelachvili, 1991e).

### 2.4.3.3 Partition coefficient $\log P$

The hydrophobicity of an uncharged molecule can be experimentally assessed by calculating the natural logarithm of the degree of its partition in a mixed water and organic solvent solution. The measure for the degree of the partition is  $\log P$ . *n*-octanol is most often used as a reference for the organic phase, beside cyclohexane. The  $\log P$  value will increase with the number of non-polar groups in the molecules as an effect of the group's preference to reside densely packed within the organic phase (Fersht, 1999). It has been observed that the  $\log P$  of a molecule made up of various substituents is the sum over each substituent's  $\log P$  value. The additive characteristics is exploited by computational tools to calculate the  $\log P$  for any molecule from a library of fragments with empirical determined  $\log P$  values (Ghose, et al., 1998). More recently, new approaches with atom-additive methods have been developed that, similar to force fields, assign each molecule atom an atom type with an associated atomic  $\log P$  value and apply correction factors to account for intramolecular interactions (Wang, et al., 2000a). The hydrophobicity of charged molecules can be more appropriately assessed with the distribution coefficient  $\log D$  that takes into account all neutral and charged forms of the molecule and varies with pH of the aqueous phase.

$\log P$  and  $\log D$  are important physicochemical descriptors for drug discovery programs, where both are used excessively in Comparative Molecular Field Analyses (CoMFA) (Testa, et al., 1996). Furthermore, they occur in Lipinski's rule of five to assess the 'druglikeness' of lead compounds as oral-active drugs (Lipinski, et al., 2001) and guide the judgement of a drug candidate's Absorption, Distribution, Metabolism, Excretion and Toxicity properties (ADME-Tox) (Tetko, et al., 2006).

#### **2.4.3.4 Hydrophobic interaction**

The hydrophobic effect on two non-polar molecules solvated in water causes both molecules to approach and bind each other. The union of both molecules is entropically favourable as it reduces the total surface area accessible to the water molecules (by the size of the interface area that the bound molecules share with each other) and the number of water molecules needed to form the cage-like structure around both molecules (Israelachvili, 1991e). The accumulation of non-polar molecules is further favoured by dispersion forces (see introduction to section 2.4) between the molecules (Fersht, 1999) and a pressure imbalance that occurs when water molecules are displaced from the interface of approaching hydrophobic surfaces (Ball, 2008).

## **2.5 Surface and shape of molecules**

*Shape* is fundamental to many processes in biology and chemistry. Comparative zoologists have demonstrated in the early 20<sup>th</sup> century that the shape and size of organisms is formed not only by evolutionary processes but also by mechanical constraints (McMahon, 1973; Thompson and Bonner, 1992). The shape description of macroscopic objects, such as animal bones or protein crystals, is straightforward as they usually have discrete size and shape.

However, for atoms and molecules the concept of a discrete shape breaks down (Mezey, 1993). At the nanoscopic scale, the shape of atoms and molecules is defined by the electrons that spin around their nuclei. The location of electrons however can only be defined with continuous cloud-like fuzzy electron distributions that prevent the determination of a discrete shape for an atom or molecule. A further complication arises for flexible molecules that adopt myriad conformations under physiological conditions (see Chapter 3). An approximation to both problems that often suffice for molecular mechanics purposes is to treat the atoms in molecules as spherical balls that are fixed at a low energy conformation with sphere radii equal to the element's *van der Waals radius*. The van der Waals radius, also called hard sphere radius, of an atom can be defined as the distance at which the force applied to a second atom to approach the first atom becomes repulsive.

## 2.5.1 Molecular surface description

Given the approximation of spherical balls for atoms, the surface of a molecule is often represented as the *van der Waals surface*, *molecular surface* or the *solvent accessible surface* (Richards, 1977). The van der Waals surface is the union of covalently bound spherical atoms in a molecule, where each atom type has a specific van der Waals radius. Molecular and solvent accessible surfaces are calculated by rolling a probe sphere over the van der Waals surface of a molecule. The inward-facing surface of the probe sphere produces the molecular surface, which is a smooth version of the van der Waals surface and often used to visualise the complementarity between molecules. The solvent accessible surface on the other hand is computed by tracing the centre of the probe sphere, thereby highlighting areas that are accessible to the whole probe sphere (Lee and Richards, 1971). The radius of the probe sphere influences the appearance of both surfaces: a smaller probe sphere will emphasise more details, whereas a larger probe sphere will show gross surface characteristics. Usually the average radius of a water molecule with 1.4 Å is used as the probe sphere radius.

Different representations exist for the visualisation of the above surface models. *Voronoi tessellation* visualizes molecule surfaces (Richards, 1974) by dividing the space occupied by all atoms on the protein surface into a set of polyhedra that are referred to as *Voronoi cells*. Each atom corresponds to the cell's centroid. The space between neighbouring centroids is divided such that any region within the cell is closer to the cell centroid than to any other centroid. A related tessellation is the *Delaunay triangulation* that divides the protein surface into a set of triangles. *Delaunay triangulation* can be obtained from Voronoi tessellation by connecting all centroids with a common Voronoi cell edge (Poupon, 2004). Connolly's *dot surface* spreads dots, which are tangent to a single atom or a set of atoms, over the molecule thereby allowing a transparent view of the molecule (Connolly, 1983b). Another method to visualise molecular surfaces splits up the solvent accessible surface into concave spherical triangles, saddle shape rectangles and convex spherical regions (Connolly, 1983b). Each of the three surface pieces is analytically described and used to calculate the surface area and the volume of the molecule (Connolly, 1983a). Furthermore, the van der Waals surface of proteins and other molecules can be visualized by contouring a density grid according to a convolution between the atom coordinates and an atom centric Gaussian function (Grant, et al., 1996; Grant and Pickup, 1995). The Gaussian function is given by:

$$\rho = \rho_0 e^{-kr^2} \quad , \quad (2.28)$$

with the density value  $\rho$  at a grid point  $r$  distance away from the atom centre, the density  $\rho_0$  at the atom centre and a constant  $k$  (Laskowski, 1995; Stockwell, 2005).

Surface representations are in particular important for the visualisation of various physicochemical properties of molecules. For example, the electrostatic potential and the hydrophobicity of a molecule can be calculated on the molecule's surface and represented using a colour scheme (Kinoshita, et al., 2002). Such surface representations can be informative in function prediction methods e.g. where large patches of highly positive electrostatic potentials on protein surfaces can sometimes indicate DNA binding sites. In

Chapter 4, the electrostatic potential and the hydrophobicity of a protein will be mapped on the van der Waals surface of its ligand to assess their complementarity.

## 2.5.2 Molecular surface comparison

Although the visualization of molecular surfaces is well established, their comparison is just the opposite. Molecular surface comparisons have some advantage over atom coordinate comparisons. Firstly, they are independent of the atomic constellation that lies beneath the surface, and secondly they can compare the surfaces of binding sites and ligands, and thus are appropriate for small molecule docking studies. Although many methods have been published in the literature (for a comprehensive overview see (Hofbauer, et al., 2004; Pickering, et al., 2001; Via, et al., 2000) and references therein), most of them are similar in approach. In broad terms surface comparison methods discriminate surfaces either by local or global features. Most methods for local feature comparison implement various forms of graph matching (Kinoshita, et al., 2002; Pickering, et al., 2001), geometric hashing (Rosen, et al., 1998) or compare local and global curvatures on the surface of molecules (Cosgrove, et al., 2000; Exner, et al., 2002). Although the comparison of global features fails to discriminate local dissimilarities, they are usually faster by several orders of magnitude and are therefore ideal for fast surface comparisons in large databases (Iyer, et al., 2005). Among the global feature comparison methods are those that compare a set of sorted distances capturing the shape and the physicochemical properties on molecular surfaces (Ballester and Richards, 2007; Binkowski and Joachimiak, 2008; Weisel, et al., 2007). Others calculate distances between expansion coefficients of Fourier descriptors (Cai, et al., 2002a; Perez-Nueno, et al., 2008; Ritchie and Kemp, 2000) or compare three-dimensional moments (Mansfield, et al., 2002) and moment invariants (Sommer, et al., 2007).

### 2.5.2.1 Graph matching

Any set of points can be represented by a graph, with each point being a graph node and the distance vectors between the nodes being the edges. In the case of surface comparison, the points correspond to vertices or dot points on the molecular surface. In order to compare two surfaces with each other, their graphs are compared and all nodes that have a similar spatial location and/or physicochemical property are extracted to form a new *association graph*. Given such an association graph, the best match between both surfaces corresponds to the maximum clique in the association graph, i.e. the largest subset of nodes that are all connected with each other in a pair wise manner (Bron and Kerbosch, 1973). This problem can be computationally demanding since every additional node increases the computation time by  $N^2$ . For protein surfaces in particular the computation is unfeasible as a surface can have more than 10.000 vertices/nodes (Kinoshita, et al., 2002). To reduce the complexity the *eF-site* database (Kinoshita, et al., 2002) divides the protein surface into sub-surfaces that are a maximum of 12 Å large and limit the size of the association graph to a maximum of 2000 nodes. Pickering and coworkers reduce the complexity by comparing only binding site surfaces with each other (Pickering, et al., 2001). In any case, once the maximum clique is found a similarity score can be calculated that relates the size of the maximum clique to the number of nodes in the smaller molecular surface.

### 2.5.2.2 Geometric hashing

The geometric hashing algorithm (Brakoulias and Jackson, 2004; Rosen, et al., 1998) for molecule surfaces consists of two operational stages: A preliminary pre-processing stage and a recognition stage. The preliminary stage itself consists of four steps that create a database of hash tables, where each hash table represents a single molecule surface:



1. Set an orthogonal 3D reference coordinate system on the surface such that a selected triplet of three vertices, being non-collinear to each other, lie in the  $xy$ -plane of the coordinate system.
2. Determine the spatial position of each remaining vertex on the surface according to the reference-coordinate system.
3. Store information about the identity of the triplet, the location of each remaining vertex (quadruplet) and if required physicochemical properties calculated at the fourth vertex in a hash table.
4. Repeat step 1-3 for all other triplet combination on the surface.

Once the database of hash tables is built, the recognition stage can begin by applying the same approach as above to a query molecule surface. However, instead of storing the quadruplets in a hash table, they are checked for their existence in the database. If a hash table exceeds a user-defined minimum hit value, the molecules are aligned to each other and a heuristic iterative matching algorithm is employed to expand the number of matching vertices in the neighbourhood of the hit.

Similar to graph matching, geometric hashing is computationally too demanding to be executed on each vertex of a molecular surface and requires a reduction in the number of vertices. In the implementation of Nussinov and coworkers (Rosen, et al., 1998), the reduction is achieved by replacing all vertices on convex, concave or saddle-shaped surface faces by a single sparse critical point. Further reduction of complexity is realized by considering only reference coordinate systems that are set by vertices satisfying a minimum and maximum distance constraint.

### 2.5.2.3 Spherical harmonics

Fourier descriptors with spherical harmonic basis functions are well suited for the description,

visualisation and comparison of molecular shapes (Leicester, et al., 1994). It is important to mention that Fourier descriptors are not comparing surfaces but rather the shape of molecules. In the framework of this thesis the *shape* of a molecule is defined as follows:

*The shape of a molecule is the geometric property of the volume, which the molecule occupies in space and is independent of the molecule's location, orientation and scale. Two molecules are said to have the same shape if and only if after appropriate translation, rotation, and scaling operations both molecules overlap with their entire volumes. The similarity of the shapes decreases with decreasing overlap of the volumes.*

For reviews about state-of-art shape comparison techniques, see (Iyer, et al., 2005; Veltkamp, 2001; Zhang and Lu, 2004).

Spherical harmonic functions  $Y_{lm}$  are probably best known as the orbital shape determining functions and solutions to the angular part of Schrödinger's equation for a one-electron atom. In general, spherical harmonics functions  $Y_{lm}$  satisfy any *Laplace equation* on the surface of a sphere, i.e. in a spherical coordinate system:

$$\nabla^2 S(r,\theta,\phi) = 0 \quad , \quad (2.29)$$

with

$$S(r,\theta,\phi) = R(r) Y_{lm}(\theta,\phi) \quad (2.30)$$

making them essential in physical problems involving partial differential equations within electromagnetism, gravity, mechanics or hydrodynamics. Their attractive properties when dealing with rotations, spherical averaging procedures or smooth surface representations on the sphere have led to their extensive use in protein crystallography. For example, in molecular replacement, spherical harmonic functions have been used as rotation functions

(Navaza, 2006) to align unknown with known protein structures (see section 2.2.1.1.5) or as means to calculate the generalized scattering factor (see section 2.2.1.2.2) of bound atoms with perturbed electron densities (Maslen, et al., 2006). Furthermore, they have been applied as shape descriptor in docking studies for protein-protein interaction (Ritchie and Kemp, 1999; Ritchie and Kemp, 2000), virtual screening of protein-ligand interactions (Cai, *et al.*, 2002a; Yamagishi, *et al.*, 2006) and in induced-fit flexible docking studies (Yamazaki, et al., 2009). As a visualisation technique, the same functions have been utilised to describe and compare various geometrical and physicochemical properties of molecular surfaces on proteins and small molecules (Dlugosz and Trylska, 2008; Duncan and Olson, 1993; Lin and Clark, 2005; Max and Getzoff, 1988).

Spherical harmonics  $Y_{lm}(\theta, \phi)$  are smooth i.e. infinitely differentiable, complex functions of two angle variables  $\theta$ ,  $\phi$  and two indices  $l$  and  $m$ . In quantum mechanics terminology  $l$  is the angular quantum number that runs from 0 to  $\infty$  and determines for an electron its angular momentum and the number of local minima in  $Y_{lm}$ .  $m$  on the other hand is the magnetic quantum number that runs from  $-l$  to  $l$  and determines how the electron moves in a magnetic field. The spherical harmonic functions  $Y_{lm}(\theta, \phi)$  can be factorized into a  $\theta$ -dependent term, which are the associated Legendre polynomials  $P_{lm}$  and into a  $\phi$ -dependent term, which is a complex-exponential function:

$$Y_{lm}(\theta, \phi) = N_{lm} P_{lm}(\cos\theta) e^{im\phi} \quad , \quad (2.31)$$

with  $N_{lm}$  being the normalization factor:

$$N_{lm} = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} \quad . \quad (2.32)$$

Illustrations and a list of functional values for low order spherical harmonics can be found at (Weisstein, 2009b).

### 2.5.2.3.1 Associated Legendre polynomials

*Associated Legendre polynomials* satisfy the  $\theta$ -dependent term in Laplace's equation and are generally defined as

$$P_{lm}(x) = (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \quad , \quad (2.33)$$

where  $P_l(\theta)$  are the *ordinary Legendre Polynomials* that are solutions to the ordinary Legendre's differential equation. *Rodrigues' formula* given by:

$$P_l(x) = \frac{1}{2^l l!} \left( \frac{d}{dx} \right)^l (x^2 - 1)^l \quad , \quad (2.34)$$

can be used to derive  $P_l(\theta)$  of any degree  $l$ , where the degree is the value of the polynomials highest power.

It can be shown (Kreyszig, 1999a) that the ordinary and associated Legendre polynomials are both a *complete set* of *orthogonal* functions in the interval  $-1 \leq x \leq 1$ , i.e.:

$$\int_{-1}^1 P_{lm}(x) P_{nm}(x) dx = \frac{(l+m)!}{(l-m)!} \int_{-1}^1 P_m(x) P_n(x) dx = \begin{cases} \frac{2(l+m)!}{(2n+1)(l-m)!} \quad , & \text{for } n = l \\ 0 \quad , & \text{for } n \neq l \end{cases} \quad (2.35)$$

where the term complete refers to a sufficiently amount of Legendre polynomials to represent all continuous functions in the interval  $-1 \leq x \leq 1$ . Both properties of the Legendre polynomials, orthogonality and completeness, are important for the purpose of this thesis as both allow the expansion of any function in the interval  $-1 \leq x \leq 1$  in term of Legendre polynomials.

### 2.5.2.3.2 Spherical harmonics expansion

Although spherical harmonic functions  $Y_{lm}$  are by definition complex functions, only the real part of the functions will be used in the course of this thesis and shall be referred as *surface spherical harmonics functions*  $S_{lm}$ .  $S_{lm}$  can be obtained by a linear combination of  $Y_{lm}$  with its complex conjugate  $Y_{l-m}$ . Three different cases arise depending on the sign of the magnetic quantum number  $m$  (Cai, et al., 2002a):

$$S_{lm}(\theta, \phi) = \frac{1}{\sqrt{2}}(Y_{lm} + Y_{l-m}) = \sqrt{2}N_{lm}P_{lm}(\cos\theta)\cos(m\phi) \quad \text{for } m > 0 \quad , \quad (2.36)$$

$$S_{l0}(\theta, \phi) = Y_{l0} = N_{l0}P_l(\cos\theta) \quad \text{for } m = 0 \quad , \quad (2.37)$$

$$S_{l-m}(\theta, \phi) = \frac{1}{i\sqrt{2}}(Y_{lm} - Y_{l-m}) = \sqrt{2}N_{lm}P_{lm}(\cos\theta)\sin(m\phi) \quad \text{for } m < 0 \quad . \quad (2.38)$$

Note, that for  $m = 0$  the associated Legendre polynomial  $P_{lm}(\cos\theta)$  reduces to the ordinate Legendre polynomial  $P_l(\cos\theta)$ . The normalisation factors  $N_{lm}$  guarantees that all  $S_{lm}$  are orthonormal.

Their orthonormal property makes surface spherical harmonics  $S_{lm}(\theta, \phi)$  well suited for the Fourier analysis of continuous single-valued radial function  $f(\theta, \phi)$  on a unit sphere. Under the assumption that the shape of a molecule is *star-like*, i.e. a ray from the centre of mass to the molecule's surface penetrates the surface only once, the shape can be described by a radial function and thus be expanded with surface spherical harmonics. For the expansion,  $S_{lm}(\theta, \phi)$  form a set of basis functions in the generalized Fourier series (Kreyszig, 1999b):

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l c_{lm} S_{lm}(\theta, \phi) \quad , \quad (2.39)$$

where  $c_{lm}$  are coefficients for each surface spherical harmonics, determining the weight of each harmonic on the expansion of the function. The expansion will be exact, if the first sum over the angular quantum number  $l$  is infinite. Terminating the expansion at a certain order  $l_{\max}$  yields an approximation to the function  $f(\theta, \phi)$ , where the approximation improves with increasing  $l_{\max}$  (see later results chapters). The number of coefficients and thus the number of functions for an approximation up to the order  $l_{\max}$  is given by  $(l_{\max} + 1)^2$ . The expansion returns a unique set of coefficients  $c_{lm}$  for different radial functions  $f(\theta, \phi)$  allowing them to be employed as shape descriptors. The first few coefficients have a direct meaning with respect to the geometry of the surface that is described by  $f(\theta, \phi)$ . The first order coefficient  $c_{00}$  corresponds to the weighted radius  $r = N_{00}c_{00}$  where  $N_{00} = 1/4\pi$ , the three coefficients for  $l = 1$  encode the centroid of the surface and the nine coefficients up to the order  $l_{\max} = 2$  define an average ellipsoid (Ritchie and Kemp, 1999). The expansion coefficients can be calculated from equation ( 2.39 ) as the inner product of the radial function  $f(\theta, \phi)$  and the surface spherical harmonics  $S_{lm}(\theta, \phi)$  (Cai, *et al.*, 2002a; Morris, 2006):

$$c_{lm} = \int_0^\pi \int_0^{2\pi} f(\theta, \phi) S_{lm}(\theta, \phi) d\theta d\phi \quad . \quad (2.40)$$

To compare shapes with each other, it is sufficient to compare the expansion coefficients of the shapes with a distance function, e.g. a standard Euclidean metric given by:

$$d(\mathbf{c}^a, \mathbf{c}^b) = \sqrt{\sum_{i=1}^n (c_i^b - c_i^a)^2} \quad , \quad (2.41)$$

where  $\mathbf{c}^a$  and  $\mathbf{c}^b$  are both coefficient vectors with  $n = (l_{\max} + 1)^2$  number of expansion coefficients.

### 2.5.2.3.3 Spherical $t$ -design

Spherical  $t$ -design (Weisstein, 2009a) is a set of  $N$  points  $P = \{p_1, \dots, p_N\}$  on the surface of a unit sphere, which is defined in  $d-1$  dimensions as  $S^{d-1} = \{x = (x_1, \dots, x_d) \in R^d\}$ . All points  $P$  are distributed such that an integral of any polynomial  $f$  of degree  $\leq t$  on the sphere, is equal to the average of the polynomial over the  $N$  points:

$$\frac{1}{\text{Vol}(S^{d-1})} \int_{S^{d-1}} f(x) dx = \frac{1}{N} \sum_{i=1}^N f(p_i) \quad , \quad (2.42)$$

where  $\text{Vol}(S^{d-1})$  denotes the ‘volume’ of the sphere in  $d-1$  dimensions, e.g. the surface area of a three-dimensional unit sphere (Meakin, 1998). Thus, spherical  $t$ -designs replace a full integration over the entire unit sphere with a sampling over a few points on the sphere surface. The sampling is not only computationally fast but also preserves the accuracy of the full integration over the unit sphere, as both sides of the equation (2.42) are equal. So far the existents of spherical  $t$ -designs were only proven for orders up to  $t = 12$ , but numerical evidence suggest that integration layouts for orders up to  $t = 21$  also exist (Morris, 2006).

Within the scope of this thesis, spherical  $t$ -designs have been employed as an integration layout on the molecular surfaces of small molecules and binding pockets for the spherical harmonic expansion (Morris, 2006; Morris, *et al.*, 2005). Their application simplified the computation of the coefficients by replacing the integral in equation (2.40) with a sum over the points  $P = \{p_1, \dots, p_N\}$ :

$$c_{lm} = \sum_P f(\theta, \phi) S_{lm}(\theta, \phi) \quad . \quad (2.43)$$

Spherical  $t$ -designs for various values of  $N$  and  $t$  are available on the world-wide-web from <http://www.research.att.com/~njas/sphdesigns> (Hardin and Sloane, 1996; Stockwell, 2005).

# Chapter 3

## Shape Variation in Protein Binding Pockets and their Ligands

### 3.1 Introduction

Molecular recognition is a central theme in molecular biology and arguably the primary driving force behind most processes in and between cells. The recognition procedure is based mainly on geometric and electrostatic complementarity (Pliska, 2002; Tsai, *et al.*, 2002). Enzymes are thought to have optimised their astonishing catalytic power and specificity by evolving their surfaces to complement substrate transition states. One would expect the co-evolution of substrates and enzymes to result in a fairly exclusive partnership that must somehow be reflected in both the ligand and the binding site. Therefore, it is reasonable to assume that proteins binding similar ligands have binding sites of similar geometrical and physicochemical properties.

A common assumption about the shape of protein binding pockets is that they are related to the shape of the small ligand molecules that can bind there, allowing both binding partners to approach each other and form short-range non-covalent bonds (Fersht, 1974; Grant, *et al.*, 1996; Jones, *et al.*, 1997; Sotriffer and Klebe, 2002). Many computational methods have been developed that make use of this assumption and predict small molecules for given binding sites based on their geometric complementarity to the binding site (Bock, *et al.*, 2007; Katchalski-Katzir, *et al.*, 1992; Tsai, *et al.*, 2001; Via, *et al.*, 2000; Yamagishi, *et al.*, 2006). The entire field of virtual screening and small molecule docking to target proteins is based on



molecular complementarity. Scoring functions that are applied in all methods to evaluate the binding energy of the compound to the binding site, score the geometrical and physicochemical complementarity between both binding partners (Kitchen, *et al.*, 2004). In a study on drug molecule binding sites, it was found that geometrical properties alone are among the most important binding factor for drug molecules. Physicochemical properties did not seem to play any decisive role (Nayal and Honig, 2006). Similar results have been reported for complexes of protein and pharmacological interesting small molecules (Norel, *et al.*, 1999b). However, several studies have reported that the geometrical complementarity is often far from perfect (An, *et al.*, 2005; Liang, *et al.*, 1998). These studies have revealed that ligands are sometimes only partially enclosed by their binding sites with the rest of the ligand exposed to the solvent. Furthermore, even within the enclosed region, contacts between protein and ligand often occur at relatively few points, involving strong hydrogen bond or van der Waals interactions (Babine and Bender, 1997; Cosgrove, *et al.*, 2000; Smith, *et al.*, 2006). It has also been suggested that the displacement of water molecules at binding sites and the retention of the ligand's flexibility in the bound state are both factors that contribute to favourable entropic changes in the binding process (Boehm and Klebe, 1996) and indicate a certain level of non-complementarity between protein and ligand molecules.

Similar opposing conclusions were drawn for the geometric complementarity of protein-protein interfaces. For a review on protein-protein interactions, see (Nooren and Thornton, 2003). Some studies (Gabb, *et al.*, 1997; Shoichet and Kuntz, 1991; Tsuchiya, *et al.*, 2006) have reported the inefficiency of using geometric descriptors for the docking of unbound protein structures and have highlighted the importance of additional physicochemical descriptors for reliable predictions of the protein complexes. In contrast, Norel *et al.* argued that shape complementarity would suffice for docking protein-protein complexes (Norel, *et al.*, 1999a). Jones and Thornton reported that the interface of homodimers and enzyme-inhibitor complexes is generally more complementary than the interface between antibody-antigen complexes. However the average complementarities for both protein complexes were rather low with gap indices of around 2.2 to 3.2 Å (gap index = volume of the gap between both

proteins divided by the interface solvent accessible surface area) (Jones and Thornton, 1996). Recently it has been argued that protein-protein complexes are formed not due to the complementarity at the interface between both binding partners but as a result of thermodynamic processes that are affected by the size of the entire complex and the entropy change due to complex formation (Krissinel and Henrick, 2007).

In this chapter, a global shape descriptor was used from (Morris, *et al.*, 2005) to follow up the analyses on the geometric complementarity between proteins and bound small molecules on a large data set. In the course of this chapter following questions will be addressed:

1. To what extent are binding pockets from non-homologous protein domains that bind the same small molecule similar in shape?
2. To what extent are binding pockets similar in shape to the ligands they bind?
3. Is shape or size more important when comparing binding pockets with ligands?
4. How useful is a global shape descriptor for binding sites in molecular recognition analysis and especially as a ligand predictor?

The importance of binding sites in proteins was recognised early in structural biology and led to many studies to identify and compare binding sites. For a comprehensive list of methods for the determination and comparison of binding sites, see (Bergner and Günther, 2004; Campbell, *et al.*, 2003; Gold and Jackson, 2006a; Sotriffer and Klebe, 2002; Vajda and Guarnieri, 2006; Via, *et al.*, 2000). Current approaches analysing binding sites can be roughly divided into three classes. Firstly, methods that detect cavities and geometrically match them to each other; secondly, methods that identify and compare specific geometrical patterns of amino acids in binding sites; and thirdly methods that use evolutionary information to predict the location of binding sites.

Among the methods in the first category is CavBase (Schmitt, *et al.*, 2002), which uses pseudospheres to represent the locations and physiochemical properties of the atoms

involved in molecular recognition. The spatial distribution of the pseudospheres is represented by a graph, and a clique detection algorithm is used to identify similar binding sites in other protein structures (see section 2.5.2.1). The algorithm IsoCleft (Najmanovich, et al., 2007) is based on the same clique detection technique but uses exclusively C-alpha atoms instead of pseudospheres for an initial binding site match, and only in a second stage runs a more demanding all atom comparison between binding sites. The eF-site data base (electrostatic surface of Functional-site) database (Kinoshita and Nakamura, 2003) compares graphs that represent the electrostatic potential, the hydrophobicity and the curvature of local binding site surface patches. The SitesBase server (Gold and Jackson, 2006b) exploits geometric hashing (see section 2.5.2.2) to identify equivalent atoms in binding sites that are of the same element type and occur in similar relative spatial orientations. In addition, it stores precalculated all-against-all similarities for the large majority of PDB binding sites.

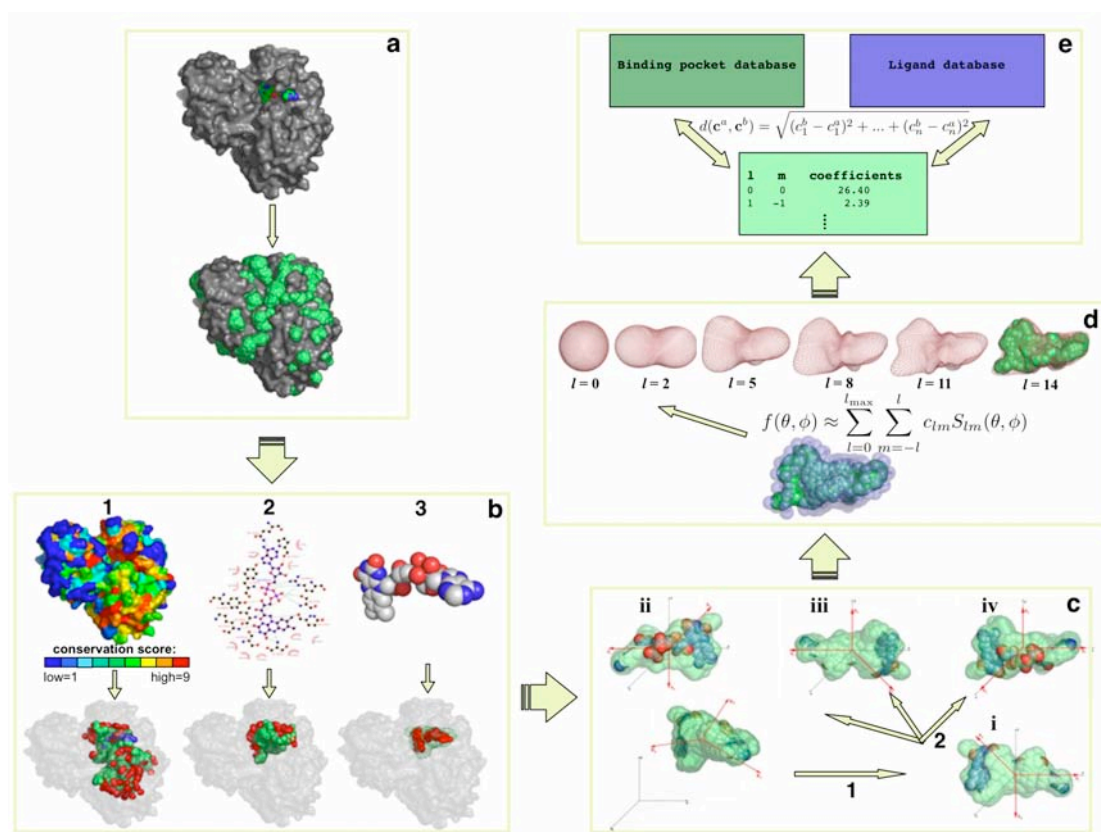
Methods from the second category are based on the fact that functionally important residues tend to maintain the same relative spatial disposition even in distantly related proteins. This is particularly true for the catalytic residues in enzymes. The best-known example is the Ser-His-Asp catalytic triad of serine proteases. In this specific case the relative positioning of these three residues are strongly conserved even in totally different structural folds (Wallace, et al., 1996). The CSA (Catalytic Site Atlas) database (Porter, et al., 2004) contains a catalogue of structural templates of 2 to 6 residues each derived from the catalytic residues of enzymes. 3D search programs like SPASM, RIGOR (Kleywegt, 1999) and Jess (Barker and Thornton, 2003) or algorithm proposed by (Besl and McKay, 1992) and (Nussinov and Wolfson, 1991) allow one to scan such templates against any query protein structure.

Finally, an example of a method from the third class which uses evolutionary information to predict the location of a protein's binding site is pvSOAR (pocket and void surfaces of amino acid residues) (Binkowski, et al., 2003a). pvSOAR uses the CASTp database (Binkowski, et al., 2003b) of protein clefts and voids and searches for similar sequence and spatial arrangement of the cavity residues for a query structure .

All the methods above involve comparison of atomic coordinates in one form or another. In the present work, a shape comparison technique is used. This avoids the problems of superposing binding sites, particularly where they are composed of different numbers of atoms and atom types. Furthermore, the consideration of shape alone allows a direct comparison of the degree of complementarity between binding pockets and ligands they bind. Many sophisticated shape description and matching methods exist, all with their strengths and weaknesses (Iyer, *et al.*, 2005; Veltkamp, 2001; Zhang and Lu, 2004). As the focus in this chapter is on 3D shapes, a method originally pioneered by Ritchie & Kemp in a series of papers (Ritchie, 1998; Ritchie, 2003; Ritchie, 2005; Ritchie and Kemp, 1999; Ritchie and Kemp, 2000) is employed where the idea of using surface spherical harmonics for protein-protein interactions and docking was developed. The idea was taken further by Cai and colleagues and applied directly to binding pockets (Cai, *et al.*, 2002a) and was later improved by (Morris, *et al.*, 2005).

## 3.2 Methods

A Java program named CleftXplorer was implemented with the aim to provide a suite of tools to explore various properties of protein binding sites and small molecules. This chapter will introduce CleftXplorer's geometric shape descriptor that is based on surface spherical harmonic functions. Chapter 4 will describe the implementation of the physicochemical properties.



**Figure 3.1: CleftXplorer algorithm for binding pocket shape description.**

Five step algorithm illustrated on flavo-hemo protein 1cqx (protein structure is grey, ligand is varicoloured, SURFNET spheres are green).

- a. Cleft localisation: Clefs in protein structure are filled with spheres by SURFNET.
- b. Cleft volume reduction: SURFNET clefs are reduced using one of three procedures (atoms used for reduction are in red colour, reduced clefs are in green colour):
  1. Conserved Cleft Model: keep SURFNET spheres next to conserved cleft region (top insert: conservation mapped on protein structure).
  2. Interact Cleft Model: keep SURFNET spheres next to protein atoms that interact with ligand (top insert: LigPlot/HBPLUS diagram).
  3. Ligand Cleft Model: keep SURFNET spheres that are in contact with ligand molecule (top insert: FAD molecule).
- c. Transformation (Cartesian coordinate axes are black coloured, eigenvectors of moment of inertia are red coloured):
  1. Transform reduced cleft to the coordinate origin and rotate according to moment of inertia.
  2. Save cleft in 4 axis-flip combinations.
- d. Spherical harmonics expansion (sample points are transparent blue coloured, approximated shape are shown as a salmon coloured mesh): Approximate surface function of each axis-flipped cleft by expanding spherical harmonics using spherical 21-design as sample points.
- e. Coefficient comparison: Scan coefficients from expansion against databases or set of pre-computed expansion coefficients of protein binding pockets or ligands via standard Euclidean distance metric. Investigate matches with smallest coefficient distance.

CleftXplorer's procedure for identifying, describing and comparing binding pocket shapes can be generally divided into five steps:

1. Identification of a binding site cleft
2. Reduction of cleft volume to where binding occurs
3. Transformation to a standard frame of reference
4. Spherical harmonic expansion of shape
5. Coefficient comparison between two shapes to quantify similarity

Ligand shapes are modelled using only steps 3-5 above. The clefts in a protein's surface are computed using SURFNET (see section 2.3.2.1), which detects protein cavities by inserting spheres of a certain range of sizes between protein atoms (see Figure 3.1a). The clefts are identified as distinct clusters of overlapping spheres and reduced in size (see next subsection and Figure 3.1b). For comparison of cleft and ligand shapes, it is necessary for the modelled shapes to be in the same orientation and coordinate frame of reference. Previous approaches used the rotational properties of the spherical harmonic functions to rotate the shapes in all orientations until the optimal superimposition was found (Ritchie and Kemp, 1999; Ritchie and Kemp, 2000). The rotation is achieved by using a Wigner rotation matrix on the coefficients and calculating the smallest distance between the respective coefficient vectors e.g. using a genetic algorithm (Cai, et al., 2002b). However, this is computer-intensive and unsuitable for database scanning. CleftXplorer speeds up the scan by pre-orientating the cleft model with three transformation operations as described in (Morris, *et al.*, 2005). The first translates the cleft model so that its centre of gravity is placed at the origin of the coordinate system (see Figure 3.1c bottom-left). The next step involves a rotation of the cleft in terms of its moments of inertia as a 'gross' shape characteristic. Therefore, the cleft model is rotated so that its moment of inertia tensor becomes diagonal with maximal values in  $x$ , followed by  $y$  followed by  $z$  (see Figure 3.1c bottom-right). However, the symmetry of the tensor cannot distinguish between objects at  $0^\circ$  and  $180^\circ$  rotation on the  $x$ -,  $y$ -,  $z$ -axes. To tackle this 'axis-flip-problem',

shape coefficients were calculated for four non-redundant combination of flips, resulting in four coefficient vectors for each cleft model (see Figure 3.1c top).

Next, CleftXplorer applies a spherical harmonics expansion on the shape of the transformed cleft. Therefore, the surface of the cleft, which is built up by the outer SURFNET spheres, is considered as a single valued (star-like) surface. In cases of a non-star-like shape, only the outermost surface points are taken into account. Furthermore, a sphere of radius 1.6 Å is rolled over the surface closing up any gaps between molecule atoms. The resulting star-like shape is considered as a spherical function on a unit sphere, with angle pairs  $(\theta, \phi)$  reflecting the domain values of the function extracted from spherical  $t$ -designs (see section 2.5.2.3.3). Using the 240 sample points of the spherical 21-design, the surface function is approximated by an expansion with real spherical harmonic functions (see section 2.5.2.3.2 and Figure 3.1d):

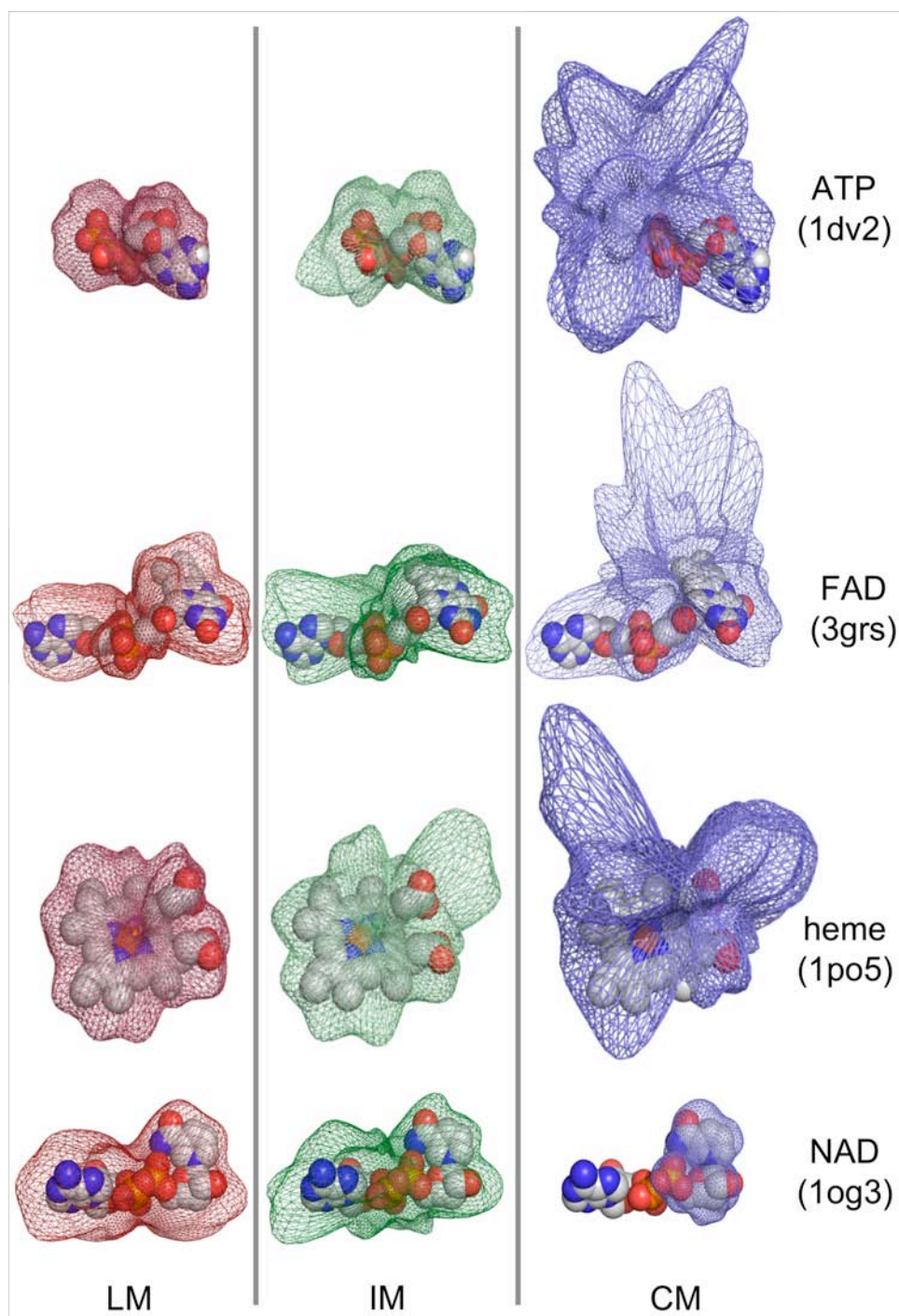
$$f(\theta, \phi) \approx \sum_{l=0}^{l_{\max}} \sum_{m=-l}^l c_{lm} S_{lm}(\theta, \phi) \quad , \quad (3.1)$$

where  $f(\theta, \phi)$  is the surface function,  $l_{\max}$  is 14,  $S_{lm}(\theta, \phi)$  are the surface spherical harmonic functions of indices  $l$  and  $m$ , and  $c_{lm}$  are the associated coefficients.

The coefficients are computed from the functional scalar product between the function and the spherical harmonics for each combination of  $l$  and  $m$  (see section 2.5.2.3.2).

The orthonormal property of the spherical harmonic polynomials guarantees a unique breakdown of the surface function into spherical harmonic functions in the expansion process and provides unique coefficients for any shape and size. The uniqueness enables the usage of the coefficients directly for comparison against other binding pocket or ligand coefficients. The standard Euclidean distance metric is used for the comparison (see section 2.5.2.3.2).

The rapid similarity calculation is the main strength of CleftXplorer and enables fast retrieval of related binding pockets or ligands from a coefficients database.



**Figure 3.2: Reconstructed shape of cleft models from different binding sites.**

The reconstruction corresponds to the order  $l_{\max} = 14$  for cleft models from ATP, NAD, heme and FAD. Associated ligands are shown as well and PQS-Ids are provided in brackets. The reconstructed shapes are visualised as a mesh and coloured according to the cleft models: CM = Conserved cleft region Model, IM = protein-ligand Interacting region cleft Model, LM = Ligand region cleft Model.



## 3.2.1 Cleft reduction

The following sub-sections are devoted to the three cleft models that are employed in CleftXplorer. Cleft models from SURFNET are often large and reach out beyond the region of the ligand location. Such clefts are neither convenient for binding pocket comparison nor for ligand docking. Therefore, all SURFNET clefts need to be reduced in size.

Three procedures were developed to reduce these initial clefts to more accurately approximate the actual shape of the ligand. All three cleft-reduction procedures provide a valid series of approximations to the real binding pocket depending on the available information. See (Glaser, et al., 2006) for a recent discussion on this topic of binding pocket localisation methods using SURFNET. An overview of the reconstructed cleft shapes for all three cleft models with their ligands is given in Figure 3.2.

All distances in the next subsections are calculated between the surfaces of Van der Waals atoms and SURFNET sphere.

### 3.2.1.1 Conserved cleft model

This approximation can be applied without any prior information about a protein-ligand interaction. The method uses the approach described in (Glaser, et al., 2006) to map phylogenetic residue conservation scores from the ConSurf-HSSP database (Glaser, *et al.*, 2003) onto the protein structure (see Figure 3.1b top-left). The basic idea of the method is that evolutionarily conserved residues are often functionally important and highlight potential ligand-binding residues when they are found within clefts (see section 2.3.2.5). By picking out only SURFNET spheres within 0.3 Å to a highly conserved residue atom with a ConSurf scores  $\geq 8$ , a cleft model can be extracted that comprises the ligand binding residues (see Figure 3.1 bottom-left). This approach is most suitable for structures solved by structural

genomics groups, where the function of the protein is unknown and no biologically relevant ligand is bound to the protein in the solved structure.

### 3.2.1.2 Interact cleft model

Another approximation of the binding pocket is obtained by keeping all SURFNET spheres within 0.3 Å of protein atoms interacting with the bound ligand (see Figure 3.1b centre). The residues are identified using HBPLUS (McDonald and Thornton, 1994). HBPLUS calculates hydrogen bonds between a protein and a ligand by looking at the distances and angles between potential hydrogen bond donors and acceptors. It also lists pairs of atoms that are in non-bonded contact. The Interact Cleft Model is of practical importance since methods already exist for predicting ligand-interacting residues (Bate and Warwicker, 2004; Laurie and Jackson, 2005; Ondrechen, *et al.*, 2001) and pharmaceutical companies as well as academia usually have high quality binding site information. Thus this approach can be used when there is no ligand bound in the available structure but the user has information about the ligand-binding protein residues.

### 3.2.1.3 Ligand cleft model

This somewhat artificial case represents the scenario of well-characterised binding pockets. Only SURFNET spheres that make contact to any ligand atom are retained (see Figure 3.1b right). This results in a very accurate, although not perfect, approximation of the ligand shape and produces a binding pocket that is obviously well suited for matching to its bound ligand. Any predictive approach will perform worse than this in getting the right shape, so this procedure corresponds to the 'best case' scenario and provides an estimate of the upper bounds on what performance can be expected for binding pockets with CleftXplorer.

## 3.2.2 Classification and data analysis

The following approaches were used to visualise and analyse the results in this chapter:

- **Distance matrices:** A distance matrix contains all-against-all pairwise coefficient distances that were calculated according to equation ( 2.41 ). These matrices give a good visual overview of the achieved classification power. A perfect classification in these plots is indicated by green squares for each ligand set in the diagonal from bottom left to top right. In the remaining rows and columns the coefficient distances should range from low to high as indicated by orange to yellow to white colouring, depending on the similarity level to the ligand set of interest. By rule of thumb coefficient distances smaller than 3 are considered as identical shapes and coloured in dark green. Coefficient distances between 3 and 5 are treated as similar, distances between 5 and 8 are regarded as dissimilar and distances between 8 and 10 are considered as highly dissimilar shapes. Coefficient distances above ten are not coloured at all and left in white. A grid on the matrices separates different ligand sets.

- **Area under Receiver Operating Characteristics Curves:** ROC (Receiver Operating Characteristic) curves (Hanley, 1982) and especially the AUC (Area Under ROC Curves) (Hanley, 1983) are well suited for the numerical comparison of classification approaches. ROC curves are used to measure the ranking quality of classifiers, by plotting the True Positive Rate ( $TPR$ ) against the False Positive Rate ( $FPR$ ) when the ordered list of classifications (in this work coefficient distances) is walked down from best to worst. Here,  $TPR = TP/(TP+FN)$  with  $TP$  = number of true positives and  $FN$  = number of false negatives and  $FPR = FP/(FP+TN)$  with  $FP$  = number of false positives and  $TN$  = number of true negatives. A diagonal ROC curve leading from the bottom left to the top right indicates a random classification where for each true hit a false hit is recovered (i.e. equal to flipping a coin). Such a curve encloses an area that corresponds to an AUC of 0.5. Conversely, the best case is a horizontal line at the top of the plot, where all true hits

are recovered before a false hit is obtained. Such a curve encloses the maximum area in the plot corresponding to an AUC of 1.0. Hence, AUC values closer to 1.0 indicate classifiers that are more able to distinguish true from false positives.

### 3.3 Data set

In order to answer the questions from the introduction to this chapter, a data set was required, which holds multiple examples of binding sites and ligands from unrelated proteins for which structural data was available. In fact, rather few binding-site/ligands complexes of this type were available in the PDB. Applying the criteria below, Data set I was compiled with 100 protein binding sites that bind one of nine ligand types (see Appendix A, Table A.1). The ligands were all of different size and flexibility, including phosphate as the smallest and most rigid molecule to ATP as flexible and middle-sized molecule up to FAD as the biggest and most flexible molecule in the data set. Following criteria were applied to derive Data set I:

1. Structural domains should be taken only from X-ray protein quaternary structures (PQS) that are thought to represent protein structures in their true biological unit (see section 2.1.2).
2. The binding sites in a ligand set should not be evolutionarily related but descend from different CATH H-levels (homologous superfamily) (Pearl, et al., 2003). In cases of homology, only the binding pocket with the highest X-ray resolution should be retained.
3. Partial, modified or incorrectly labelled ligands should be discarded, by comparing each ligand against the reference compound for that ligand's three-letter residue identifier in the MSD-ligand-chemistry database (MSDchem) (Golovin, et al., 2004).
4. Binding sites of only cognate ligands should be considered (see section 2.1.3). For enzymes a biologically relevant ligand was defined as one involved in the protein's

enzymatic reaction as given by the protein's EC number (Bairoch, 2000). For non-enzymes the protein's UniProt entry (Apweiler, et al., 2004) was checked for any information about its cognate ligand(s).

5. Each ligand set should have at least 5 members. (The number 5 was chosen arbitrarily but was deemed sufficient for assessing the success rate of assigning binding pockets to their ligand sets)

The intersection between two data sets in the literature, first (Stockwell and Thornton, 2006) and second (Nobeli, *et al.*, 2003) assisted the derivation of Data set I. The first data set ensured the achievement of the first three rules, whereas the second data set verified the fourth rule. Additional to both data sets manual searches were carried out to overcome two deficiencies of both data sets; namely that the first data set was missing all binding sites having no CATH domain assignments, whereas the second data set was missing all non-enzyme structures.

Binding sites without a CATH assignment were tackled by querying the Cathedral server (Pearl, et al., 2003) with the protein structure holding the binding site of interest and assigning to it the CATH code of the closest fold. The second deficiency was approached by scanning the appropriate three-letter residue identifier (e.g. FMN) and the ligand name (e.g. flavin) in the protein's UniProt entries. All hits were manually checked to avoid false positives.

The final data set, Data set I, comprises 100 protein binding pockets that bind either AMP, ATP, FAD, FMN, glucose, heme, NAD, phosphate or steroids (estradiol and dehydroepiandrosterone) (see Table A.1). It should be noted here that the data set covers only a tiny fraction of the chemical space that proteins are able to recognise and bind. 40% of PDB entries are enzymes and thus it is not surprising that most of the ligand molecules in Data set I are cofactors. The limited data in the PDB on non-homologous protein binding sites of other small molecules such as lipids, peptides, vitamins or glycans made it impossible at the time of this work to compile a more comprehensive list of ligand sets. Nevertheless,

despite its small size I believe that Data set I contains sufficient examples for an initial test on the complementarity of the geometrical and physicochemical properties between protein binding sites and their ligand counterparts.

## 3.4 Results and discussion

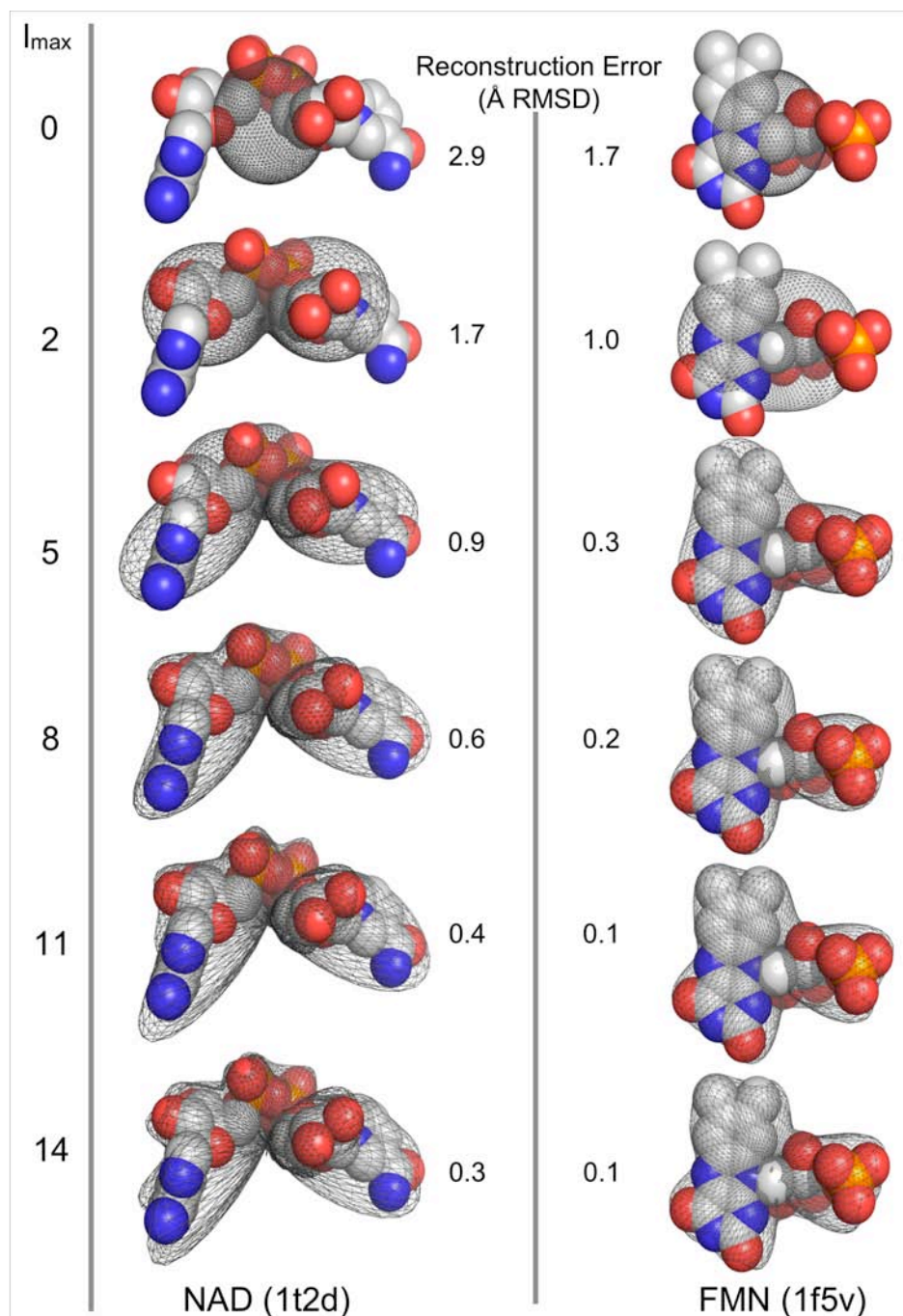
Before applying the spherical harmonic functions for the approximation of molecular shapes, the shape reconstruction must be compared to the molecules that were used to define it. Thus, first a number of quality checks of the shape description and comparison method are presented followed by a more detailed discussion on the biological implications of the results.

### 3.4.1 Shape reproduction quality and comparison metric

#### 3.4.1.1 Reconstruction error

Any function on the unit sphere can be reconstructed to any arbitrary error threshold by a linear combination of spherical harmonic functions to different orders. In Figure 3.3 such reconstructions are shown for two ligand molecules. Depending on the application, the spherical harmonics expansion can be terminated at an appropriate order, e.g. to roughly capture the overall shape of a small molecule an expansion up to  $l_{\max} = 6$  is sufficient; for highly non-central distributions an expansion order of several hundred may be necessary. The effects of series termination on the error of the binding pocket shape reconstruction are visualised in the two examples of Figure 3.3. Additionally, reconstruction errors are provided

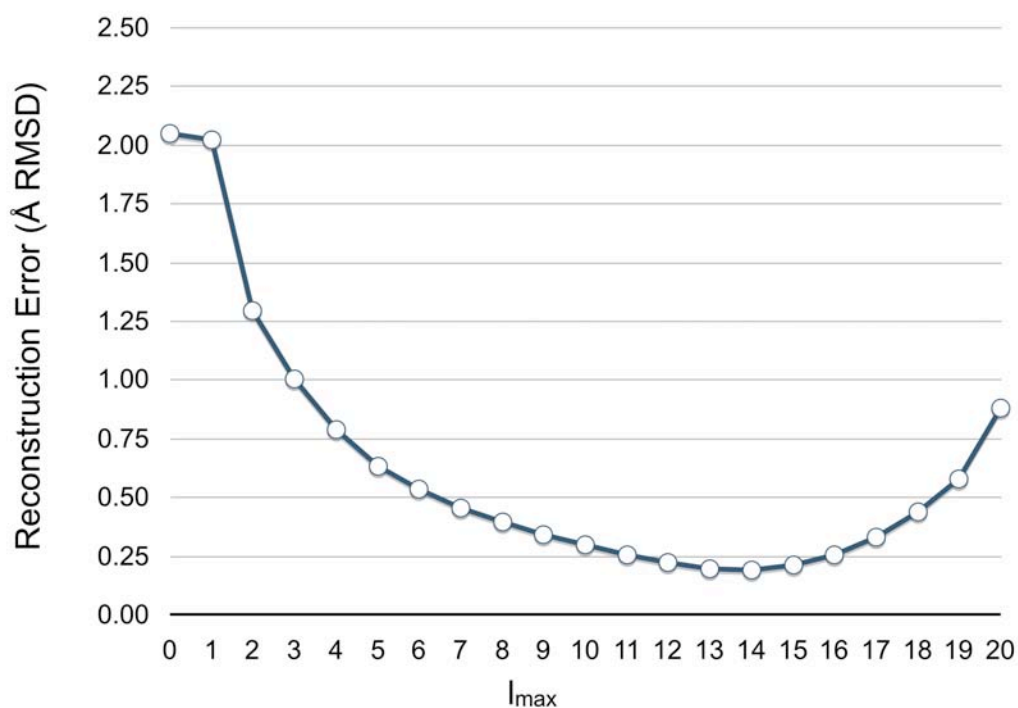
in the figure reflecting the RMSD (see equation ( 2.15 )) between 240 sample points from the spherical 21-design and their reconstructed values.



**Figure 3.3: Shape reconstruction with spherical harmonics.**

Various approximations of the shapes (black coloured mesh) for NAD and FMN with different degrees of termination in the spherical harmonics series expansion. Reconstruction errors are provided corresponding to RMSD values between the ligand shape and the reconstructed shape. (NAD was extracted from PQS structure 1t2d; FMN was extracted from PQS structure 1f5v).

In Figure 3.4, the reconstruction error is shown as a function of the expansion order. Mathematically one would expect the error to decrease smoothly with increasing expansion order, which is indeed the case for integration methods that are accurate into higher orders. Spherical designs as proposed by (Morris, 2006), however, have a limited region of applicability for integration in expansion space and are not yet algebraically proven to exist (see section 2.5.2.3.3) for those high orders that have been used in this work. The limitation leads to increasing numerical errors for spherical harmonic orders higher than  $l_{\max} = 14$  for the employed spherical 21-design. As the most accurate reconstruction of the shapes is demanded, whilst keeping a fast integration, all the cleft models and ligand shapes in Data set I were expanded to order 14, which according to the plot has an average error of 0.188 Å.



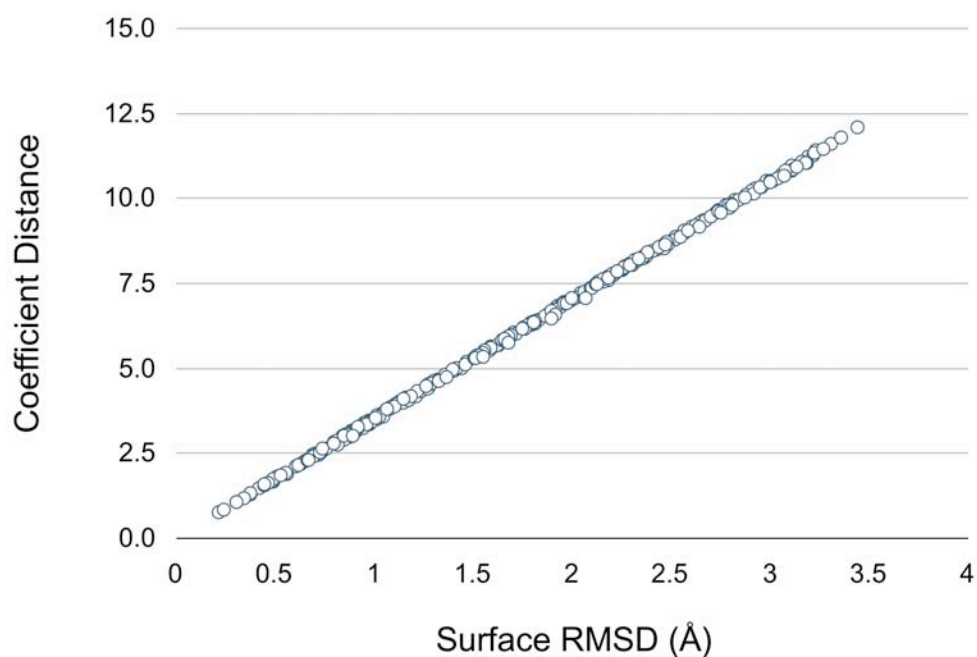
**Figure 3.4: Error while shape reconstruction with spherical harmonics.**

Reconstruction error for different degrees of termination in the spherical harmonics series expansion. The error was measured as RMSD between original sample point and reconstructed value. The rising error after  $l_{\max} = 14$  is due to spherical  $t$ -designs that are suitable only for the expansion to certain order and start to accumulate numerical errors when used for higher orders.



### 3.4.1.2 Comparison to surface RMSD

The difference between two shapes is calculated using the standard Euclidean metric in coefficient space. To assess whether the resulting coefficient distances indicate similarity or dissimilarity, they were plotted against surface RMSD (see Figure 3.5). The surface RMSD follows the common standard RMSD calculation in equation ( 2.15 ) but instead of using atomic coordinates, the 240 spherical 21-design sample points were used. The plot in Figure 3.5 shows a high correlation of  $R^2 = 0.99$  between the coefficient distance and the surface RMSD and thus allows the translation of any required RMSD into a coefficient distance with the ratio of 1 : 3.54. Experience shows that a coefficient distance of under 3 gives visually almost identical shapes and a distance below 5 corresponds to similar shapes.



**Figure 3.5: Coefficient distance correlation to surface RMSD.**

Diagram shows the high correlation between spherical harmonics expansion coefficients of the order  $l_{\max} = 14$  and surface RMSD with a ratio of 3.54 : 1. The surface RMSD was calculated similar to the structural RMSD, except that sample points on the shape surface were used instead of atom coordinates. The correlation coefficient is  $R^2 = 0.99$ .

Furthermore, the strong correlation between the coefficient distances and the surface RMSD values confirms that a weighting of the expansion coefficients is not required. The standard coefficients are already able to sufficiently distinguish dissimilar from similar shapes and sizes.

## 3.4.2 Shape Variation

After having assessed the accuracy of the spherical harmonics implementation in CleftXplorer, shape coefficients to the order  $l_{\max} = 14$  were calculated for all 100 ligands and cleft models in Data set I and compared to each other using the standard Euclidean metric.

### 3.4.2.1 Ligand conformations

It should be kept in mind that any recognition process of binding pockets or ligand shapes has to deal with conformational variance of both the protein (and therefore also the binding pocket) and the ligand. In a non-homologous data set containing unrelated proteins that may have evolved different strategies for binding the same ligand, one can expect different conformations and therefore different shapes for every flexible ligand. A robust shape-based classification of such a data set is therefore likely to be difficult. However, a working shape descriptor should be able to pick up conformational similarities for the same ligands and differences between different ligands.

The average shape similarity of all identical ligands in Data set I is 3.6 coefficient distance, which corresponds to a surface RMSD of approximately 1 Å, see Table 3.1. As such, the shape variation for individual ligands is low but mainly related to the flexibility of the ligand molecules. Four of the nine ligand sets (glucose, phosphate, steroid, AMP) can be considered as rigid molecules with an average distance of less than 3 (surface RMSD < 0.9 Å). Three of

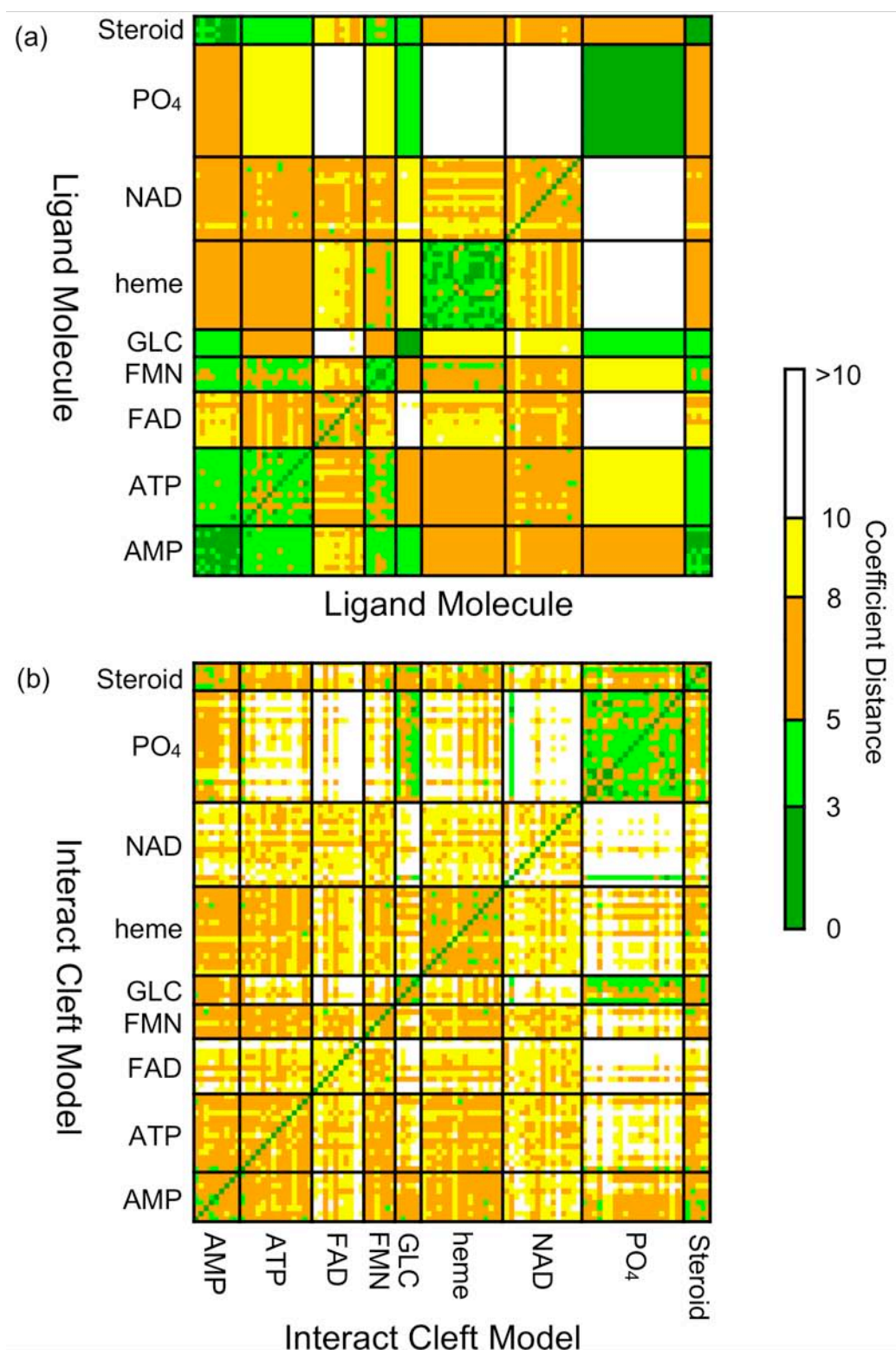
nine (heme, FMN, ATP) are slightly flexible with an average coefficient distance of less than 5 (surface RMSD < 1.4 Å).

Only NAD and FAD with their multiple single bonds are highly flexible molecules and occupy significantly different conformations, which confirms the results of (Stockwell and Thornton, 2006) that flexible ligands adopt a wide range of conformations when bound by non-homologous proteins. The bound conformation often differs significantly from the global energy minimum conformation in the unbound state (Nicklaus, et al., 1995).

**Table 3.1: Statistics on coefficient distances.**

<b>Ligand Set</b> <i>(set size)</i>	<b>Avg</b> <i>coeff dist</i>	<b>Std dev</b> <i>coeff dist</i>	<b>Min</b> <i>coeff dist</i>	<b>Max</b> <i>coeff dist</i>
<i>a) Statistics for ligand molecules</i>				
GLC (5)	1.2	0.2	0.6	1.5
PO <sub>4</sub> (20)	1.2	0.2	0.4	1.9
Steroids (5)	1.5	1.0	0.2	2.4
AMP (9)	2.4	0.5	1.1	3.9
Heme (16)	3.3	0.6	1.6	5.6
FMN (6)	3.8	0.6	2.3	4.6
ATP (14)	4.3	0.7	1.4	6.2
NAD (15)	6.8	0.9	4.5	9.8
FAD (10)	7.1	1.0	3.8	9.4
<b>Total (100)</b>	<b>3.6</b>	<b>1.9</b>	<b>0.2</b>	<b>9.8</b>
<i>b) Statistics for protein–ligand interacting reduced cleft models</i>				
PO <sub>4</sub> (20)	4.6	0.9	2.2	8.2
Steroid (5)	5.4	0.8	4.1	6.3
GLC (5)	5.6	1.4	3.6	7.6
AMP (9)	6.1	0.8	4.5	8.3
Heme (16)	6.5	1.0	3.8	8.9
FMN (6)	7.1	0.6	5.5	8.2
ATP (14)	7.4	0.7	5.2	10.1
FAD (10)	8.8	0.3	6.6	11.9
NAD (15)	9.0	1.8	6.2	13.7
<b>Total (100)</b>	<b>6.6</b>	<b>1.7</b>	<b>2.2</b>	<b>13.7</b>

The coefficient distances are ordered by their average for (a) ligand molecules and (b) protein-ligand interacting reduced cleft models. Surface RMSD values can be obtained by dividing the coefficient distances by 3.54 (see correlation in Figure 3.5).



**Figure 3.6: All-against-all coefficient distance matrices for Data set I.**

Matrices of all-against-all coefficient distances visualising the shape (dis)similarity between (a) ligand molecules and (b) protein-ligand interacting region cleft model shapes. The coefficient distances are coloured from green to orange to yellow reflecting low, intermediate and high coefficient distances. Coefficient distances higher than 10 are left out (white). The ligand sets are separated by a grid and labelled on the left of the figure and on bottom of each matrix.

The shape similarity between the ligand sets can be assessed in the distance matrix of Figure 3.6a. The matrices in Figure 3.6 show the all-against-all shape comparisons for (a) all the ligand molecules in Data set I and (b) all the binding pockets defined by the Interact Cleft Model (see section 3.2.1.2). The coefficient distances are colour coded according to the similarity level they reflect, ranging from dark green for highly similar shapes (low coefficient distances) to yellow and white for very different shapes (high coefficient distances). From Figure 3.6a it becomes evident that almost all ligand sets can be distinguished from each other just based on their shapes. Apart from FAD, NAD and partially ATP, all ligand sets are more similar to themselves than to other ligand shapes. The squares in the diagonal in the matrix from bottom left to top right contain lower coefficient distances than the rest of the matrix. The most distinct example is the ligand set of phosphate molecules. These are least similar to the large FAD, NAD and heme molecules (white rectangle in the matrix), highly dissimilar to AMP, ATP and FMN (yellow rectangles), dissimilar to steroid molecules (orange rectangles) but reasonably similar to glucose molecules (bright green rectangles).

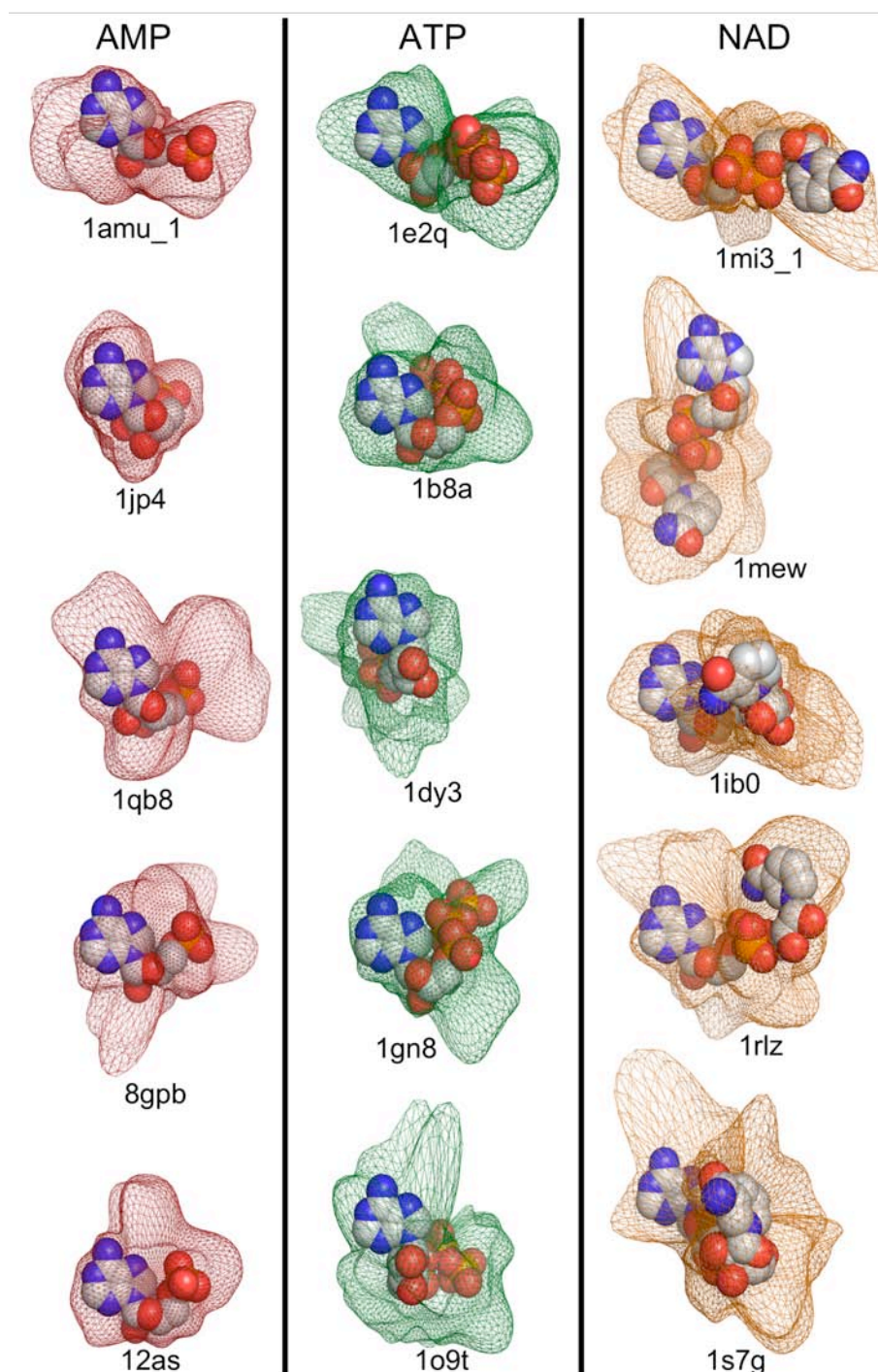
**Table 3.2: AUCs for various classification approaches on Data set I.**

<b>Cleft model</b>	<b>Cleft vs. Cleft</b>	<b>Cleft vs. Ligand mol</b>	<b>Ligand mol vs. Cleft</b>	<b>Ligand mol vs. Ligand mol</b>
<i>a.) Comparison with standard shape coefficients incorporating size and shape</i>				
Conserved	0.53	0.54	0.52	0.92
Interact	0.77	0.63	0.56	
Ligand	0.85	0.69	0.59	
<i>b.) Comparison with normalised shape coefficients corresponding to shape only</i>				
Conserved	0.52	0.52	0.55	0.87
Interact	0.64	0.64	0.73	
Ligand	0.74	0.68	0.83	
<i>c.) Comparison with the size of the shapes, which corresponds to the zeroth order in the spherical harmonics expansion</i>				
Conserved	0.53	0.51	0.51	0.94
Interact	0.73	0.51	0.51	
Ligand	0.76	0.52	0.51	
Average area under receiver operator curves (AUC) for different comparisons with different cleft models. Different cleft models in the rows are related to comparison combinations between cleft model and ligand molecules in the columns				

To assess how well shape information alone can identify a ligand, each ligand set in Data set I was 'predicted' by the ligand to which its shape matched most closely. From the set of obtained true and false positives and their match scores, a ROC (Receiver Operating Characteristics) curve was plotted and the area under the curve (AUC) calculated (see for details section 3.2.2). Values of AUC close to 1.0 correspond to perfect performing predictors whereas values close to 0.5 suggest that the predictor performs no better than random. Table 3.2a shows the AUC values obtained for various cleft and ligand molecule comparisons for the different cleft models. The ligand vs. ligand molecule comparison gives an AUC of 0.92, which shows that shape alone is a good but not a perfect predictor. Closer investigation of the AUC values of each ligand set reveal that rigid ligands are perfectly classified whereas flexible ligands like FAD, NAD and partially ATP adopt a wide variety of conformations and complicate the prediction of their ligand type from shape alone. Nevertheless, FAD and NAD molecules achieve an AUC value of about 0.75 and ATP scores an AUC value of 0.87. Thus, despite highly flexible, they retain a signature that allows them to be distinguished from other ligands.

### **3.4.2.2 Binding pocket shape diversity in ligand sets**

Of more practical interest, is how variable and identifiable are the binding pockets. The same analyses were performed on the three cleft models (Table 3.1b, Table 3.2a, Figure 3.6b). Briefly, the investigations showed that binding pockets do vary their shapes in non-homologous proteins just like the ligand molecules.



**Figure 3.7: Shape variation of protein binding pockets and ligands.**

Diversity of binding pocket shapes shown for 5 examples of AMP, ATP, and NAD. The binding pockets at the top are manually chosen. The other binding pockets are the most different ones to the manually chosen top binding pockets. Binding pockets correspond to protein-ligand interacting region cleft model and are oriented according to the adenine ring of their bound ligand and represented by a spherical harmonic reconstruction of the order  $l_{\max} = 14$ . PQS-Ids of associated protein structures are provided below each binding pocket.

Figure 3.7 shows a few examples of how the shapes of different protein binding pockets binding the same ligand vary. The binding pockets are modelled as Interact Cleft Models and oriented by superposing the adenine rings of the molecules. The cleft models at the top were manually chosen and the four models with the highest coefficient distance are displayed below. The high shape variation of Interact Cleft Models is numerically expressed by an average coefficient distance of 6.6 (surface RMSD  $\sim 2$  Å), see Table 3.1b. Nevertheless, despite this low shape similarity of the binding pockets, the average AUC value of 0.77 for the Interact Cleft Model in Table 3.2a suggests that there must be partial common shape information in each ligand set.

### 3.4.2.3 Binding pocket shape vs. ligand shape

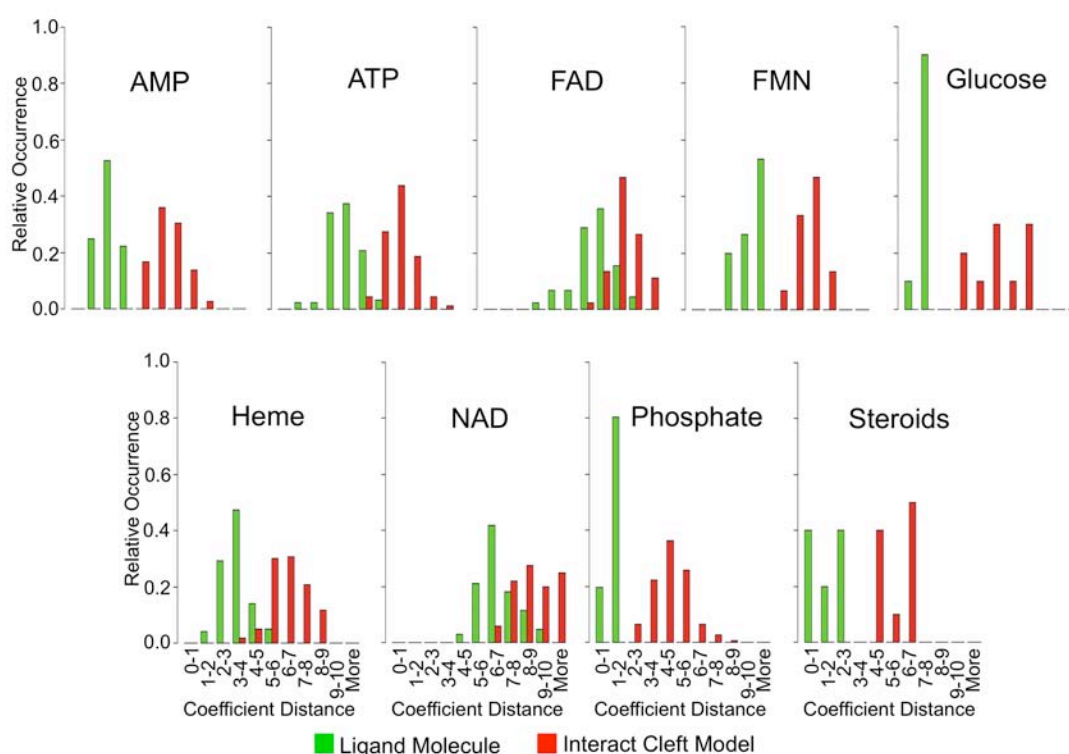
The average coefficient distances for the binding pockets are correlated to a certain extent with the flexibility of the bound ligand. Binding pockets binding rigid ligands like glucose and steroid molecules tend to have lower coefficient distances than binding pockets binding flexible ligands like FAD and NAD, see Table 3.1b. More interestingly, for binding pockets higher shape distances are observed than for ligand shapes, as shown in the coefficient distance distributions of Figure 3.8. This suggests that binding pockets are more variable in their shapes than their ligand counterparts. The binding pocket distances are so high that for more than half of all ligand sets the most similar binding pockets are less similar than the most different ligand shapes. The coefficient distances for the ligand molecules range between 1 and 4, whereas Interact Cleft Models show distances around 5 to 8. However, for FAD, NAD and partially ATP and heme the distributions overlap.

Additionally, for every ligand in Data set I, one can order all the other ligands by their coefficient distance from it. The rank of the first ligand, belonging to the same ligand set in each case, is plotted in the histogram in Figure 3.9. The green bars show that in 89% of the cases the closest ligand belongs to the same ligand set. Repeating the same procedure for



the Interact Cleft Model (red bars) reveals that the percentage at the first rank drops to 44%. When each Interact Cleft Model is compared to all ligand molecules (orange bars) the percentage at the first rank drops even to 27% and with more than half of the first true hits being beyond the rank order of 10.

Detailed examination of the crystallographic structures of the proteins shows that a perfect fit of the ligand into its binding site is never achieved. Not every ligand atom makes contact with the protein. Consequently, there is always space between parts of the ligand and the protein like a 'buffer zone'. Figure 3.10 illustrates that the buffer zone is partially occupied by crystallographic observable water. Thus on average, the Ligand and Interact Cleft Models are about 3 times larger in volume than their bound ligands (Table 3.3). Visual examples are given in Figure 3.2 and Figure 3.7. The difference in size means that the binding pocket of a small phosphate is able to accommodate an AMP, ATP, steroid or glucose molecule and this makes it impossible to match the ligands to their binding pockets on the basis of shape

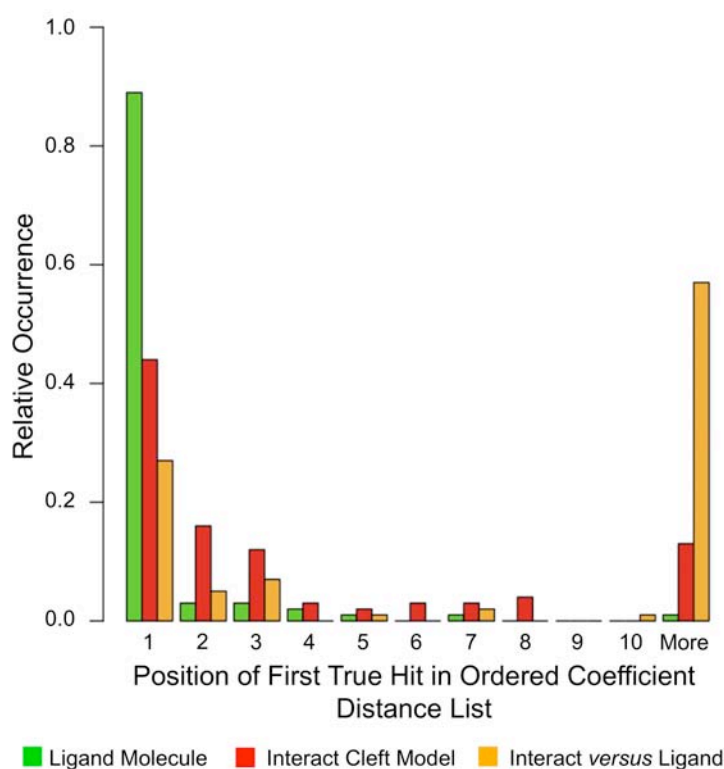


**Figure 3.8: Histograms comparing binding site and ligand coefficient distances.**

Distribution of the coefficient distances for each ligand set. Green and red bars show the relative occurrence of the coefficient distances for ligand molecules and protein-ligand interacting region cleft models, respectively.

similarity alone. This is reflected in the maximum AUC value of just 0.69 for the cleft vs. ligand comparison in Table 3.2a.

The higher shape variation of binding pockets compared to their ligand counterparts might give evidence that geometrical complementarity alone is in general not sufficient to drive ligand recognition in binding sites. According to the hypothesis of enzyme-transition-state complementarity (Benkovic and Hammes-Schiffer, 2003) protein binding sites are most complementary to the transition state of their ligand and not to their substrate or the product molecules (see section 2.3.2). A transition state cannot be observed in X-ray crystal structures (see Chapter 1). Therefore, a perfect complementarity between a protein and its ligand molecule can necessarily not be detected.



**Figure 3.9: Histogram of first true hits in coefficient distance calculations.**

Histogram of the relative occurrences of the positions that hold the most similar ligand set member. The positions are determined by ordering each coefficient distance list and recording the position of the first hit that belongs to the same ligand set when the list is walked down from best to worst. Green coloured bars illustrate the histogram for ligand molecules; red coloured bars show the histogram for protein-ligand interacting region cleft models and orange coloured bars display the histogram for the Interact Cleft Model vs. ligand molecule comparison.

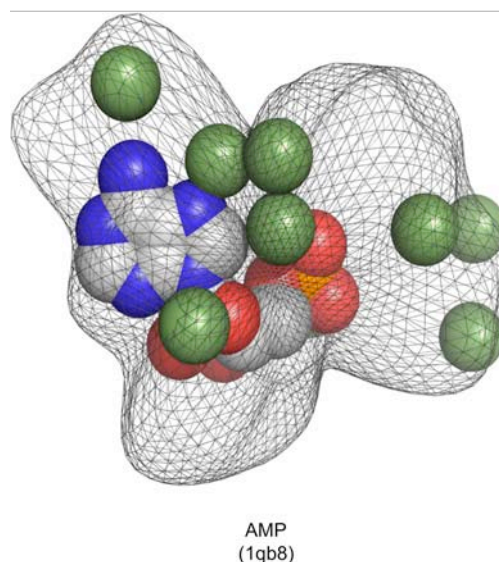
**Table 3.3: Statistics on the volume of Interact cleft models.**

<i>Ligand Set</i> (set size)	Avg Lig Vol [Å <sup>3</sup> ]	Avg Vol [Å <sup>3</sup> ]	Std Dev Vol [Å <sup>3</sup> ]	Min Vol [Å <sup>3</sup> ]	Max Vol [Å <sup>3</sup> ]
PO <sub>4</sub> (20)	73	445	118	168	797
GLC (5)	156	590	203	416	912
Steroids (5)	280	903	171	607	1144
AMP (9)	290	1097	156	774	1579
ATP (14)	400	1416	186	822	1723
FMN (6)	402	1443	265	1196	1879
Heme (16)	610	1507	209	1031	2030
NAD (15)	562	1809	305	486	2340
FAD (10)	688	2099	224	1580	2507
<b>Total (100)</b>	<b>395</b>	<b>1279</b>	<b>515</b>	<b>168</b>	<b>2507</b>

The ligand sets are ordered according to the average volume of their ligands in the second column.

### 3.4.2.4 Shape vs. size

Although spherical harmonics expansion is an approach for shape description, size is intrinsically incorporated into the expansion coefficients and therefore the previous results contain both shape and size. To highlight the importance of shape alone, a normalisation on all coefficient vectors was performed. As the zeroth order of the spherical harmonic coefficients reflects the general size of a shape (see section 2.5.2.3.2), the division of all coefficients by the zeroth order coefficient, places the shapes on the same scale and thereby removes the influence of different sized objects. Table



**Figure 3.10: Buffer-zone and water molecules in binding sites.**

Not every ligand atom contacts a protein atom and thus leaves space between parts of the ligand and the protein. The space is partially occupied by crystallographic observable water molecules. An example is shown on the AMP binding pocket of PQS entry 1qb8, with the reconstructed pocket shape shown as a black coloured mesh, the ligand shown in varicolour and the oxygens of the water molecules shown as green coloured spheres.

3.2b and Table 3.2c show the AUC values for the classification using normalised coefficients (only shape incorporated) and zeroth order coefficients (only size incorporated) respectively. From the AUC values it can be observed that shape plays the main role in the cleft vs. ligand comparison and vice versa. For the clefts vs. clefts and ligands vs. ligands comparison size seems to outweigh the performance of shape alone. This is not remarkable, as the ligands in Data set I are almost all distinguishable by size. However, except for the cleft vs. cleft comparison of the Interact Cleft Models, it is remarkable how little the performance differs when using only shape for the classification.

In fact, the size difference between binding pockets and ligands accounts for the failure of the shape comparison method to match binding pockets to their ligands as described in the previous subsection. With the normalisation, the size is excluded and a successful matching solely on shape becomes possible. As a result, the AUC value for the ligand vs. cleft comparison rises to a maximum of 0.83 (see Table 3.2b). Interestingly, the cleft vs. ligand comparison still gives relatively low AUC values, which is caused mainly by the FAD and NAD ligand sets. The average coefficient distances using normalised coefficients for FAD and NAD binding pockets are smaller than for their ligands, due to imperfect complementarity.

### **3.4.2.5 Performance of cleft models**

The poor performance of the Conserved Cleft Model is mainly caused by enzymes in Data set I that have at least two binding pockets next to each other (one for the cofactor and one for the substrate). As both binding pockets are important for the function, both will be highly conserved. Thus reducing the SURFNET spheres via conservation still results in a larger merged cleft model, consisting of the cofactor and substrate binding pocket. This is a common problem and at least 27 ligands in Data set I are known to be cofactors for which a 'combined' binding pocket was obtained. The selection of conserved residues is further complicated by the fact that not only binding site residues are highly conserved but also

residues that are essential for the structural integrity of a protein, e.g. both cysteines in a disulphate bond. Chelliah and co-workers were able to differentiate between both types of conserved residues by adding further restraints to the conservation calculation (Chelliah, *et al.*, 2004). Their powerful discriminator was based on environment specific substitution tables that for each amino acid listed substitution likelihoods in dependency to the local environment of the amino acid in a folded protein. Although the application of such substitution tables is likely to improve the performance of the Conserved Cleft Model, it would not solve the problem of obtaining distinct cleft models for binding sites that are located next to each other.

Another issue is the divergence of substrates in some large protein families like the SDR protein family (Oppermann, *et al.*, 2003), where the binding site is not more conserved than the rest of the protein. In these and similar cases the Conserved Cleft Model contains only a portion of the binding pocket (see Conserved Cleft Model of NAD binding pocket in Figure 3.2).

It is also important to note the number of homologous proteins used to calculate the sequence conservation and their sequence similarity. Few sequence homologs will result in an unreliable conservation score and therefore in an unreliable binding pocket prediction. Furthermore, PSI-BLAST, which is used by ConSurf does not distinguish between orthologous and paralogous sequences in its search for homologous proteins. The distinction is however important. While both sequences originate from the same ancestor, only the former has a high probability to observe the same function and thus support the discovery of functionally important residues in a protein. Paralogous proteins are the result of gene duplication events that can withdraw the evolutionary pressure from the gene duplicates to maintain their function enabling them to evolve new protein functions. As a consequence, paralogous proteins are inappropriate for identifying functionally important residues but are rather beneficial for determining sequence and structure based factors that can lead to alternations of protein functions (Dunbrack, 2002).

### 3.4.3 Limitations and problems

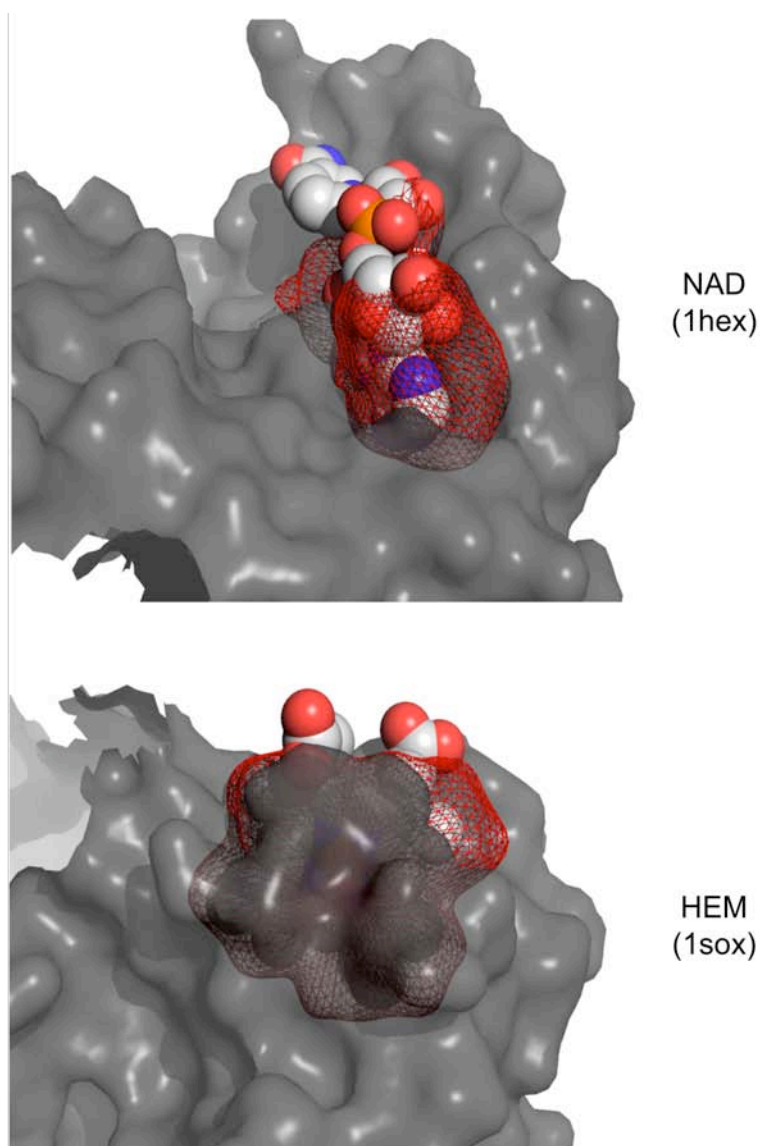
#### 3.4.3.1 Binding pocket prediction

The main obstacle of the approach presented here, is the accuracy of the cleft model and in particular the binding pocket prediction step for cases where no information about the location of a binding site is known. A number of approaches exist (see section 2.3.2), but generally, an accurate solution remains unavailable.

Other problems involve some general characteristics of protein structures. For example, loop regions are often missing in crystallographic structures due to their flexibility, making it difficult to predict the binding pocket for those built up partially by loops. Nine protein structures in Data set I feature missing loops close to binding sites. Furthermore, many protein structures are solved as part of functional assessment experiments, where functionally relevant amino acids are mutated to study their effects on the protein structure and function. Mutations are often performed on ligand interacting residues resulting in a slightly different binding pocket shape. Such mutations are found in 26 protein structures in Data set I. Other more technical problems involve the accuracy of X-ray structure coordinates. The median value of the estimated standard deviation for atoms in crystallographic structures ranges from 0.1 to 0.5 Å (DePristo, et al., 2004; Laskowski, 2003). Neither SURFNET nor HBPLUS account for these uncertainties in their algorithms, which leads to missing SURFNET spheres in some of the cleft models. Furthermore the crystalline arrangement of the proteins in crystals often constrains protein molecules to adopt a non-biological conformation at the crystal contacts (Jacobson, et al., 2002). Such constraints can affect the structure of binding sites as shown for the active site of a homodimeric cytoplasmic malate dehydrogenase (Birktoft, et al., 1989).

### 3.4.3.2 Partially bound ligands

Some ligands are bound only partially inside a binding pocket with their other end protruding into the solvent, such as the NAD and the heme group in Figure 3.11. As the spherical harmonic functions work globally on the whole shape they are not well suited for local shape matching. Finding the correct ligand in such cases will not succeed. However, if the partial



**Figure 3.11: Partially occupied binding pockets.**

Two examples for a partially bound ligand to its protein. The protein is represented as a transparent surface coloured in grey, the reconstructed binding pocket shape (protein-ligand interacting region cleft models) is shown as a red coloured mesh and the ligands are varicoloured. The top example shows an NAD (PQS-Id: 1hex) from which only the front part is surrounded by amino acids. The bottom example displays a heme group (PQS-Id: 1sox), which protrudes to the solvent with its two carboxyl-groups.

bound state is a common picture for the entire protein family, a cleft vs. cleft comparison could help to find a homologous family member.

### **3.4.3.3 Star-like shapes and rotational variance**

There are some minor problems related to properties of the spherical harmonic. These functions are suitable for describing the global surface of star-like shapes. However, binding pockets and ligands are not always star-like in shape. In cases where the ray from the centre of gravity to the surface penetrates the surface more than once, the outermost surface point was used to approximate the global shape. This can bring some loss of shape information but should not change the matching results significantly.

Furthermore, the coefficient vectors are not rotationally invariant. Although obtaining coefficient vectors for all four axis-flip-combinations solved the flipping-problem, it is still possible that a rotationally invariant shape descriptor will improve the results.

### **3.4.3.4 Single property descriptor**

The molecular recognition of a ligand is induced by physicochemical properties in addition to the shape, such as electrostatic potential and hydrophobicity. Including such features in the cleft models and the ligands might improve the observed results (see section 4.3.5).

## **3.5 Conclusions**

In this chapter, a fast and efficient spherical harmonics shape descriptor was employed to compare binding pocket and ligand shapes (Figure 3.2 to Figure 3.5). It was shown that the



shape descriptor is able to reflect the conformational state of the ligands allowing correct classification of rigid ligands, but poor classification of highly flexible ligands (Figure 3.6, Table 3.2).

In addition, it was shown that the assumption about proteins binding similar ligands having similar geometrical properties is only partially true (Figure 3.7). As expected, the similarity is closely related to the flexibility of the ligand molecules. The binding pockets are observed to be more variable in their shapes than their bound ligand molecules with a difference in their average coefficient distances of 3.0, which corresponds to 0.9 Å surface RMSD (Figure 3.8, Table 3.1). This difference in shape variation between the cleft models and ligand molecules shows that shape complementarity in general is not sufficient to drive molecular recognition alone and requires additional physicochemical properties (see Chapter 4).

Furthermore a 'buffer zone' can be found between the ligand and ligand interacting protein atoms which is partially occupied by water molecules so that on average binding pockets tend to be 3 times larger than their bound ligand molecule (Figure 3.10, Table 3.3).

The normalisation procedure of the standard spherical harmonic coefficients enabled the investigation of the contribution of shape and size to the classification performance (Table 3.2). Shape alone outperforms the contribution of size alone in the classification, but size does surprisingly well when comparing clefts to clefts and ligands to ligands. However, the molecular sizes of the ligand sets in this study were almost all distinguishable, which would not be the case if all metabolites were considered.

The relationship between classification performance and accuracy of the cleft models points towards the need for a good binding pocket model (Table 3.2). The random classification of the conserved cleft regions proved that residue conservation does not provide sufficiently accurate binding pocket models and cannot be used for function prediction. However, the

global shape descriptor combined with the Interact Cleft Model is an elegant descriptive method for comparing binding pocket shapes in protein families.

## Chapter 4

# On the Diversity of Physicochemical Environments Experienced by Identical Ligands in Binding Pockets of Unrelated Proteins

## 4.1 Introduction

It is generally assumed that the molecular recognition between protein receptors and ligand molecules requires in addition to geometrical complementarity, the physicochemical complementarity between both binding partners (Tsai, et al., 2002). In particular, the complementarity derived from long-range electrostatic interactions is believed to be the driving force for molecular interactions. For example, proteins binding an adenine or guanine were found to discriminate both molecular moieties on the basis of electrostatics (Basu, et al., 2004). Members of the copper zinc superoxide dismutase protein family attract positively charged metal ions into their binding sites through a highly negative electrostatic field (Livesay, et al., 2003) and DNA-binding proteins attract negatively charged DNA molecules through positively charged patches on their protein surface (Tsuchiya, et al., 2004). Different studies have been performed to quantify the electrostatic complementarity between binding partners. In particular, early works by Nakamura and coworkers (Nakamura, et al., 1985a; Nakamura, et al., 1985b; Tsuchiya, et al., 2006), Chau & Dean (Chau and Dean, 1994a; Chau and Dean, 1994b; Chau and Dean, 1994c) and Naray-Szabo (Gerczei, *et al.*, 1999; Naray-Szabo, 1989; Naray-Szabo and Nagy, 1989) analysed the electrostatic lock-and-key model of

single ligand and protein binding sites by mapping the Molecular Electrostatic Potential (MEP) of both binding partners on a set of reference points on their surfaces. The MEPs on the reference points were eventually compared following the host-on-guest and guest-on-guest model. These studies came to the conclusion that long-range effects of partial charges are essential for creating complementary electrostatic environments between binding site and ligand, but that partial charges on protein and ligand lack face-to-face complementarity. Similar conclusions were drawn for protein-protein interfaces (McCoy, et al., 1997).

It has been suggested that physicochemical interactions in natural complexes lack exact complementarity (Kangas and Tidor, 2001) most likely to avoid irreversible ligand binding. In evolutionary terms, one could state that proteins have evolved to bind their ligand counterparts just as much as necessary to perform their biological function. Once a protein has established its function, a further increase in binding affinity becomes unnecessary and with respect to the irreversible binding even disadvantageous. In this context, Ledvina and coworkers showed that some phosphate receptors, sulphate binding proteins, flavodoxin structures and DNase proteins exert an intense negative electrostatic potential at their binding sites despite binding a highly negative ligand (Ledvina, et al., 1996). The proteins stabilize the anion charges by van der Waals interactions and an extensive local hydrogen-bonding network comprising main chain NH groups and hydroxyl side chains. The question remains open whether in their evolutionary past these enzymes were binding ligands with complementary electrostatic potentials. Similarly, Herschlag and colleagues' study of the electrostatic and geometric complementarity in the oxyanion hole of ketosteroid isomerase concluded that electrostatic complementarity makes only a modest contribution to the enzyme's catalytic mechanism which is primarily driven by geometrical complementarity (Kraut, et al., 2006). Nakamura listed four reasons for the absent of electrostatic complementary in some protein ligand complexes: (1) the ligand interacts mainly with the solvent, (2) hydrophobic interactions are strong, (3) dissociation of the ionisable protein residues was incorrectly assigned and (4) additional ionic ligands were affecting the experienced electrostatic potential of the ligand (Nakamura, et al., 1985a).

To gain a wider perspective on physicochemical complementarity, the physicochemical characteristics of the 100 protein binding sites in Data set I were analysed. Countless applications have been developed in the past for analysing active and binding sites, mainly in the field of rational drug design. Most of the programs like LigBuilder (Wang, et al., 2000b), MCSS (Miranker and Karplus, 1991), PocketFinder (An, et al., 2005), Q-SiteFinder (Laurie and Jackson, 2005) are derivatives of the program GRID (Goodford, 1985). GRID, developed by Goodford and coworkers, analyses the physicochemical properties of protein binding sites by computing the interaction energy between various spherical probes and the atoms of the protein using a standard potential energy function. Other programs provide different information about a given binding site. For example, GRASP (Nicholls, et al., 1991), although primarily a powerful protein structure visualization software, has a built-in Poisson-Boltzmann Solver which allows the calculation and visualisation of the electrostatic potential in and around the protein binding site. The program HINT (Kellogg, et al., 1991) calculates the hydrophobicity of a molecule using experimental octanol/water partition coefficients and constructs a hydrophathy field or complementarity map for a protein binding site. The CASTp (Binkowski, *et al.*, 2003b) database is a repository on protein clefts and voids and provides area and volume measurements for each cavity. I have developed CleftXplorer (see Chapter 3), which uses spherical harmonic expansion coefficients to analytically describe the shape and size of binding sites and ligands molecules (see Chapter 3). To obtain a more complete picture also of the physicochemical properties within binding pockets, the shape descriptor in CleftXplorer was extended by various physicochemical descriptors that characterise electrostatic charged-charged interactions, hydrophobic interactions, hydrogen bonds and van der Waals interactions (see 4.2 Methods).

Despite differences in their algorithms, all methodologies that aim to identify a ligand that binds to a given protein binding site or, indeed, identify what the cognate ligand might be, assume that ligands and binding sites exhibit geometric as well as physicochemical complementarity. If this assumption is correct, binding sites that bind the same cognate ligand

should share very similar properties. Here, such an analysis is reported revealing a surprising diversity of properties among binding pockets binding the same ligand.

## 4.2 Methods

### 4.2.1 Calculating protein physicochemical properties on ligand molecules

The computer program CleftXplorer (see Chapter 3) was modified so that in addition to describing a 3D shape of a binding pocket or a ligand molecule using spherical harmonics, it could describe physicochemical properties mapped onto the surface of that shape. The physicochemical properties included electrostatic potentials, hydrophobicity scores, hydrogen bond donor, acceptor and van der Waals potential energies. These were computed using standard software packages, supplemented by own code as described below. For the purpose of this chapter, the physicochemical properties were not mapped on the surface of the shapes, but rather onto the atom centre coordinates of each ligand in the data set, thereby following a similar approach to GRID (Goodford, 1985). CleftXplorer's algorithm is simple and consists of three steps:

- 1.) Remove all crystallographically observed water molecules from the protein structure file. The computed electrostatic potential and hydrophobicity score implicitly account for surrounding water molecules.
- 2.) For each ligand, identify and remove those polypeptide chains and Neighbouring Chemical Compounds (NCCs) that lie within 9 Å of any ligand atom. NCCs correspond to HET groups such as metals, cofactors and coenzymes. The cut-off

distance of 9 Å is in accordance with the typical interaction ranges of non-bonded interactions (Mackerell, 2004). This speeds up the calculations while having negligible effects (Basu, et al., 2004; Nielsen and McCammon, 2003) on the calculated properties.

- 3.) At the centre of each ligand atom, calculate the value of each physicochemical property, using the atoms of the retained protein chains and NCCs.

## 4.2.2 Electrostatic potential

The electrostatic potentials were calculated by solving the nonlinear Poisson-Boltzmann equation (see section 2.4.2.2.1) using APBS (Baker, et al., 2001) (version 1.0). A grid size of 1 Å was chosen and the finite difference discretisation technique (Klapper, et al., 1986) was applied. Calculations were run at room temperature with a counter-ion concentration of 0.1 M. The continuum solvent model was allocated a dielectric constant of  $\epsilon_S = 78$ . Protein and ligand atoms as well as NCCs were assigned a dielectric constant of  $\epsilon_M = 4$ . Hydrogen atoms were added to the protein structure using the program REDUCE (Word, et al., 1999) (version 3.13) which optimizes the protein's hydrogen bond network, whilst flipping ASN or GLN side chain amides where appropriate. Protonation states of histidine residues were approximated using PROPKA (Li, et al., 2005) at a pH given for the mother liquor in the PDB or in the primary literature of the crystal structure. Partial charges and atom radii from the PARSE parameter set (Sitkoff, et al., 1994) were assigned to all protein atoms using the automated procedure of PDB2PQR (Dolinsky, et al., 2004) (version 1.3.0). The reported potentials from APBS were converted from  $kT/e$  to  $kcal/mol \cdot e$  by applying the conversion factor of 0.592.

The ligands in Data set I as well as all NCCs were assigned Pauling's van der Waals radii (which are also used in the PARSE parameter set). All ligands in the data set were left uncharged in the binding sites, to account for the reduction of the screening effect by the

ligand molecule. Partial charges for the NCCs were obtained from the Merck Molecular Force Field (MMFF94) (Halgren, 1996) as implemented in an in-house modified version of the Chemistry Development Kit (CDK, version 2.0.2) (Steinbeck, et al., 2003) for Java. The correctness of the MMFF94 partial charges was checked against the MMFF94 validation suite (<http://www.ccl.net/cca/data/MMFF94/>). For simplicity, metal ions were assigned partial charges equal to their formal charge following the approach of MMFF94. Iron ions in heme molecules were treated as an integral part of the molecule and assigned a partial charge of +2e irrespective of their oxidation state. Hydrogen atoms were added automatically to the ligands and NCCs at pH 7 using OpenBabel (Guha, et al., 2006) (version 2.1.1) and in the few cases where the automatic assignment failed were added manually using PyMOL (version 1.0) (DeLano, 2002). The spatial position of the hydrogen atoms were optimized within the binding site using REDUCE.

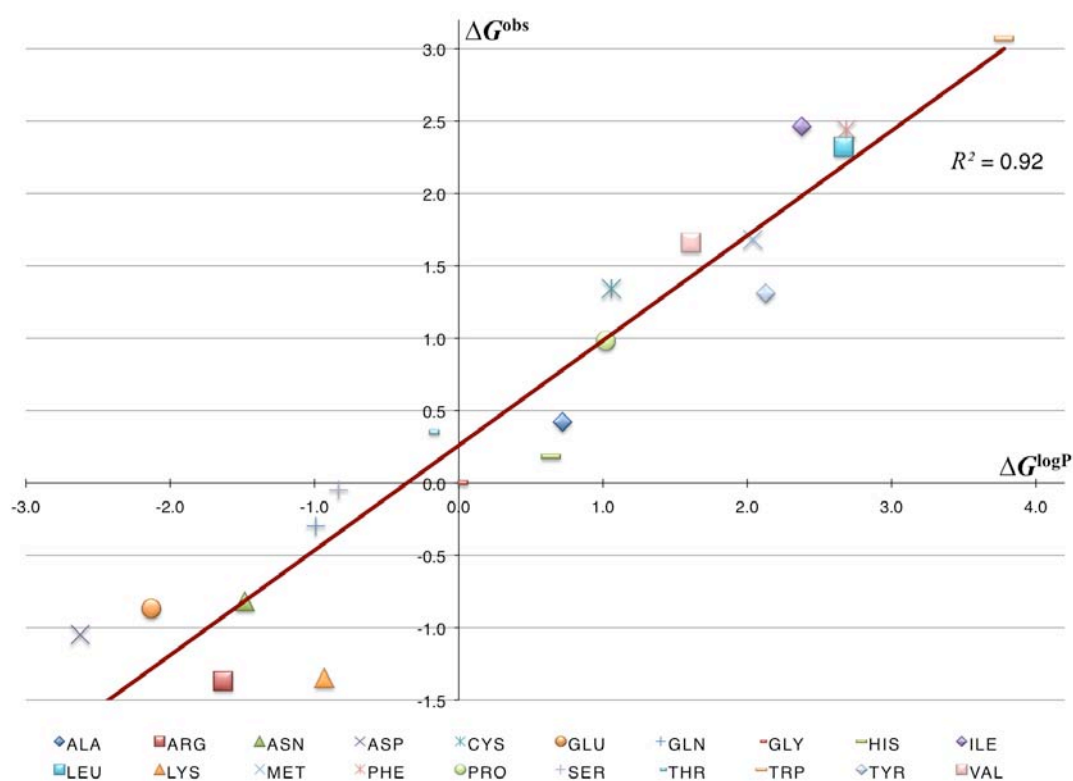
### 4.2.3 Scoring the hydrophobic environment

The calculation of CleftXplorer's hydrophobicity scores follows the approach of HINT (Kellogg, et al., 1991) and the hydrophobicity potential of (Fauchère, et al., 1988) in relating the hydrophobicity score to octanol-water partition coefficients ( $\log P$ ) values. CleftXplorer differs in that it uses atomic based rather than fragment based  $\log P$  values. This makes CleftXplorer applicable to a wider spectrum of molecules especially those for which HINT has no fragment parameters. The atomic  $\log P$  values were calculated using a modified version of the XlogP (Wang, et al., 2000a) algorithm in the CDK. Modifications were required to add  $\log P$  values for charged molecules as by default,  $\log P$  values are defined only for neutral molecules (see section 2.4.3.3). Some of the ligands in Data set I, as well as some of the NCCs contain charged moieties (e.g.  $\text{PO}_4$  groups), which increase the solubility of the molecule in water. If the molecule is treated as neutral, its hydrophobicity will be overestimated. To reduce the  $\log P$  value for charged atoms, a correction factor of -1.083 was introduced that corresponds



to half of the zwitterion correction factor used in the XlogP algorithm for the hydrophobicity calculation of amino acids. Metal ions were assigned an arbitrary  $\log P = -3$ , assuming that metal ions in solution minimally interfere with the water network. For protein atoms  $\log P$  values were calculated considering each amino acid within the tripeptide Gly-X-Gly, where X is the amino acid of interest. The values were calculated relative to glycine, following the method of calculating the Hansch  $\pi$ -constant (Fauchère and Pliska, 1983), by subtracting the mean atomic  $\log P$  value for glycine atoms from each of the atomic  $\log P$  values giving an atomic  $\log P^{rel}$  value. Based on  $\log P^{rel}$ , the energy of transfer from water to organic solvent,  $\Delta G^{logP}$ , was inferred following the approach of (Eisenberg and McLachlan, 1986) by applying the Boltzmann equation from statistical thermodynamics:

$$\Delta G^{logP} = \log P^{rel} \cdot 2.30RT = \log P^{rel} \cdot 1.36 \text{ kcal/mol} \quad , \quad (4.1)$$



**Figure 4.1: Scatter plot of  $\Delta G^{logP}$  versus  $\Delta G^{obs}$**

Scatter plot comparing for protein amino acid their XlogP based solvation energies  $\Delta G^{logP}$  (as computed in this work) with published experimental solvation energies  $\Delta G^{obs}$  (Eisenberg and McLachlan, 1986).

where  $R$  is the Boltzmann factor and  $T$  is room temperature in Kelvin. The XlogP based 'hydrophobicity energy' correlates with the experimental free energy of transfer (Eisenberg and McLachlan, 1986) with  $R^2 = 0.92$ . A similar approach to the molecular lipophilic potential (MLP) (Heiden, *et al.*, 1993; Mancera, 2007) was used to map the 'hydrophobicity energy' to any point in space with the following sigmoid potential function that was suggested by Levitt (Brylinski, *et al.*, 2006; Levitt, 1976):

$$f(dist^{rel}) = 1 - 0.5 \left( 7dist^{rel\ 2} - 9dist^{rel\ 4} + 5dist^{rel\ 6} - dist^{rel\ 8} \right) \quad , \quad (4.2)$$

where the value of  $dist^{rel}$  ranges from 0 to 1 and corresponds to the distance between a ligand and protein or NCC atom divided by the cut-off distance of 9 Å.

To calculate the total Hydrophobic Environment Score (HES) at a ligand's atom centre, it was further necessary to sum up the product between the distance function  $f(dist^{rel})$  and the energy of transfer  $\Delta G^{logP}$  over all  $n$  neighbouring atoms. Following equation expresses this simple sum and gives the final equation for HES:

$$HES = \sum_{i=1}^n f(dist^{rel})_i \cdot \Delta G_i^{logP} \quad . \quad (4.3)$$

## 4.2.4 Other physicochemical properties

CleftXplorer calculates van der Waals potential energies using the 'buffered 14-7' potential equation from MMFF94 (Halgren, 1992). Hydrogen bond donor and acceptor potential energies are calculated according to the knowledge-based potential energies from (Kortemme, *et al.*, 2003). These orientation dependent hydrogen-bonding potentials were derived from geometric features found in a large set of high-resolution protein structures. For more details, see the cited publications.

## 4.2.5 Average properties and their variation

Once all physicochemical properties for all ligands in a ligand set have been calculated, the physicochemical property scores on the same atoms in the same ligand in different binding sites were averaged and their standard deviation computed. For this purpose, an atom-mapping table was created mapping corresponding atoms between ligand A and ligand B, where both were from the same ligand set. In cases where the molecular moiety had rotational symmetry, atoms at similar spatial positions were associated to each other, after the molecules were manually aligned on the neighbouring moieties. Atoms in the phosphate ligand set were mapped randomly as a unique mapping due to symmetry was not possible. However, the small size of phosphate molecules should generally result in similar physicochemical property scores over the whole molecule.

In addition to the standard deviation, the relative number of sign-changes in the property scores for each atom was calculated as a further mean to assess the variation of the physicochemical properties in binding sites. The Sign Change Ratio (SCR) for an atom  $i$  in a ligand is given by

$$SCR_i = 1 - \frac{\max(N_i^+, N_i^-)}{N_i^{total}} . \quad (4.4)$$

$N_i^+$  is the number of positive score values,  $N_i^-$  the number of negative score values and  $N_i^{total}$  is the size of the ligand set. The functional range of SCR is between 0 and 0.5, with 0 denoting no variation (all corresponding atoms have equal sign), and 0.5 indicating highest variation (one half of corresponding atoms are positive, the other half negative).

## 4.2.6 Data set

The analysis in this chapter was performed as in Chapter 3 on the Data set I (see Appendix A, Table A.1). As mentioned in the previous chapter, nine protein structures in Data set I have loop regions in the proximity of their binding sites for which, due to insufficient experimental data, no atom coordinates are available in the PDB file. These loops regions probably exhibit high flexibility and contribute insignificantly to the binding process. Therefore, no attempt was made to include them in the physicochemical property calculations.

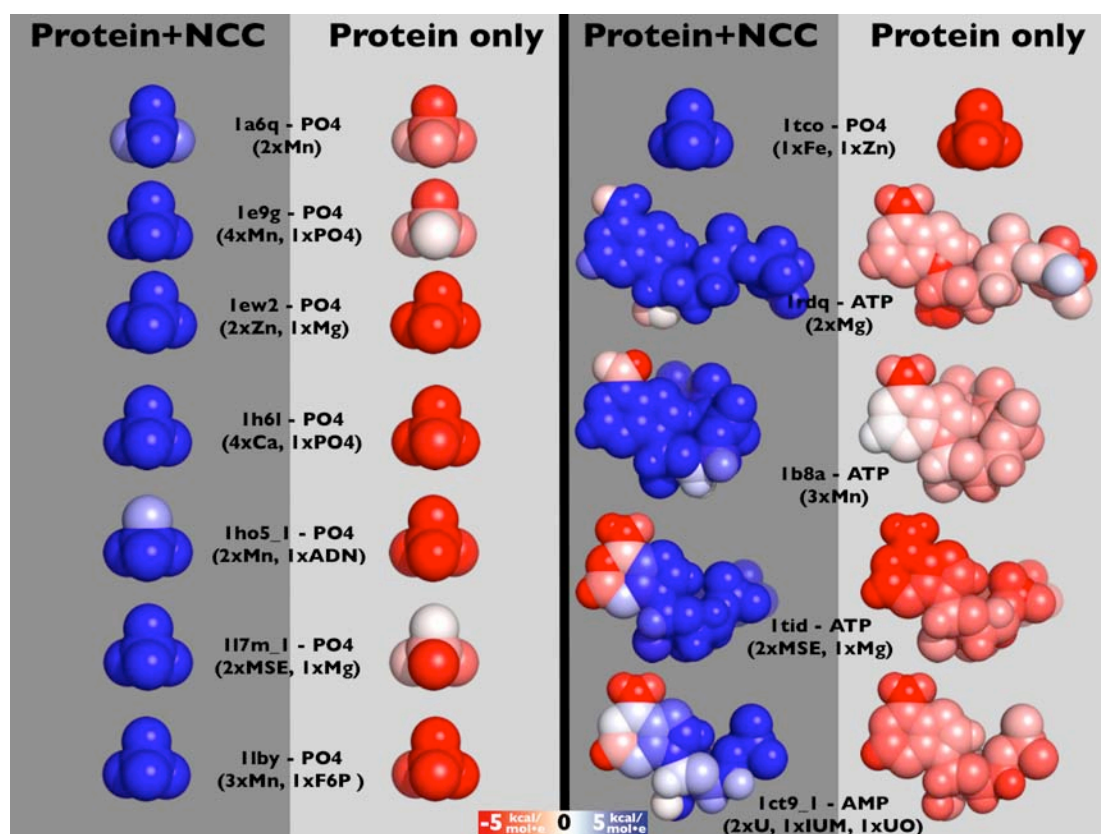
## 4.3 Results

All calculations shown in this section were performed with all five physicochemical properties, i.e. electrostatic potentials, hydrophobicity scores, hydrogen bond donor, acceptor and van der Waals potential energies. However, the discussion of the results focuses on the ElectroStatic Potential (ESP) and the Hydrophobic Environment Score (HES) due to their importance and discriminative power in molecular recognition events.

### 4.3.1 Factors affecting the electrostatic potential

The protein electrostatic potentials on ligand molecules were tested with and without neighbouring chemical components (NCC). NCCs were defined as all HET groups (e.g. metals, cofactors or coenzymes) within 9 Å distance to a ligand in the PQS protein structure (see section 2.1.2). Most common NCCs were found to be charged groups.

In almost all the binding pockets the electrostatic potential was significantly changed by the addition of NCCs. Figure 4.2 shows some extreme examples of the electrostatic potential calculated with and without NCCs for those ligands that experience a potential sign change for at least 80% of their atoms. For example, in the human and bovine serine/threonine phosphatases 1a6q and 1tco respectively, there are highly repulsive forces between the negatively charged phosphate groups and the protein until the effect of the NCC inverts the electrostatic potential. Of the 100 ligands in the data set, 67 had a total of 144 NCCs, of which 63 were metals mostly close to AMP, ATP or phosphate molecules. For the remaining 33 ligands (mostly HEM, NAD and steroid) no NCC was found in the PQS entry but their

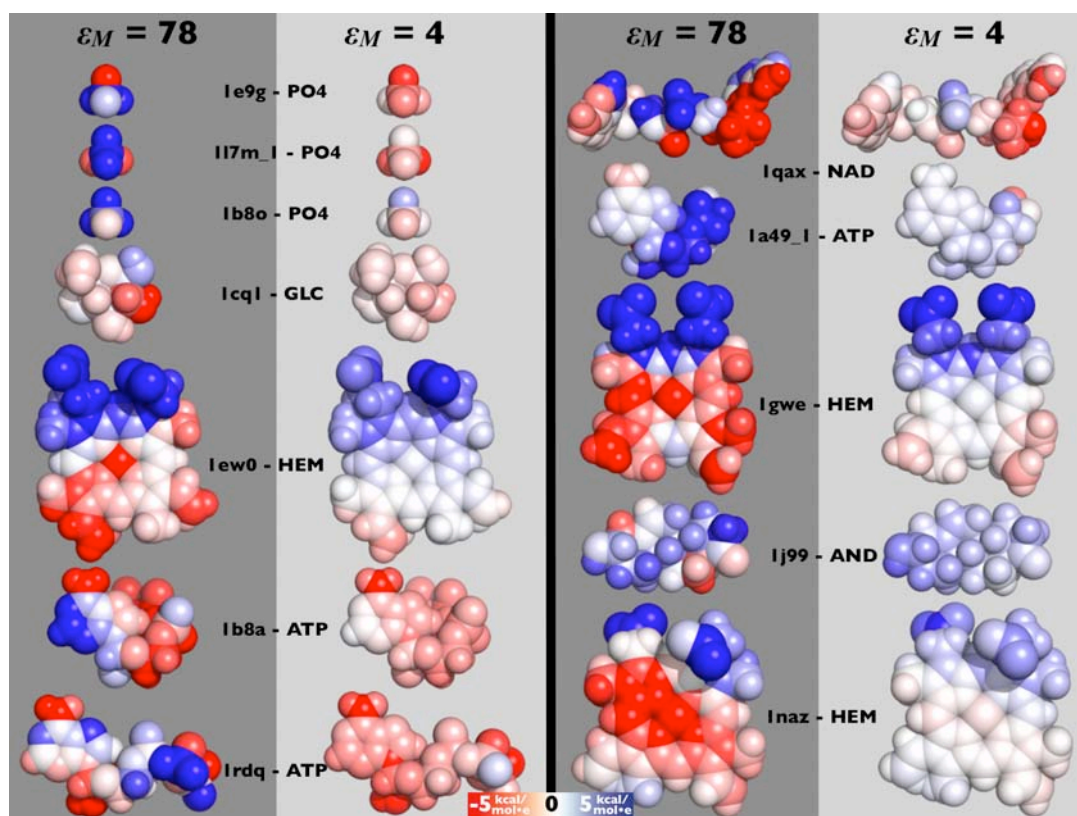


**Figure 4.2: Influence of NCCs on protein's electrostatic potentials**

Electrostatic potentials energies from the protein mapped on ligands with and without Neighbouring Chemical Compounds (NCC). NCCs are all small molecules in the proximity of 9 Å to the ligands above. Depicted are those ligands from Data set I that experience for at least 80% of their atoms a potential sign change upon including NCCs in the potential calculation. The ordering from top left to bottom right is according to the percentage of atoms exhibiting a sign change. The electrostatic potentials are coloured from red to white to blue for potential energies below -5 kcal/mol to 0 kcal/mol to 5 kcal/mol. PQS identifier of the protein structures and the PDB three letter code of the cognate ligand together with the PDB three letter code of the NCCs are given between each pair.

presence *in vivo* can, of course, not be excluded. From these results, it appears that a protein on its own often lacks the required physicochemical properties for the binding of its cognate ligand and requires the assistance of inorganic NCCs (Thompson and Simonson, 2006). For this reason NCCs were included for all the rest of the calculations in this chapter.

Further significant changes in the electrostatic potentials were observed when the dielectric constant of a ligand was changed from the solvent dielectric constant of  $\epsilon_S = 78$  to the protein dielectric constant of  $\epsilon_M = 4$ . This change reduced the screening of electrostatic charges around the ligand by the solvent and allowed charges farther away from the ligand to affect the electrostatic potential at the ligand site. The exclusion of the solvent screening sometimes changes the sign of the potential but more often only strengthens the interaction energy while preserving the sign. Only eleven ligands in Data set I experienced a sign change for more than 25% of their atoms (see Figure 4.3). The best strategy for computing charges and the choice of dielectric is still a matter under debate (Mackerell, 2004; Ponder and Case, 2003), but since the low dielectric constant seems more physically realistic, it was adopted throughout this chapter. For statistics on the influence of partial charges and radius of solvent on Poisson-Boltzmann electrostatics, see (Shen and Wendoloski, 1996).



**Figure 4.3: Influence of dielectric constant on protein's electrostatic potentials.**

Electrostatic potentials from the protein mapped on ligands without neighbouring chemical compounds but different dielectric constant assigned to the ligands. The potentials on the right were calculated with the ligand atoms having a dielectric constant of  $\epsilon_M = 4$ , which excludes the electrostatic screening effect of the solvent at the ligand positions. The potentials on the left mimic an empty and fully solvated binding pocket and were calculated setting the ligand's dielectric constant to the solvent dielectric constant of  $\epsilon_S = 78$ . Depicted are those ligand molecules in Data set I that experience a electrostatic potential sign change upon dielectric constant change for at least 25% of their atoms. The ordering from top left to bottom right is according to the percentage of atoms exhibiting a sign change.

## 4.3.2 Physicochemical properties of proteins in ligand binding sites

### 4.3.2.1 Electrostatic potential on ligands

Figure 4.4, Figure 4.7 and Figure 4.9 show examples of the ESP experienced by all ATP, NAD and heme ligands, respectively, in Data set I. For the ESP of the remaining ligand sets, see Figure 4.10.

**Table 4.1: Average and standard deviation of physicochemical properties.**

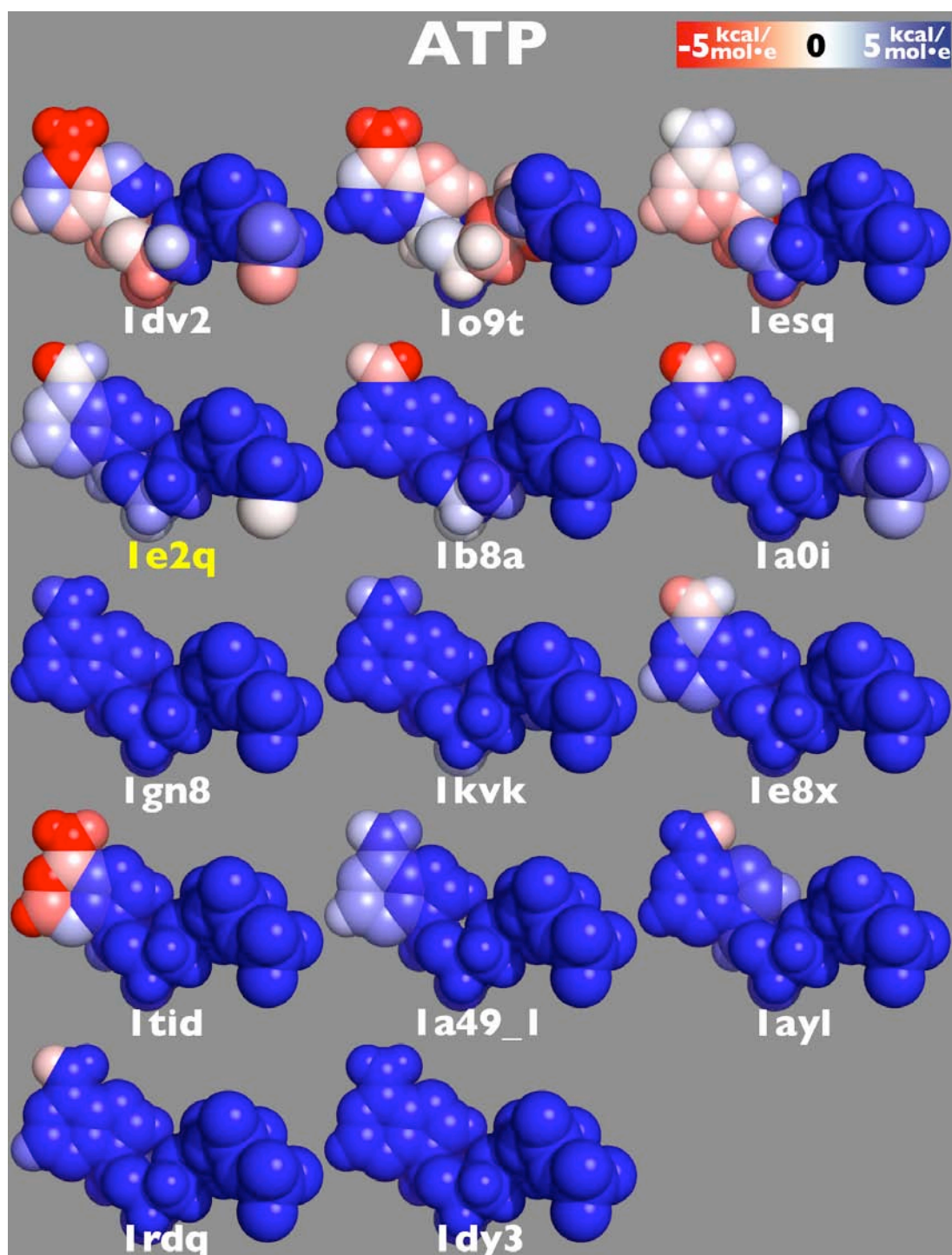
Property	All	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO4	Steroid
Electrostatic Potential (kcal/mol·e)	1.96 ±6.75	1.53 ±5.87	11.33 ±9.13	5.27 ±5.93	4.81 ±7.44	-7.17 ±4.65	-3.11 ±8.30	0.03 ±6.41	17.14 ±14.64	-1.94 ±3.93
Hydrophobicity Environmental Score (HES)	0.70 ±1.53	-0.65 ±2.53	-1.02 ±2.00	0.94 ±1.21	0.00 ±1.39	-0.45 ±1.83	2.30 ±1.46	0.39 ±1.22	-3.65 ±4.24	2.87 ±1.19

Average and standard deviation of the experienced electrostatic potential and hydrophobicity environmental scores for different ligand sets. The lowest and highest average score values for each property is coloured purple and green respectively.

### 4.3.2.2 Adenosine-5'-triphosphate (ATP)

An ATP molecule consists of an aromatic adenine ring, a ribose sugar and a negatively charged triphosphate group. For recognition and binding, the protein is expected to provide a positive electrostatic potential for complementarity to the aromatic ring and in particular the triphosphate tail. Figure 4.4 confirms this expectation showing positive electrostatic potentials for all ATP binding sites.

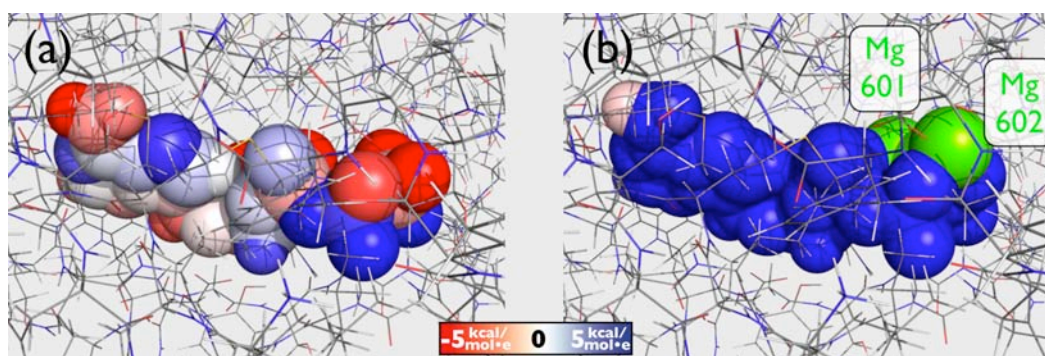




**Figure 4.4: Electrostatic potentials experienced by ATP molecules.**

Electrostatic potentials that ATP molecules ‘experience’ within their binding sites. The ligands are horizontally ordered from top left to bottom right with increasing positive potential. The potential are coloured from red to white to blue for values below -5 kcal/mol·e to 0 to values above 5 kcal/mol·e. The PQS-ID of the protein structure in which the ligand was found is given below each ligand. Note that to make the visual comparison easier, the electrostatic potentials of each molecule were mapped on the representative conformation of 1e2q, which has a yellow label.

An inspection of the ATP binding sites revealed that the high positive potentials on the ATP molecules (on average 11.33 kcal/mol·e, see Table 4.1) were caused primarily by metal ions that are coordinated to one or two phosphates at ATP's triphosphate tail. Metal ions were found coordinated in all ATP binding sites except in the DNA ligase (1a0i) and the biotin carboxylase (1dv2). The function of those metals varies but often they act as catalytic stabilizer for transition states of the enzyme-substrate complex (Gonzalez, et al., 2003; Larsen, et al., 1998) or as charge neutralizers between charged protein and phosphate groups (Schmitt, et al., 1998; Zheng, et al., 1993). In particular, the latter function is important with respect to the molecular recognition of ATP as metal ions can shield negative charges between the protein and ATP and lock the otherwise loose triphosphate tail of ATP to the protein (Bilwes, *et al.*, 2001; Masuda, *et al.*, 2004). The charge neutralizing effect can increase ATP's binding affinity by several orders of magnitude, as in the case of the bovine cAMP-dependent protein kinase (Armstrong, et al., 1979). The mouse homologue of the same protein, 1rdq, in Data set I shows, after the exclusion of the divalent magnesium ion from the ESP calculation, negative repulsive potentials towards the ATP molecule. Only when the metal ion is included in the calculation, are repulsive forces neutralized and become attractive (see Figure 4.5). Similar neutralizing effects of metal ions were observed in the PQS structures 1a49, 1b8a, 1e8x, 1o9t and 1tid.



**Figure 4.5: Influence of metal ions on ATP's experienced electrostatic potential.**

Electrostatic potential of the mouse cAMP-dependent protein kinase (PQS-Id 1rdq) as experienced by the cognate ligand ATP (a) without and with (b) two coordinated magnesium ions (green coloured spheres). The potential are coloured from red to white to blue for values below -5 kcal/mol·e to 0 to values above 5 kcal/mol·e

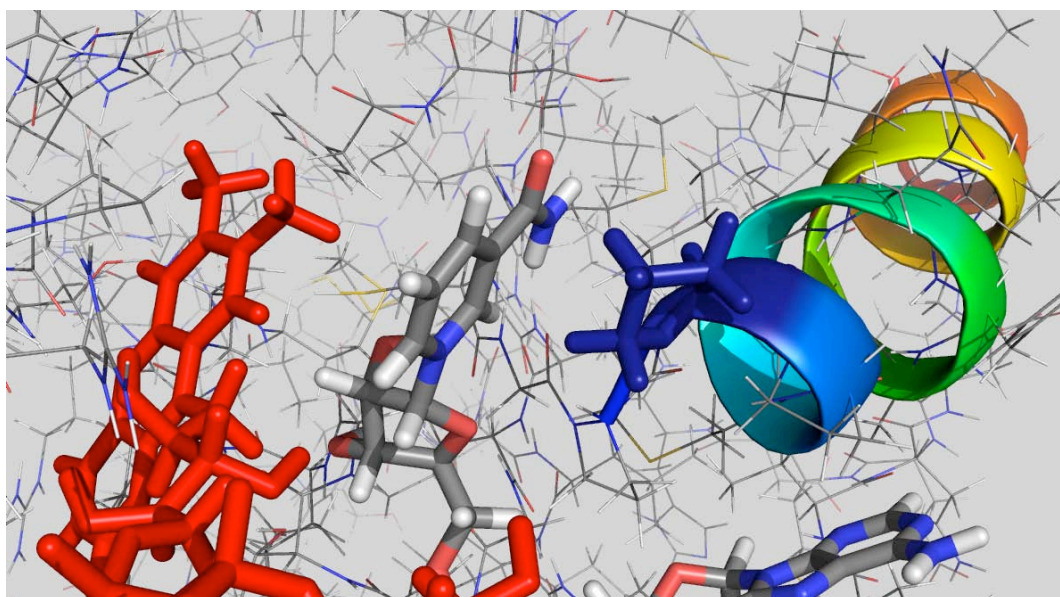
Not all ATP molecules require a metal ion to bind to their receptor protein. As mentioned above, ATPs in the binding sites of the DNA ligase (1a0i) and the biotin carboxylase (1dv2) were found without a metal ion or any other NCCs. Although they are important for the catalytic reaction of these enzymes magnesium ions are not essential for ATP binding. A detailed inspection of their binding sites revealed that the proteins created attractive positive electrostatic potentials (see Figure 4.4) towards the ATP phosphate tails through phosphate binding motifs. These motifs are characterized by positive electrostatic fields originating from positively polarized N-termini of  $\alpha$ -helices and/or positively charged lysine and arginine side chain and/or ionic hydrogen bonds ( $O^- - HN$ ) between backbone NH groups and phosphate oxygen ions (Hirsch, et al., 2007).

### 4.3.2.3 Nicotinamide adenine dinucleotide (NAD)

NAD molecules consist of three functional groups, namely the mainly aromatic adenosine moiety, the negatively charged pyrophosphate group and the aromatic nicotinamide ring that can carry a formal positive charge when oxidised to  $NAD^+$  (Smith and Tanner, 2000). NAD molecules can have various functions in enzyme reactions. In Data set I, the two most prominent functions are the redox function in oxidoreductases (1hex, 1ib0, 1jq5, 1mew, 1mi3\_1, 1o04\_1, 1qax, 1t2d, 2npx) and the group transfer function in transferases (1ej2, 1og3, 1s7g, 1tox\_1, 2a5f). As a redox partner in NAD-dependent dehydrogenases, a  $NAD^+$  molecule supports the oxidation of a substrate by accepting a hydride group ( $H^-$ ). As a substrate in ADP-ribosyltransferase reactions,  $NAD^+$  molecules are cleaved into an ADP-ribosyl and a nicotinamide group, where the former is transferred to an arginine group within the active site as part of a posttranslational modification and the latter is released into the solvent. Both enzyme reactions are distinct from each other and should result in distinct ESP on the NAD molecules, in particular on the nicotinamide moieties. For dehydrogenases, the nicotinamide moiety should experience negative ESP due to the  $H^-$  group transfer, while in ribosyl-transfer reactions the positively charged nicotinamide moiety should experience

repulsive positive ESP. Figure 4.7 shows the ESP of all NAD molecules in Data set I ordered by the total ESP experienced by the molecules from top left to bottom right. NAD molecules that are functioning as oxidation partners are labelled green while molecules acting as substrates in transferase reactions are labelled orange.

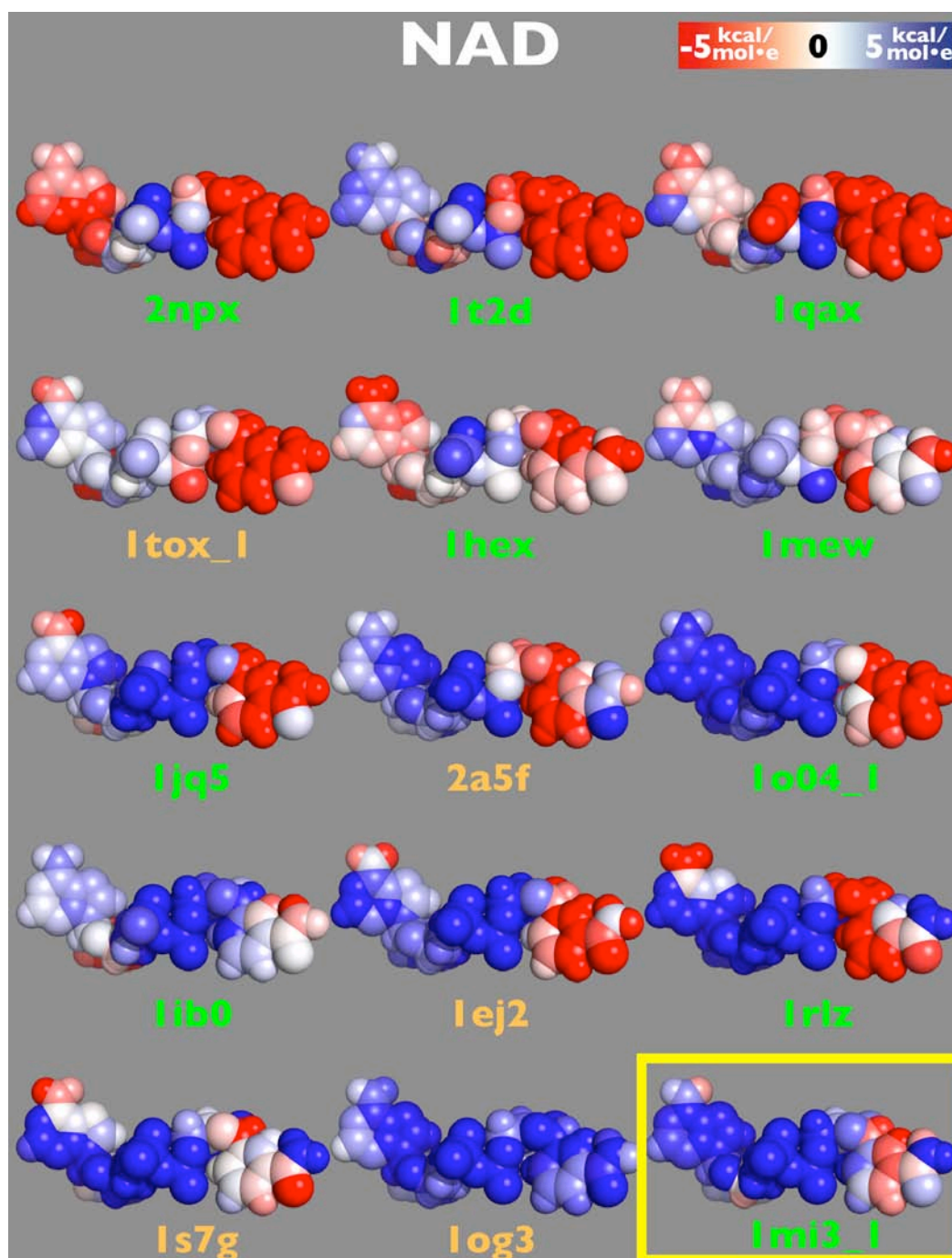
All NAD molecules bound to oxidoreductases confirm the expectation and show a negative ESP at the majority of their nicotinamide moiety with the exception of 1ib0, a cytochrome b5 reductase, which experiences positive ESP. The repulsive ESP was found to be a functional necessity, as 1ib0 in contrast to the other oxidoreductases is a reductase and not a dehydrogenase. In reductases NAD molecules become oxidized i.e. provide rather than accept a  $H^+$  group. To ease the transfer of the  $H^+$  group from NADH to the neighbouring cofactor FAD in 1ib0, a positive ESP was found to expand from a positively polarized N-terminus of an adjoining  $\alpha$ -helix to the positively partial charged succinimidyl group in the flavin moiety of the FAD molecule (see Figure 4.6).



**Figure 4.6: Functional necessary repulsive forces in cytochrome b5 reductase 1ib0.**

The Cytochrome b5 reductase structure 1ib0 is the only oxidoreductase in Data set I that exposes the nicotinamide moiety of its substrate NADH (vary-coloured molecule) with a positive electrostatic potential. The positive potential is created by the positive polarized N-terminus of an adjoining  $\alpha$ -helix, and is functionally required to ease the transfer of a  $H^+$  group from the NADH molecule to a neighbouring FAD molecule. The nicotinamide moiety is stacked between a flavin group of a neighbouring FAD molecule (red coloured) and a proline residue (blue coloured).

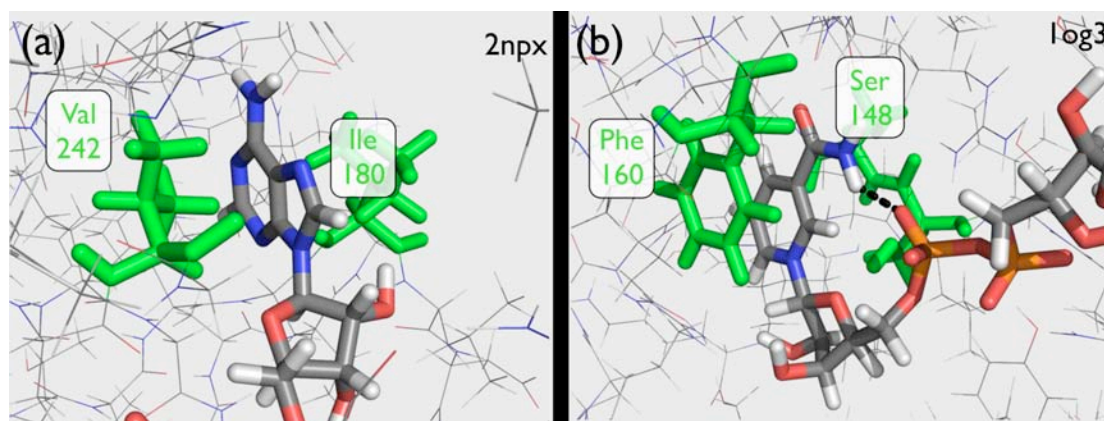
The transferase structure of the ADP-ribosyltransferase 1og3 and partially the structure of the NAD-dependent deacetylase 1s7g exert as expected electropositive potentials towards the oxidized nicotinamide moiety of  $\text{NAD}^+$ . The structure of the nicotinamide mononucleotide adenylyltransferase 1ej2 however exerts negative ESP. Although this might seem in contrary to 1og3 and 1s7g, it conforms to the function of 1ej2 to synthesise  $\text{NAD}^+$  molecules. While transferases discard the nicotinamide moiety for which repulsive forces are advantageous, 1ej2 tightly binds the nicotinamide moiety in order to fuse it to an ATP substrate molecule. For the negative ESP of the diphtheria toxin structure 1tox and in the cholera toxin structure 2a5f, however, no functional basis could be found. Both proteins cleave the nicotinamide moiety and thus should exert an electropositive potential and not as observed a complementary attractive electronegative potential. It remains to be tested whether the large conformational changes that both proteins undergo upon NAD binding (Bell and Eisenberg, 1996; O'neal, *et al.*, 2005) lead to changes in the protein environment that allow the protein to form repulsive ESP against the positively charged nicotinamide moiety.



**Figure 4.7: Electrostatic potentials experienced by NAD molecules.**

Electrostatic potentials that NAD molecules ‘experience’ within their binding sites. The ligands are horizontally ordered from top left to bottom right with increasing positive potential. The potential are coloured from red to white to blue for values below -5 kcal/mol·e to 0 to values above 5 kcal/mol·e. The PQS-Id of the protein structure in which the ligand was found is given below each ligand. Note that to make the visual comparison easier, the electrostatic potentials of each molecule were mapped on the representative conformation of 1mi3\_1, which has been encircled with a yellow box. NAD molecules bound to oxidoreductases are labelled green; those bound to transferases are labelled orange.

For the bacterial NADH peroxidase (2npx) repulsive potentials were found for the adenine moiety of NADH (2npx is the only protein in Data set I having bound NADH. All other NAD proteins have bound the oxidized NAD<sup>+</sup>). The negative electrostatic field is created by six negatively charged glutamate and aspartate amino acids in close proximity. A close inspection of the residues around the nicotinamide reveals that the repulsive potentials are overridden by aromatic interactions to a hydrophobic valine and an isoleucine side chain (see Figure 4.8a). Similar aromatic interactions were found for the nicotinamide group of 1og3 (see Figure 4.8b), where  $\pi$ - $\pi$  stacking interaction to a phenylalanine from one side and an aromatic-backbone-amide interaction from the other side provide the necessary binding free energy for the nicotinamide moiety of the NAD molecule. The importance of aromatic interactions in Data set I will be described later in the text. Similar observations were done on adenine groups that were found to bind protein binding sites primarily with intermolecular hydrogen bonds together with  $\pi$ - $\pi$  stacking and cation- $\pi$  interactions (Denessiouk, *et al.*, 2001; Mao, *et al.*, 2004).



**Figure 4.8: Aromatic interactions compensate repulsive electrostatic interactions.**

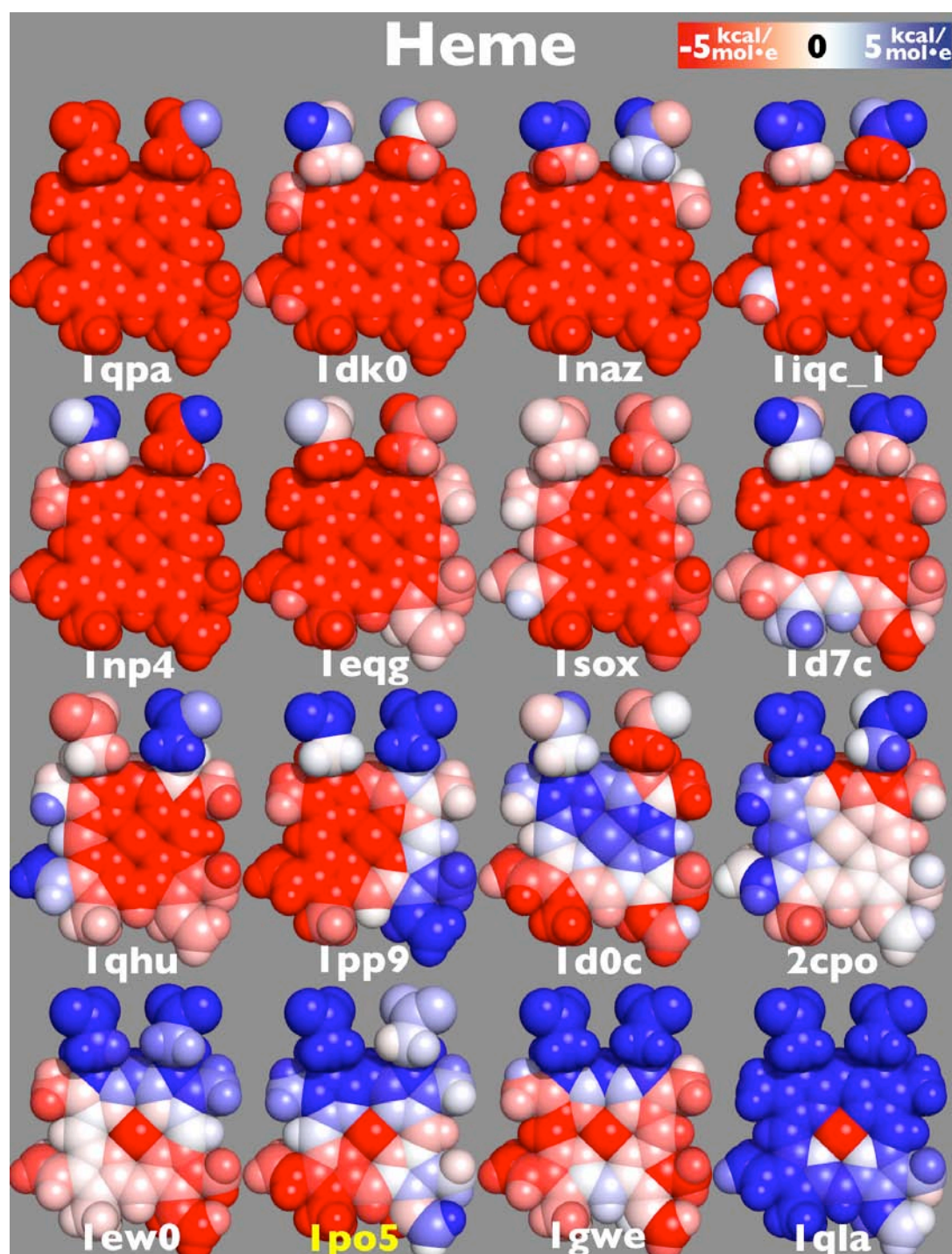
(a) The repulsive electrostatic potentials of NADH peroxidase towards the adenine moiety of NADH are compensated via  $\pi$ -CH interactions between the adenine moiety (varicoloured) and Ile 180 (green coloured) and Val 242 (green coloured). (b) ADP-ribosyltransferase stabilises the binding of NAD<sup>+</sup> via a  $\pi$ - $\pi$  interaction with Phe160 (green coloured),  $\pi$ -NH interaction with the backbone amide of Ser148 (green coloured) and an internal hydrogen bond (black dashed line) between the amide group and the phosphate group.

#### 4.3.2.4 Heme type B

The heme molecule consists of an aromatic porphyrin ring system with a central coordinated iron ion and two negatively charged propionate groups sitting 'on top' of the porphyrin ring system. Arginine residues form ionic bonds to the propionate groups and anchor the heme molecule to the protein structure, whilst histidine and less frequently methionine residues coordinate the heme's iron ion and affect the electrochemical character of the iron metal ion.

Based on these characteristics one would expect positive electrostatic potential around the propionate groups and negative potential in the porphyrin ring system around the iron ion. Visually this prediction seemed true for the PQS structures 1dk0, 1iqc\_1, 1naz and 1np4. However the remaining heme binding sites, in particular 1d0c, 1ew0, 1po5 and 2cpo, show repulsive positive electrostatic fields around the heme's porphyrin group, causing the heme molecules to have a diverse range of ESP values with an average of -3.11 kcal/mol·e and a standard deviation of 8.30 kcal/mol·e (see Table 4.1). Heme molecules are bound by more than 20 different protein folds (Schneider, et al., 2007) leading to a large diversity of environments around the molecules. In depth inspection of the protein environments around the heme molecules show that, like NAD's nicotinamide moiety, repulsive electrostatic potentials are overridden by aromatic interactions. Aromatic interactions are well known to contribute strongly to the binding energy of a heme molecule (Reedy and Gibney, 2004; Roberts and Montfort, 2007; Schneider, et al., 2007). An important electrochemical characteristic of heme molecules is their reduction potential which can vary between -550mV (in hemophore HasA) and +362mV (in cytochrome *f*) versus the standard hydrogen electrode (Reedy and Gibney, 2004). Relating the electrostatic potential to the reduction potential is not straightforward as the reduction potential is influenced by many different factors such as the type of the heme's axial ligands, heme burial, local charges etc. (Gunner and Honig, 1991; Mao, et al., 2003; Reedy and Gibney, 2004). Nevertheless, my calculations show that there is a modest negative correlation ( $R^2 = 0.5$ ) between ESP and the experimental reduction potential.





**Figure 4.9: Electrostatic potentials experienced by heme molecules.**

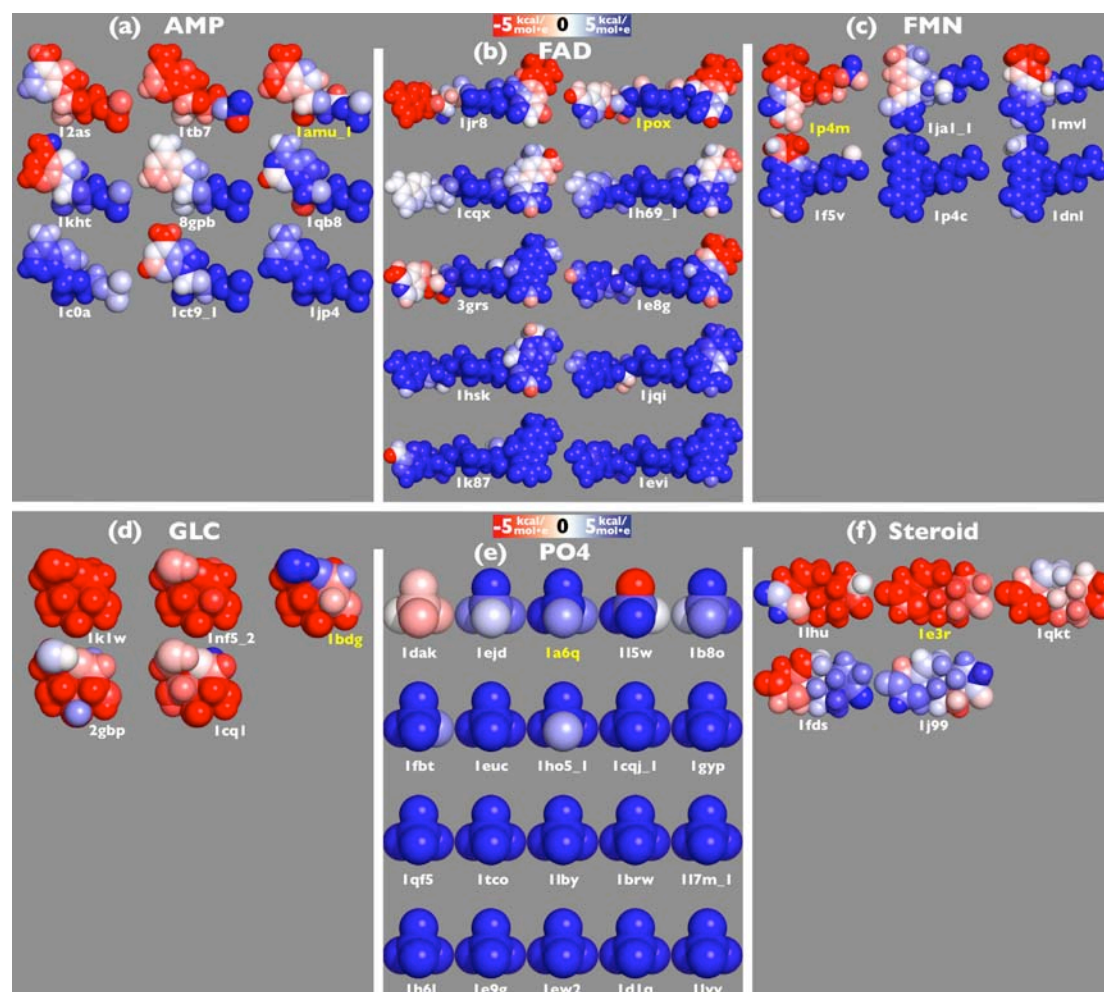
Electrostatic potentials that heme molecules ‘experience’ within their binding sites. The ligands are horizontally ordered from top left to bottom right with increasing positive potential. The potentials are coloured from red to white to blue for values below  $-5$  kcal/mol·e to 0 to values above  $5$  kcal/mol·e. The PQS-ID of the protein structure in which the ligand was found is given below each ligand. Note that to make the visual comparison easier, the electrostatic potentials of each molecule were mapped on the representative conformation of 1po5, which has a yellow label.

### 4.3.2.5 Remaining ligand sets

AMP binding sites often generate positive electrostatic potentials towards their ligand's phosphate group (see Figure 4.10). Five of the proteins, namely 1amu, 1ct9\_1, 1jp4, 1qb8 and 1tb7 had at least one metal ion bound within 9 Å proximity inducing a positive potential on the phosphate tail. The AMP molecule in asparagine synthetase (12as) compensated the repulsive global negative electrostatic potentials through a local network of eight hydrogen bonds. The flavin moiety in FAD as well as in FMN experienced both negative and positive potentials in Data set I. The phosphate group in both ligands was however always attracted by positive potentials. In the riboflavin kinase (1p4m), the pyrophosphate of the product ADP located next to FMN caused the repulsive negative potential at the phosphate group. The negative potential on the aromatic ring of the same ligand was compensated by aromatic interactions.

Glucose binding sites were electrostatically most negative in Data set I with an average electrostatic potential of -7.17 kcal/mol·e (see Table 4.1). In contrast, phosphate binding sites were most positive with an average potential of 17.14 kcal/mol·e (see Table 4.1). Metal ions were found within 9 Å proximity in nine cases (1a6q, 1e9g, 1ew2, 1h6l, 1ho5\_1, 1l7m\_1, 1lby, 1qf5, 1tco) generating positive potentials around the PO<sub>4</sub> molecules. An interesting exception was the binding site of the bacterial dethiobiotin synthetase (1dak). This particular binding site apparently completely lacked the expected physicochemical properties to bind a phosphate molecule. Not only did it exert repulsive negative potentials, but it also lacked any compensating polar interactions and/or hydrogen bonds towards its ligand. Also molecular recognition based purely on van der Waals interaction could be ruled out as the phosphate molecule was solvent accessible on up to 70% of its molecular surface. It may be possible to explain these discrepancies by examining the experimental data for this protein structure. Unfortunately, the structure factors for 1dak were not available at neither the PDB nor the Electron Density Server (Kleywegt, et al., 2004), preventing further investigations of this case. And finally, as for heme molecules, the molecular recognition of steroid molecules is driven by

hydrophobic and aromatic interactions (Wallimann, et al., 1997) that override the observed repulsive ESP in steroid binding sites in the data set.



**Figure 4.10: Electrostatic potentials experienced by remaining ligand sets.**

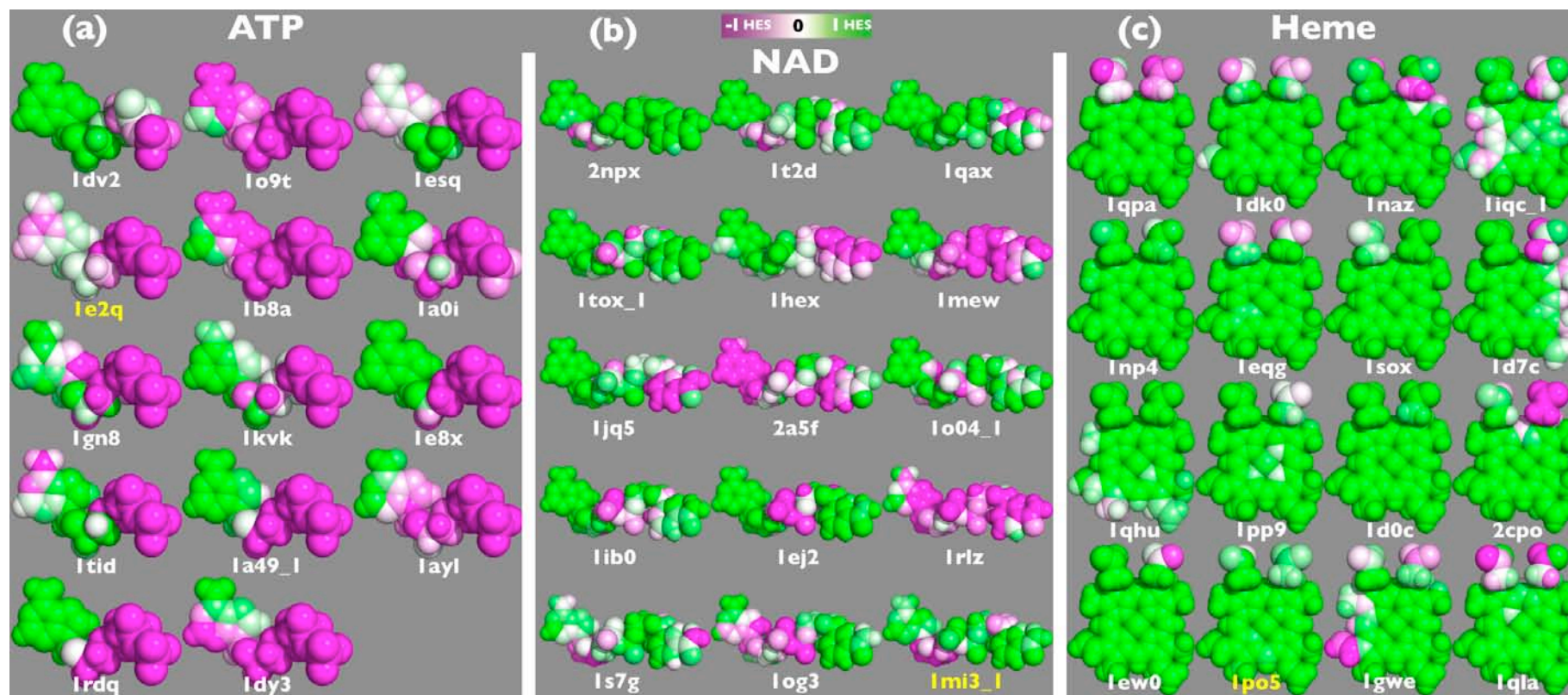
Electrostatic potentials that (a) AMP, (b) FAD, (c) FMN, (d) GLC, (e) PO4 and (f) steroid molecules 'experience' within their binding sites. The ligands are horizontally ordered from top left to bottom right with increasing positive potentials. The potentials are coloured from red to white to blue for values below -5 kcal/mol to 0 to values above 5 kcal/mol. The PQS-Id of the protein structure in which the ligand was found is given below each ligand. Note that, to make the comparison easier, each type of molecule is represented by a single conformer and may not be the conformer in the given PQS file. The representative conformers are those labelled in yellow.

### 4.3.3 Non-electrostatic interactions between protein and ligand

The results above show that binding events are not solely directed by classical electrostatic interactions between partially charged atoms, as there are many examples of non-complementary electrostatic potentials between protein and bound ligand. Thus other, non-electrostatic forces, acting between ligand and protein can result in negative binding free energies and lead to the binding of the ligand to its binding site. With the exception of glucose and phosphate, all the ligands possess at least one aromatic ring complex and most of the variation in terms of electrostatic potential is observed at these ring systems. Ligand moieties that are highly charged, like the phosphate groups, experience rather constant complementary electrostatic fields. Close inspections of the binding site ligand complexes in Data set I showed that almost all aromatic ring groups undergo various forms of aromatic interactions. Most of these interactions were of a hydrophobic nature towards aliphatic hydrocarbons like valine, leucine or isoleucine side chains, forming  $\pi$ -CH interactions (Tsuzuki and Fujii, 2008; Tsuzuki, *et al.*, 2000b). The second most often observed aromatic interactions were  $\pi$ - $\pi$  interactions between aromatic ring groups mostly from phenylalanine and tyrosine residues (Hunter, *et al.*, 2001). Other aromatic interaction types were  $\pi$ -cation interactions between aromatic rings and positively charged side chains of arginine and lysine (Cauet, *et al.*, 2005) and  $\pi$ -HN interactions with backbone amide groups that act as a hydrogen bond donor towards an aromatic ring (Levitt and Perutz, 1988). Also observed were  $\pi$ -proline interactions (Toth, *et al.*, 2001). Aromatic interactions, in particular  $\pi$ -CH and  $\pi$ - $\pi$ , are in general dominated by London dispersion and hydrophobic interactions (Luo, *et al.*, 2001; Marsili, *et al.*, 2008). Electrostatic interactions play only a secondary role and only contribute where aromatic interactions involve protein hydroxyl or amine groups that have a high electrostatic dipole moment (Tsuzuki, *et al.*, 2000a). Accurate dispersion energy calculations require sophisticated and computationally demanding quantum chemistry

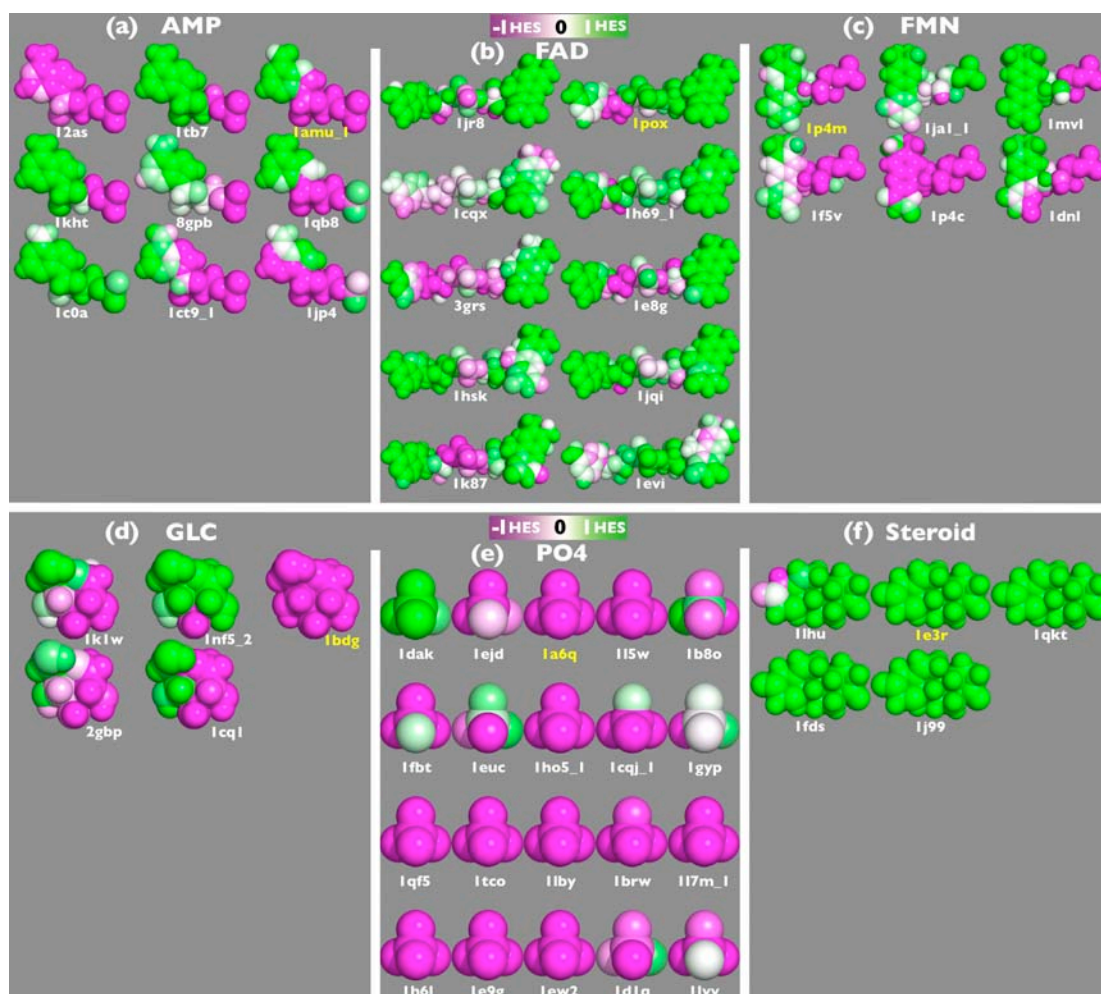
calculations (Luo, et al., 2001) that are beyond the scope of this work. Instead, the hydrophobicity of the ligand environment within protein binding sites was analysed.

Figure 4.11 depicts the Hydrophobic Environment Score (HES) (see section 4.2.3) for the binding sites in Figure 4.4, Figure 4.7 and Figure 4.9, coloured from magenta to white to green for -1 to 0 to 1 environment score for polar, neutral and hydrophobic environments, respectively. From Figure 4.11a, it is clear that aromatic ring systems are generally located in a hydrophobic environment. In particular, the variation observed for ESP for heme binding sites does not occur for HES. All heme molecules are located entirely in hydrophobic binding pockets whether facing attractive or repulsive ESP. Similar dominant HES were observed for steroid molecules and flavin moieties (see Figure 4.12). The adenine moieties of ATP and NAD were usually found in a hydrophobic environment whereas their charged phosphate groups tend to face a polar environment. The NAD molecules (PQS Ids 1ibo, 1hex, 1mew, 1og3, 1qax, 2npx) that were found to experience repulsive electrostatic forces at their adenine and nicotinamide moieties simultaneously form attractive aromatic-hydrophobic interactions.



**Figure 4.11: Hydrophobicity experienced by ATP, NAD, heme molecules.**

Hydrophobicity experienced by (a) ATP, (b) NAD and (c) heme molecules within their binding pockets. The level of experienced hydrophobicity is given by the Hydrophobicity Environmental Scores (HES) (see section 4.2.3) and is coloured from magenta to white to green for polar environments with less than -1 HES to 0 to hydrophobic environments with values above 1 HES. The ligands are ordered as in Figure 4.11. The PQS-Id of the protein structure in which the ligand was found is given below each ligand. The representative conformer for each set has a yellow label.



**Figure 4.12: Hydrophobicity experienced by remaining ligand sets.**

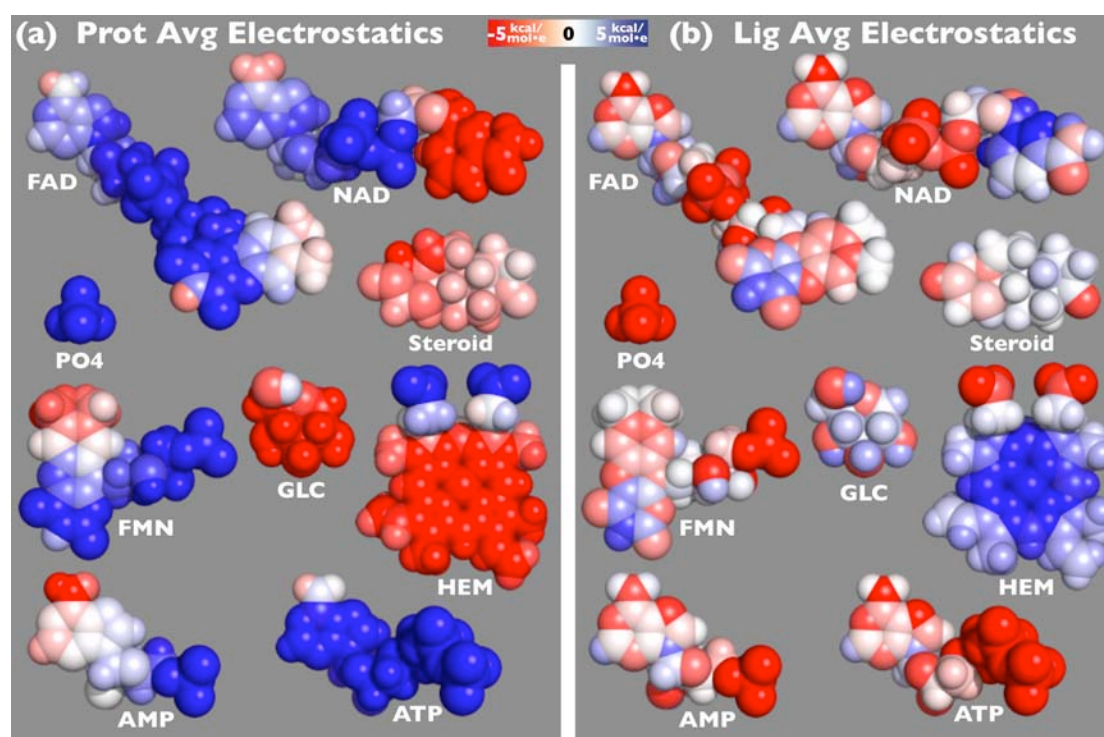
Hydrophobicity experienced by (a) AMP, (b) FAD, (c) FMN, (d) GLC, (e) PO<sub>4</sub> and (f) Steroid molecules within their binding pockets. The level of experienced hydrophobicity is given by the Hydrophobicity Environmental Scores (HES) (see section 4.2.3) and is coloured from magenta to white to green for polar environments with less than -1 HES to 0 to hydrophobic environments with values above 1 HES. The ordering of the ligands is adopted from Figure 4.10. The PQS-Id of the protein structure in which the ligand was found is given below each ligand. The representative conformer for each set has a yellow label.

#### 4.3.4 Average and variation of physicochemical properties

Despite differences in the conformations of a ligand, one can calculate the average ESP and HES and its standard deviation for each ligand atom in the protein binding pocket. The

average atomic ESP and HES were calculated for each ligand set and plotted in Figure 4.13 and Figure 4.14. Next to the figures are depicted ESP and HES generated by the ligands in isolation. In contrast to many single cases (see NAD in Figure 4.7 and heme in Figure 4.9), the average values of the physicochemical properties show complementary characteristics between protein and ligand for the majority of the cases. From Figure 4.13 it becomes evident that usually electrostatically positive protein potentials are neutralised by electrostatically negative ligand potentials and vice versa.

A closer look to the ESP in Figure 4.13a reveals average positive electrostatic potentials for all adenine groups and all phosphate groups, whether as separate entities or part of larger molecules. Note that despite the similarity between AMP and ATP, the latter is usually bound in pockets with higher positive potential due to the larger number of metal ions coordinated to the phosphate tail of ATP. The highly hydrophobic heme and steroid molecules, as well as glucose and the nicotinamide moiety in NAD, are mainly surrounded by negative potential.



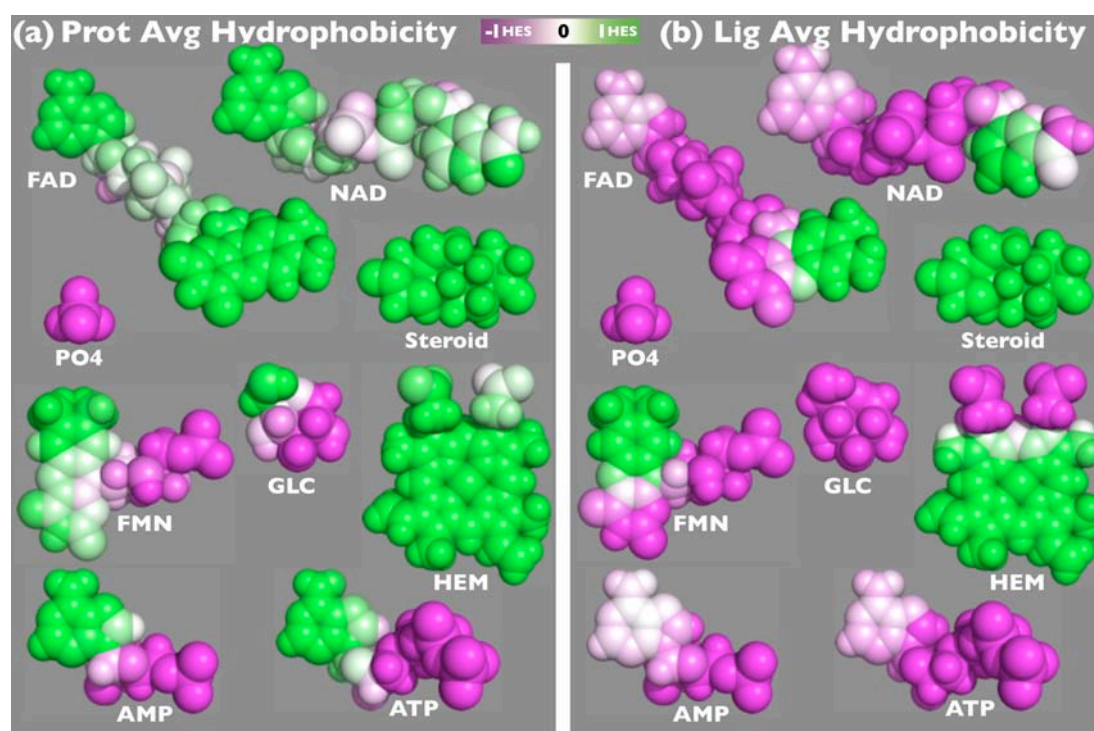
**Figure 4.13: Average electrostatic potentials for Data set I.**

(a) Average electrostatic potentials that each ligand set in the data set 'feels' within protein binding sites.  
(b) Electrostatic potentials calculated for each ligand molecule in isolation. The electrostatic potentials are coloured from red to white to blue for values below  $-5$  kcal/mol to 0 to above 5 kcal/mol.



Figure 4.14a shows that steroids and the porphyrin core of heme molecules are bound in very hydrophobic environments together with the flavin moieties in FAD and FMN. Due to the large number of  $\pi$ -CH and  $\pi$ - $\pi$  interactions involving adenine moieties in the data set, these moieties experience large HES despite being of moderate polar nature.

The comparison of the averaged electrostatic potentials from the proteins (Figure 4.13a) and the ligands (Figure 4.13b) on each ligand atom reveals a correlation of only  $R^2 = 0.25$ . The correlation for HES is higher with  $R^2 = 0.66$ . Both numbers are relatively small and indicate rather little interdependence between the protein's or ligand's physicochemical properties. However they still exceed the correlation between all 100 individual binding-site/ligand complexes, with  $R^2 = 0.14$  for ESP and  $R^2 = 0.35$  for HES, demonstrating that the average physicochemical properties show a higher complementarity than individual protein-ligand



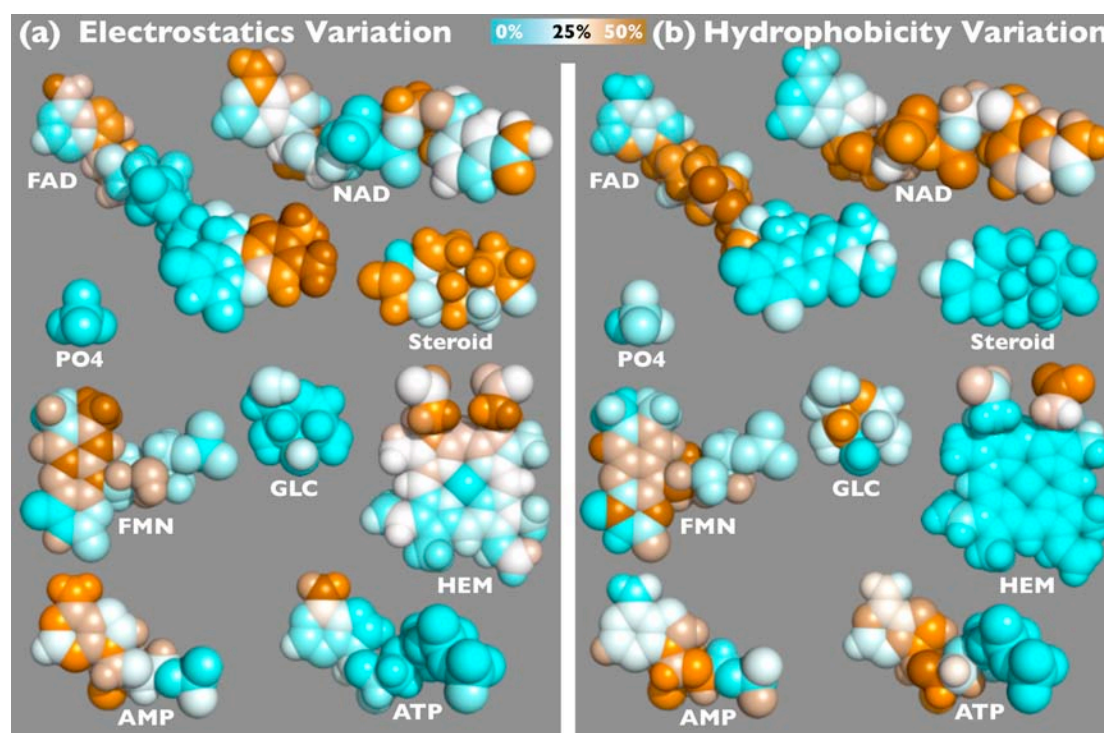
**Figure 4.14: Average hydrophobicity scores for Data set I.**

(a) Average hydrophobicity that each ligand set in the data set 'experiences' within protein binding sites. (b) Hydrophobicity of the ligand molecule in isolation. The level of experienced hydrophobicity is given by the Hydrophobicity Environmental Scores (HES) (see section 4.2.3) and is coloured here from magenta to white to green for polar environments with less than -1 HES to 0 to hydrophobic environments with values above 1 HES.

pairs.

The standard deviation of the absolute ESP and the HES are given in Table 4.1. ESP shows large variations with standard deviations several times larger than the average value. The highest variation is seen for phosphate binding pockets having a standard deviation of 14.64 kcal/mol·e, whilst the lowest variation is seen in steroid binding pockets with 3.93 kcal/mol·e. The variation of HES is generally less and is around four times smaller than the average standard deviation of ESP.

Although the absolute values of ESP and HES have a high standard deviation and thus are very variable, their values are consistently of the same sign. Figure 4.15 shows for each ligand atom the Sign Change Ratio (SCR) (see section 4.2.5) for ESP and HES. Cyan colour indicates no variation (0%) for a ligand atom with score values all with the same sign; the



**Figure 4.15: Variation of physicochemical properties in protein binding pockets.**

Variation of physicochemical property scores measured by the sign change ratio of the scores at each atom. Variation is coloured from cyan to white to orange for relative sign-changes of 0% to 25% to 50%. 0% represents no score variation for the property. 50% illustrates highest variation with half of the ligand set having positive and half of the ligand set having negative score values.

highest variation (50%) is coloured orange reflecting half positive and half negative score values. According to Figure 4.15 half of all ligand atoms in Data set I experience electrostatic potential and hydrophobicity scores of low SCR whereas half experience high SCR. Hydrophobic moieties in ligand molecules often occupy equivalent hydrophobic environments, like the flavin groups in FAD, FMN or the porphyrin core in heme molecules or steroid molecules, but have their electrostatic environments change. Similar observations can be made for central phosphate groups in FAD and NAD. Thus, the interactions are on average complementary and vary widely in their magnitude.

### 4.3.5 Comparison of properties between ligand sets

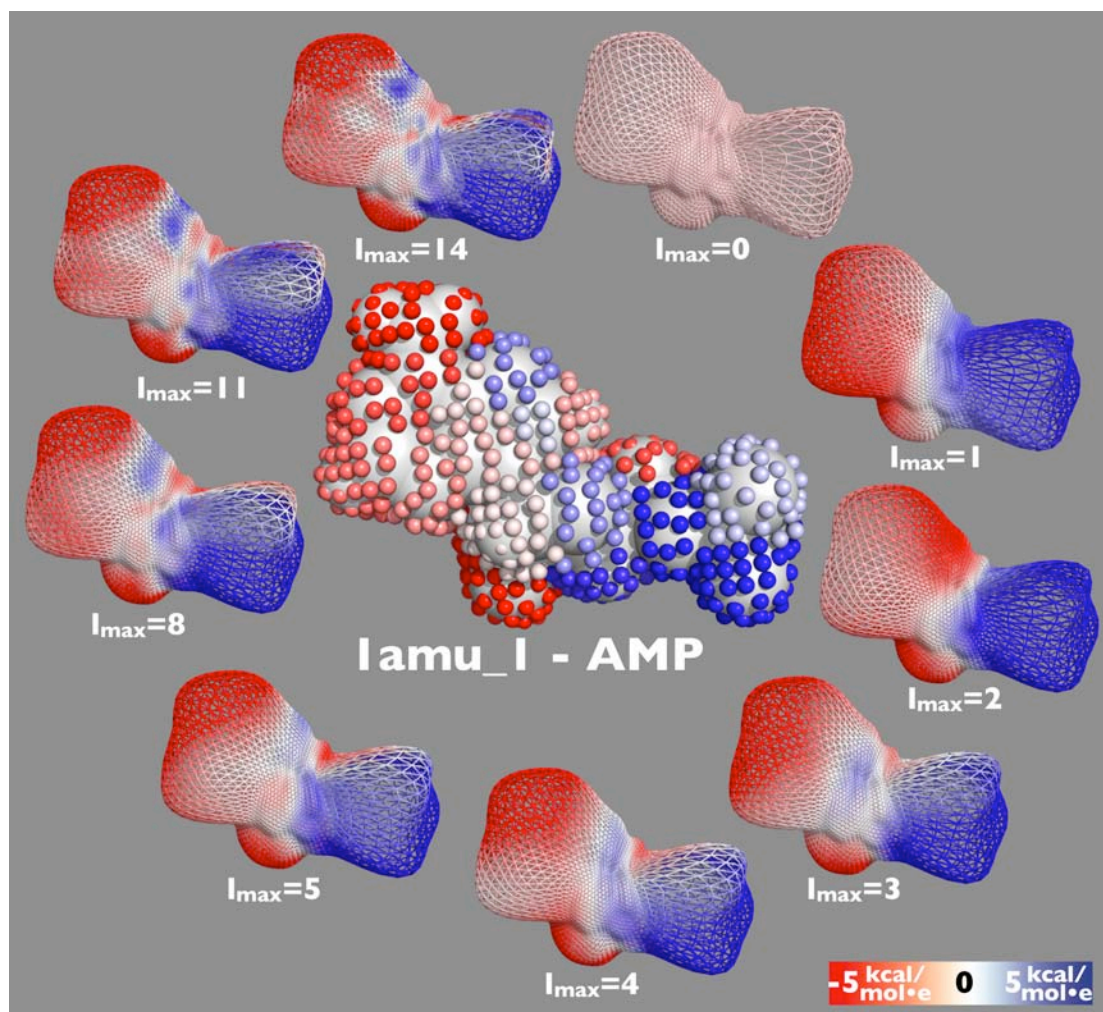
The calculation of ESP and HES for each ligand atom allowed the comparison of physicochemical properties within ligand sets, but not among ligand sets, as there are no complete atom-atom correspondences between different ligands. However, to answer questions such as whether e.g. AMP binding pockets possess similar physicochemical properties to steroid binding pockets, or whether physicochemical properties of binding pockets allow the prediction of its ligand, it was essential to compare ligand sets with each other.

In Chapter 3, I have presented a framework for the description and comparison of ligand and cleft shapes employing spherical harmonics (Morris, 2006; Morris, *et al.*, 2005). For the problem stated above the same framework was utilised to describe the distribution of the physicochemical properties that ligand experience from their proteins. The property scores were considered as a function on a unit sphere allowing their expansion with spherical harmonic functions. Note that in such expansions  $l_{\max} = 0$  gives a weighted average property score and  $l_{\max} = 1$  describes the general tendency of the score distribution (see section 2.5.2.3.2 and Figure 4.16). The expansion leads to a coefficient vector of 225 numbers that

uniquely and analytically describe the distribution of the ESP and HES independent of shape and size (disregarding the implicit incorporation of shape and size when measuring the physicochemical properties on the ligand molecule). A Euclidean metric applied on pairs of coefficient vectors was employed to compare pairs of physicochemical property distributions. In order to compare the physicochemical properties independent of the ligand conformations, the ESP and HES experienced by each ligand set member were mapped on a ligand set representative, similar to Figure 4.4. Furthermore, to satisfy the observations that the property scores vary less in their sign than in their absolute values, a cut-off was applied on ESP below -1 kcal/mol·e and above 1 kcal/mol·e. The same cut-off was utilised on HES scores below -1 and above 1. For the shape description of the binding pockets a new cleft model, the convex cleft model, was implemented as a fourth cleft model to the already existing Conserved, Interact and Ligand Cleft Models in CleftXplorer. The Convex Cleft Model bases on the Interact Cleft Model but in addition excludes all SURFNET spheres that are not enclosed by the convex hull (see section 2.3.2.2) of the ligand interacting protein atoms. The Convex Cleft Model leads generally to a reduction of the 'buffer zone' volume with a higher shape complementarity between cleft and ligands.

It should be stressed at this point that the comparison of physicochemical properties between ligand sets is in general troublesome. Suppose an ATP and a phosphate molecule bind at overlapping surface areas on a protein, with the phosphate's binding site being a subset of the ATP's binding site. As all previous pictures in this chapter have depicted, the physicochemical properties of a protein can largely vary within short distance. In our approach, which describes the van der Waals surface of a molecule as a single valued function  $f(\theta, \phi)$ , this can lead to different property scores at equivalent sample points  $(\theta, \phi)$  for both molecules. Furthermore, the assessment of the contribution of each molecular descriptor to the molecular recognition of a ligand set, requires the independent comparison of each descriptor. The physicochemical properties must be therefore detached from size and shape (disregarding the implicit incorporation of the shape when measuring the physicochemical properties on a ligand molecule). This however can lead to biologically meaningless results in

particular for ligand molecules that cannot sterically fit to each others binding site but experience similar physicochemical properties. I was aware of these problems, however performed the subsequent analysis to demonstrate the power and the diversity of spherical harmonic functions but primarily in the hope that some general tendencies about molecular recognition could be inferred.



**Figure 4.16: Spherical harmonics reconstruction of electrostatic potential distribution.**

Reconstruction of the electrostatic potential distribution that AMP experiences in the bacterial peptide synthetase (PQS-Id: 1amu\_1) with spherical harmonic functions. The potentials are measured at the atom centres and mapped on the dots of the van der Waals surface of the AMP (centre of picture). The potential reconstruction is shown on the shape reconstruction of AMP ( $l_{\max} = 14$ ) with different degrees of termination in the spherical harmonics series expansion, going clockwise from low quality with  $l_{\max} = 0$  to high quality with  $l_{\max} = 14$ .

The spherical harmonic expansion was employed to test whether the ligand sets in Data set I experience discriminative pattern of physicochemical properties within protein binding sites. The 'Area Under the receiver-operator-characteristics Curve' (AUC) is a convenient way to measure the performance of such a classification attempt (see section 3.2.2). Table 4.2 lists the AUC values for the physicochemical properties, for the ligand and binding pocket shapes as well as for their linear combination. Of all the properties analysed, the geometric properties stand out with an average AUC value for the ligand molecules of 0.93 and for the Convex Cleft Model of 0.87. Note, that the small difference in the AUC values here and in Table 3.2 are due to the hydrogen atoms that were added to the molecules in this chapter. The AUC values for the geometric properties are more than 0.11 units higher than for HES (AUC = 0.76) and ESP (AUC = 0.74). Looking more carefully at the AUC values of individual ligand sets, it is evident that FAD, HEM and PO4 tend to have unique physicochemical features with AUC values above 0.75 for both ESP and HES, which allows them to be fairly easily discriminated from the remaining ligands in the data set. In contrast, AMP and FMN molecules have the least discriminative properties with the lowest AUC values for all physicochemical properties. It turns out that the pattern of ESP and HES on both AMP and FMN molecules often resembles ATP and PO4 binding sites.

Table 4.2 holds in addition to the AUC values of the single descriptors, the AUC values of the linear combinations of the same descriptors. For each binding site, comparison the linear combination with weight factors set to one was calculated over the ESP and HES coefficient distances. The resulting new scores were reordered and the AUC value was recalculated.

The linear combination of the physicochemical properties without shape information had minor effects on the AUC values, most often scoring in-between the AUC values of the single physicochemical descriptors. However, for the FAD ligand set an increase of about 0.1 AUC units was detected, which was caused mainly by the complementary variation of ESP and HES in the FAD ligand set. According to Figure 4.15, ESP varies highly at the adenosine and flavin moiety, which however is highly conserved for HES. Combining both properties into a

single coefficient distance merges the conserved moieties of ESP and HES and leads to an average AUC of 0.92 (see Table 4.2). Shape plays a minor role in the ligand set of FAD with an AUC of 0.74 for the convex cleft model. Similarly, for the ATP ligand set the classification based on shape improved when physicochemical properties were included. In contrast, for the AMP, ATP, FMN, NAD, PO4 and Steroid ligand sets the linear combination of the physicochemical properties increased by at least 0.1 AUC units when shape information was added. The superior classification performance of linearly combined properties gives evidence that molecular recognition is in general a mutual cooperation of various geometrical and physicochemical factors and not induced by a single property, however with a proneness towards shape complementarity. This conclusion is in line with previous work (Jiang, et al., 2002).

**Table 4.2: AUCs for geometrical and physicochemical properties.**

Descriptor	All	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO4	Steroid
Ligand shape	0.93	0.99	0.90	0.80	0.94	1.00	1.00	0.82	1.00	1.00
Cleft shape	0.87	0.90	0.80	0.74	0.85	0.93	0.97	0.78	0.95	0.93
Hydrophobicity	0.76	0.54	0.79	0.81	0.65	0.67	0.94	0.63	0.79	0.95
Electrostatic	0.74	0.59	0.71	0.76	0.60	0.94	0.78	0.68	0.85	0.51
Electro + Hydrophobic	0.80	0.56	0.80	0.90	0.62	0.95	0.91	0.71	0.87	0.80
Cleft Shape + Electro + Hydrophobic	0.92	0.84	0.92	0.88	0.85	0.98	0.98	0.83	0.98	0.97

Area under Receiver operator characteristic Curve (AUC) for geometrical and physicochemical properties and their linear combination for all ligand sets in Data set I. The lowest and highest AUC for each property is coloured purple and green respectively. Red coloured numbers highlight the highest AUC among the physicochemical properties of each ligand set.

## 4.4 Discussion

The starting point for this work was to explore the validity of the assumption that binding sites, which bind the same ligand, achieve their selectivity by providing a similar physicochemical environment. To analyse this ligand-binding pocket relationship, semi-empirical calculations were performed to compute and then compare the electrostatic and hydrophobic environments in 100 protein binding sites. The binding pockets were analysed in groups with each group member binding the same ligand. The results reveal large variations in protein environments experienced by each ligand. These large quantitative differences indicate the absence of perfect physicochemical complementarity between binding site and ligand. This work gives evidence that - similar to the concept of convergent evolution of gene and protein function - nature has evolved multiple binding solutions for the same ligand. For some proteins it may not matter how the ligand binds to its receptor, but merely that it binds and so a diverse range of strategies of binding the same ligand have evolved. However, the imperfect complementarity raises the question as to how proteins utilise physicochemical forces to distinguish between different small molecules in their environment.

For promiscuous enzymes, such discrimination is not required. These enzymes are able to perform their function on various different ligand molecules (Copley, 2003; Khersonsky, *et al.*, 2006) that cannot all display physicochemical complementarity to the binding site. In fact, enzymes with new functions might have evolved from ancestral proteins that were promiscuously binding a second substrate molecule. In the course of time, mutations at the active site might have caused the enzyme to lose its affinity towards its original substrate and become selective only for the second substrate (Khersonsky, *et al.*, 2006). Promiscuity can also be advantageous for the survival of a species under varying environmental conditions, in particular for prokaryotes. It could give the organism the ability to maintain its function under contrary conditions, thereby guaranteeing the survival of the organism (James and Tawfik, 2003). The variation that was observed in this thesis in the geometrical and physicochemical properties of protein binding sites supports the role of promiscuity in protein



evolution. The partial complementarity between binding site and ligand would allow proteins to bind their original substrate molecule and maintain their original function. However, at the same time they could explore other substrate molecules with similar binding affinities that would allow them to eventually alternate their biochemical function within a cell.

But for non-promiscuous enzymes, the contrasting local concentrations of ligands and proteins in cell compartments will have a large effect on the collision frequency and therefore on the overall binding of both binding partners. Albe *et al.* have estimated that substrate molecules have double the concentration as their protein counterparts (Albe, et al., 1990) and Kurland and colleagues showed that eukaryotic proteins have a variety of distributions across cell compartments (Kurland, et al., 2006). Unfortunately data on the local protein–metabolite concentration within cells are not available for the majority of proteins for which we have 3D structural models in the PDB. However, such data is essential for the analysis and simulation of cellular processes.

Regardless of the simplicity of the approach in this chapter, it was possible to observe some general principles of molecular recognition, including the contribution of the hydrophobic interactions to molecular recognition (Davis and Teague, 1999; Gilson and Zhou, 2007; Gruber, *et al.*, 2007). In this work, the hydrophobicity was observed to vary relatively little among each set of ligands (see Table 4.1 and section 4.3.4). In contrast, the electrostatic potential varies widely and is highly influenced by neighbouring chemical compounds. The importance of entropic energy contributions is also supported by the ‘buffer zone’, i.e. the free space between a binding pocket and its bound ligand (see Figure 3.10), which can stabilize entropically the protein-ligand complex by allowing the ligand and the binding site residues to retain some of their vibrational motion and flexibility (Boehm and Klebe, 1996). It should be pointed out that the correlation of the computed properties with experimental observations (e.g. binding constants and redox potentials) is modest at best. In some cases, violations against the principle of molecular complementarity turned out to be a prerequisite for the protein’s biochemical function (see section 4.3.2.3). Herein, no considerations were paid to

the binding specificity and discrimination power of proteins towards their cognate ligands, which adds a further level of complexity to the analysis of the protein ligand interaction *in silico*. In particular most allosteric and induced-fit enzymes lack a correlation between binding affinity and specificity due to the smaller magnitude of  $\Delta\Delta G$  as compared to the free binding energy  $\Delta G$  (Schneider, 2008). Consequently, a protein can demonstrate specificity towards its ligand despite having a low complementarity and thus binding affinity. Furthermore, recent results on *in vivo* proteins have shown that partially or fully intrinsically disordered proteins are common, in particular among cell signalling proteins and transcription factors, with each undergoing a transition from a disordered to an ordered structure upon ligand binding (James and Tawfik, 2003). In this respect, the “New View” model for protein folding was briefly introduced in section 2.3.2.4. According to the “New View” model a disordered protein exists in different isomeric states. Recent experiments on antibody-antigen complexes showed that each of the isomeric states could bind different antigens with high specificity but low affinity (James, *et al.*, 2003). In particular it was found that the shape complementarity in antibody-antigen interfaces is lower than in other protein-protein interfaces (Jones and Thornton, 1996). The lack in shape complementarity permits an antibody to recognise spontaneously new antigens that have not been seen so far in the antibody’s evolutionary history (Uversky, *et al.*, 2005), which defines the strength of the immune system against pathogenic intruders.

The observations from this chapter confirm results found by some earlier studies that were made on limited data sets (see references in the introduction to this chapter), yet the myth of exact complementarity remains. The observations have consequences for a number of applications, particularly those related to function prediction and virtual screening. Function prediction methods from structure are frequently based on detecting similarities between annotated and functionally unknown binding sites. The very basis of these approaches is, however, challenged if complementarity is not a given. Such methods must cope with this variation and include a probabilistic term in their similarity search that accounts for the observations in this chapter that the same ligand can bind in one binding site via electrostatic interactions, in another via entropic contributions or in a third via purely van der Waals

interactions. A workaround to the probabilistic description could be the utilization of '3D consensus binding profiles'. Such profiles would represent the physicochemical environment that a ligand on average experiences in various proteins. A similarity search would then involve the comparison of the physicochemical properties of a potential ATP binding site against a set of 3D consensus binding profiles. Not just would this decrease the screening time for a whole database, but it also might increase the prediction accuracy, as the average properties tend to have higher complementarity to the ligand properties than single binding-site/ligand-complexes.

Despite well-developed theory and extensive progress in molecular simulations computational approaches to calculate both enthalpic and entropic energies are still limited in accuracy (Gilson and Zhou, 2007). The accurate computation of these terms is, however, necessary to understand the fine balance of binding forces underlying protein-ligand interactions and to make reliable predictions of binding energies. Scoring functions implemented in docking applications typically employ a rather simplistic model of molecular recognition to allow the screening of a myriad binding poses (Jain, 2006). As a result, the average accuracy of docking applications lies within 1.5 – 2 Å root mean square deviation in about 70-80% of cases. Most of these success are cases in which ligand and protein are relatively rigid (Sousa, et al., 2006). Docking calculations performed by Angelo D. Favia with a MM-GBSA model (Lyne, et al., 2006) on Data set I showed unfavourable positive energies of up to +170 kcal/mol for 20 protein ligand complexes (see Table 4.3). The results in this chapter show that computational shortcuts based on complementarity can be dangerous and that more sophisticated models would be required for accurate predictions of binding energies and binding poses that take into account all factors and forces currently known to contribute to molecular binding. In particular desolvation energies (Koehl, 2006) at the protein binding site and entropy gain/loss of ligand and binding site molecules (Gilson and Zhou, 2007) will need addressing. This complexity will provide a challenge for computational biology, be it for the accurate calculation of binding free energies or the derivation of more empirical approaches.

Table 4.3: Binding free energies calculated for Data set I

PDB id	Ligand Code	MM-GBSA (kcal/mol)	PDB id	Ligand Code	MM-GBSA (kcal/mol)	PDB id	Ligand Code	MM-GBSA (kcal/mol)
1qf5	PO4	-525.15	1jq5	NAD	-86.49	1d7c	HEM	-37.97
1tco	PO4	-365.55	1hsk	FAD	-85.46	1jp4	AMP	-36.66
1e9g	PO4	-292.09	1tb7	AMP	-84.66	1e3r	AND	-36.41
1ew2	PO4	-259.06	2a5f	NAD	-82.88	1fds	EST	-36.03
1b8a	ATP	-242.12	1dv2	ATP	-82.08	1j99	AND	-35.95
1gn8	ATP	-216.45	1bdg	GLC	-79.72	1cq1	GLC	-34.96
1ho5_1	PO4	-212.91	1p4c	FMN	-78.19	1nf5_2	GLC	-32.19
1h6l	PO4	-206.42	1p4m	FMN	-76.40	1hex	NAD	-25.34
1lby	PO4	-191.97	1tox_1	NAD	-75.06	1esq	ATP	-23.04
1rdq	ATP	-186.30	2gbp	GLC	-74.81	1iqc_1	HEM	-6.30
1e8x	ATP	-165.08	1mvl	FMN	-71.62	1jr8	FAD	-5.62
1a0i	ATP	-162.21	1ib0	NAD	-69.71	1eqg	HEM	-3.78
1qax	NAD	-160.58	2npx	NAD	-69.46	12as	AMP	-3.51
1dy3	ATP	-148.79	1cqj_1	PO4	-64.75	1qpa	HEM	0.24
1dnl	FMN	-145.84	1gwe	HEM	-61.77	1c0a	AMP	1.06
1mi3_1	NAD	-143.30	1og3	NAD	-60.68	1a6q	PO4	2.74
3grs	FAD	-138.67	1po5	HEM	-59.94	1b8o	PO4	4.16
1qla	HEM	-122.34	1o9t	ATP	-57.48	1sox	HEM	4.70
1e2q	ATP	-118.03	1cqx	FAD	-55.69	1qb8	AMP	5.26
1brw	PO4	-117.41	1lhu	EST	-51.42	8gpb	AMP	11.04
1t2d	NAD	-115.52	1k1w	GLC	-51.03	1qhu	HEM	14.78
1k87	FAD	-115.20	1s7g	NAD	-49.62	1dk0	HEM	18.24
1ayl	ATP	-114.46	1h69_1	FAD	-47.69	1pp9	HEM	20.25
1kvk	ATP	-109.16	1qkt	EST	-46.99	1naz	HEM	20.62
1rlz	NAD	-108.89	1kht	AMP	-46.48	1l5w	PO4	30.68
1tid	ATP	-104.91	1mew	NAD	-46.25	1d0c	HEM	39.22
1amu_1	AMP	-104.18	1f5v	FMN	-46.17	1dak	PO4	51.63
1l7m_1	PO4	-102.47	1ew0	HEM	-45.53	1euc	PO4	59.25
1e8g	FAD	-101.47	1ja1_1	FMN	-45.34	1ftb	PO4	70.85
1evi	FAD	-101.11	1lyv	PO4	-44.02	1np4	HEM	72.38
1pox	FAD	-98.70	1o04_1	NAD	-42.53	1ejd	PO4	79.75
2cpo	HEM	-90.76	1d1q	PO4	-38.66	1ej2	NAD	107.63
1jqj	FAD	-90.29	1a49_1	ATP	-38.12	1gyp	PO4	171.24
						1ct9_1	AMP	*

Binding site ligand complexes are sorted according to their calculated binding free energy from low (top left) to high (bottom right). \*The binding free energy calculation for 1ct9\_1 could not be completed due to missing parameters for the uranium NCC metal ions.

## 4.5 Conclusion

This chapter expands the work in Chapter 3 on the geometrical variation in protein binding pockets and ligands by analysing the same binding-pocket / ligand-complexes with respect to their physicochemical properties. Here, the analysis continued on the same binding sites and showed that neighbouring chemical components must be included in the calculation of physicochemical properties. Metal ions in particular had the ability to invert the sign of the experienced electrostatic potentials on a ligand and to induce attractive forces where originally repulsive forces were detected (Figure 4.2). Similar sign inversions of electrostatic potentials on ligands molecules were observed when the solvent screening effect was reduced by assigning the ligands a dielectric constant of  $\epsilon_L = 4$  (Figure 4.3). Furthermore, it was demonstrated that the physicochemical properties ligands experience when bound to different binding pockets vary significantly (Figure 4.4, Figure 4.7, Figure 4.9, Figure 4.10). This high variation reflects large energy fluctuations that are sometimes functionally necessary, including changes in the sign of the potentials for corresponding atoms in a ligand set (Figure 4.15). To overcome repulsive electrostatic interactions, proteins were observed to utilise attractive aromatic interactions to their ligands (Figure 4.11, Figure 4.12). Complementarity was observed to some extent only for averaged properties in the ligand set (Figure 4.13 and Figure 4.14).

## Chapter 5

# Automated Ligand Recognition in Electron Density Maps

## 5.1 Introduction

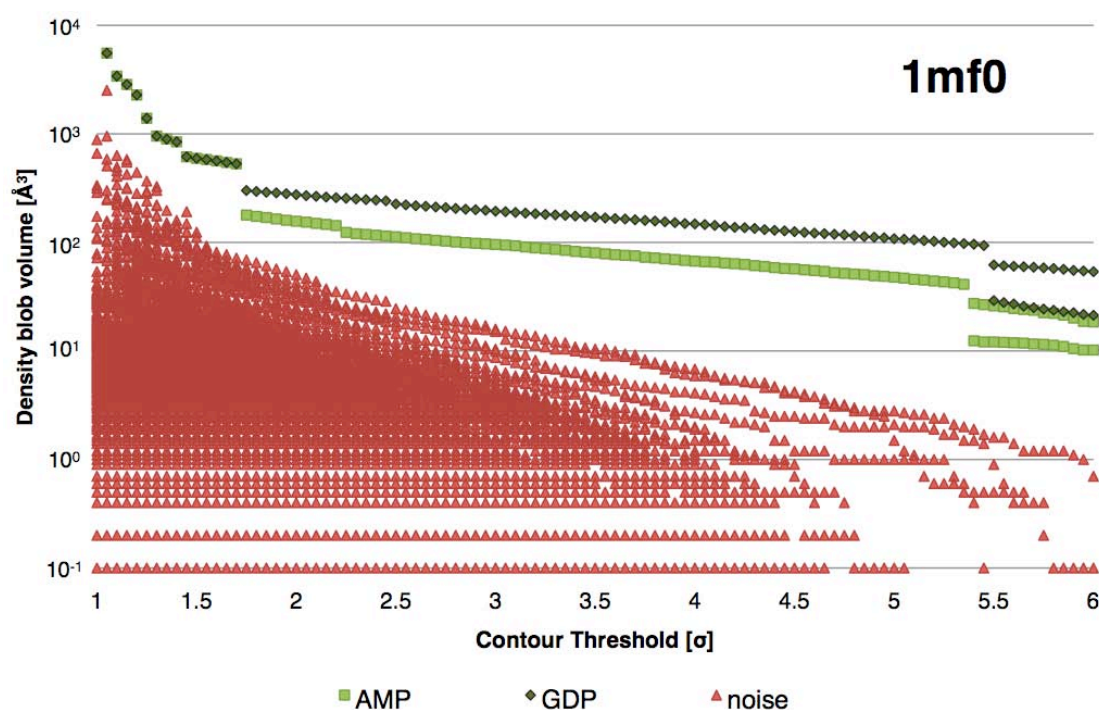
Since the beginning of automatic protein model building (see section 2.2.1.3), almost all emphasis was placed on the construction of the protein model alone. Less importance was attributed to the automatic fitting of ligand and polynucleotide molecules into electron density maps. Two of the very first programs to perform automated ligand fitting were ESSENS (Kleywegt and Jones, 1997) and X-LIGAND (Oldfield, 2001). ESSENS used a minimum function to score and find the best fit of a template ligand to a given electron density. However, this matching was computationally intensive, requiring systematic rotations of the ligand around each density grid point. With X-LIGAND the first automated flexible ligand fitting procedure was introduced. X-LIGAND first selected closed regions of density also called *density blob* at different contour thresholds, sigma ( $\sigma$ ), in a difference electron density map. Such a map represents the residual electron density of a full density map after the density of the protein model is removed. Mathematically a difference electron density map can be computed by subtracting calculated structure factors  $F_C$  of the protein model from the experimental observed structure factors  $F_O$  of the experiment ( $F_O - F_C$ ). Having obtained all density blobs in the difference electron density map X-ligand selects those blobs that have a comparable volume to the ligand molecule. Next numerous conformers of the ligand are generated on the fly and matched to the density blobs by an alignment of their principal moments of inertia. Several other methods have been introduced since then. The ligand fitting

procedure in RESOLVE (Terwilliger, 2003) follows the principle of fragment-based docking, where the ligand is first divided into rigid fragments and later step-wise constructed inside the density blobs by connecting the fragments following a density fit scoring function. BLOB (Diller, et al., 1999) uses a Monte Carlo sampling technique to scan the electron density grid for an appropriate location for the ligand molecule. Aishima and coworkers have developed a technique (Aishima, et al., 2005) that simplifies the isosurface of a density blob with medial axes, following the idea of skeletonisation of the main chain density (see section 2.2.1.3). The set of medial axes is matched to the connectivity graph of the ligand using a graph-matching algorithm. Lately, AutoSolve (Mooij, et al., 2006) and work by Wlodek and coworkers (Wlodek, et al., 2006) were introduced for the automatic ligand fitting. AutoSolve uses a genetic algorithm to optimize the location and conformation of the ligand within difference maps. The scoring function for the genetic algorithm includes in addition to a density overlap score, protein-ligand interaction terms and ligand internal energy terms to distinguish similar shaped ligand molecules and exclude energetically unfavourable ligand conformations. Wlodek *et al.* took the idea further and minimized the internal energy of an ensemble of ligand conformers before aligning ligand and density blobs in terms of their moment of inertia and matching their shapes with Gaussian volume functions.

Since version 6.1, ARP/wARP also allows the automated fitting of ligand molecules into difference electron density maps (Evrard, et al., 2007; Zwart, et al., 2004). The algorithm of the ligand-fitting module consists mainly of five steps:

1. Create a difference electron density map
2. Extract the largest blob of electron density from the difference map
3. Represent density blob with sparse set of grid points
4. Construct the ligand by assigning atom labels to the sparse-grid nodes via label-swapping
5. Refine electron density with fitted ligand

The most recent ARP/wARP version 7.0 features new functionality in the ligand fitting module (Langer, et al., 2008); important ones are filter procedures to locate the location of ligand binding sites within the second step of the general algorithm. These filter procedures aim at removing insignificant densities in the difference map that are unlikely to have been produced by the ligand of interest. Difference electron density maps always contain in addition to the density of ligand molecules a number of features that originate from experimental noise, protein model uncertainties, solvent molecules and other small and partially ordered ligands such as crystallisation agents. These densities can be recognized and removed with a *fragmentation tree filter*, which makes use of the common characteristics of non-ligand density blobs, namely their smaller volume and their rapid decay into separate smaller blobs with increasing contour threshold.



**Figure 5.1: Fragmentation tree filtering of electron density blobs**

The filtering was performed on the difference density map of Adenylosuccinate synthetase (PDB-Id: 1mf0). Density blobs of small molecules (here AMP and GDP) distinguish themselves from noise (i.e. experimental noise, model uncertainty, ions, solvent) by a high volume prolonging over a large range of contour threshold ( $\sigma$ ) and a fragmentation at only higher threshold levels.



The fragmentation tree plots the density blob volume as a function of the contour level. The highest information content is found usually between  $\sigma = 1.0$  and  $\sigma = 6.0$ . For the filtering, only those densities are selected that have a certain size and fragment only at later contour thresholds (see Figure 5.1). The fragmentation tree filter is able to significantly reduce the number of density blobs in a difference map.

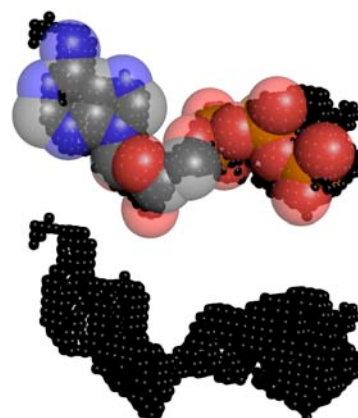
The second filter of ARP/wARP's ligand binding site locator, searches for gross-shape similarities between candidate density blobs and a ligand molecule using the following seven intrinsic geometric features:

1. Surface to volume ratio
2. Bounding box limits
3. Principal moments of inertia
4. Rotation match score
5. Eigenvalues from covariance matrix
6. Distance histogram
7. Geodesic distance histogram

Only density blobs that are geometrically most similar to the ligand of interest are passed further down to the third step in the overall algorithm. Although the seven geometric features are appropriate for characterizing the shape of a density blob or a molecule, they are not a genuine shape descriptor like the expansion coefficient of spherical harmonic functions introduced in Chapter 3 and have therefore some drawbacks. Firstly, the linear combination of all features is non-trivial. Currently, the scoring function for the geometric features consists of a sum of 36 linear combinations of the seven features. The weights for each linear combination were trained with a genetic algorithm and may require recalculation if applied to a different data set. Secondly, the features are to a certain degree correlated and not independent from each other, making it difficult to evaluate false-positive or false-negative predictions. And finally, some of the features are relatively time consuming to calculate.

In collaboration with Victor Lamzin and Gerrit Langer from EMBL Hamburg, the spherical harmonics shape descriptor of CleftXplorer was suggested as a supplement to the geometric features. In the previous two chapters, the functionality of the spherical harmonic shape descriptor was demonstrated. Thus, the question about the application of the shape descriptor to blobs in difference electron density maps was not whether the descriptor was able to capture the shape of density blobs but rather whether the density blobs would show sufficient resemblance to the ligand molecules to allow their recognition. After all, the density blob of ligand molecules in electron density maps is smaller in size and in its shape details distinct from that of the ligand molecule due to noise and partial disorder (see Figure 5.2). Spherical harmonics in ligand density recognition have the potential to exceed all methods mentioned above in their accuracy/speed performance. All of the methods above are either accurate in their shape matching but computationally exhausting or vice versa. With this approach, various experimental protocols involving the comparison of numerous entities can be realised with minimal computational cost. For example, given an electron density of an unknown ligand, a large ensemble of ligand conformers can be matched to the density. Similarly, given a density map with unidentified location of a known ligand, the whole density map can be screened for density blobs with a similar shape to the ligand, and this all in little computation time.

As a proof of principle, this chapter will show tests of the spherical harmonic shape descriptor as a filter of density blobs for the automated ligand-fitting module in ARP/wARP. An overview of its performance will be shown together with a comparison to the performance of the geometric features. The chapter will end with a discussion about successful and failed predictions and



**Figure 5.2: Electron density of an ATP.**

Electron density blob of ATP in hydrolase 1ii0 at a contour threshold  $\sigma = 1.70$ . Note the smaller volume of the black coloured density blob as compared to the varicoloured ligand and their distinct shape details. The density blob is shown with and without the ligand superimposed.

conclusions about the range of applicability of the spherical harmonic shape descriptor in automated ligand recognition.

## 5.2 Methods

The spherical harmonic shape descriptor used to capture the shape of density blobs was extracted from *CleftXplorer*, which is described in the Methods section of Chapter 3. In order to avoid repetition only a brief overview of the algorithm will be given.

### 5.2.1 Algorithm summary

All candidate density blobs were processed by a subroutine of *CleftXplorer*. In order to make the density blobs readable by the subroutine, all blobs were first converted into clusters of density grid points (see Figure 5.2 and Figure 5.3). For the remaining part of this chapter, all density blobs will be referred to as density clusters, reflecting their representation by a cluster of grid points. After having read the coordinates of the density cluster, the subroutine translated the density cluster with its centre of mass to the Cartesian coordinate origin and rotated the cluster such that its principal moments of inertia coincide with the three Cartesian axes  $x, y, z$  in the order of their eigenvalue magnitude. Having transformed the density cluster, the spherical 21-design integration layout was mapped on the outmost surface shell of the cluster. The resulting single valued radial function  $f(\theta, \phi)$  was expanded with spherical harmonics and the derived expansion coefficients were stored. The same approach was applied to the molecular surface of ligand molecules. Eventually, both sets of shapes were compared by comparing the ligand and density coefficient vectors using a Euclidean distance metric.

## 5.2.2 Performance measure

The performance of the spherical harmonic shape descriptor's ability to predict the correct density cluster for a ligand within a difference map was tested with the same statistics as in Chapter 3, the *Area Under the receiver operator characteristics Curve* (AUC). In an AUC calculation the fraction of true hits is plotted against the fraction of false hits when an ordered list of scores is ranked from best to the worst, and the area under the resulting curve is computed. In this work, electron density clusters at various contour thresholds ( $\sigma$ ) originating from the ligand in question were referred to as true hits (true/correct density clusters) and all others were labelled as false hits (false/incorrect density clusters). AUC values of 1.0 indicate perfect classification (i.e. all true positives occupy the highest ranks without exception) and 0.5 indicate random classification (with equal number of true and false positives being retrieved as the list is processed).

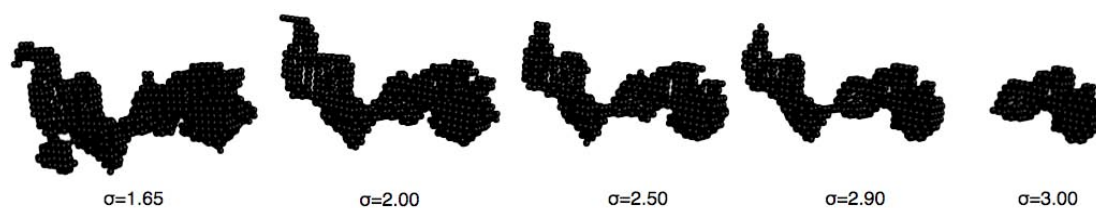
## 5.3 Data set

To test the performance of spherical harmonics for the automated ligand recognition, a data set in consultation with Gerrit Langer was compiled. For the data set all 29,742 diffraction data sets from the Electron Density Server (Kleywegt, et al., 2004) (as on the 1st April 2008) were downloaded. The associated protein structures from the Protein Data Bank (PDB) (Berman, et al., 2000) were refined without any HETATM entries, i.e. without ligand and solvent molecules, against the diffraction data. This step caused a worsening in the correlation between the experimentally observed and modelled structure factors, but conformed to real experiments where the atom coordinates of the ligand molecules are unknown and not available for the refinement routine of the protein model (see section 2.2.1.2). Subsequently, difference electron density maps were calculated with *refmac5* (Winn, et al., 2003) and filtered with the fragmentation tree approach, giving a list of candidate

density clusters. The list of candidate density clusters was redundant as it held the same density at different contour thresholds (see Figure 5.3). Protein structure models were further checked to fulfil the following criteria:

1. Resolution of density map: 2.0 Å - 2.5 Å
2. Map correlation: > 75%
3. Number of non-hydrogen atoms in Ligand: 20 - 40
4. Occupancy of Ligand: 100%
5. Ratio of true hits in the list of density cluster: 20% to 80%.
6. Number of total density clusters examined: < 400

Of the initial 29,742 density maps from the Electron Density Server, only 471 maps with 548 ligand molecules passed all filters and satisfied the above conditions. The large majority of density maps did not satisfy the above best-case scenario. A subset of 12 ligand molecules was randomly chosen from the large data set to form a small Data set III (see Appendix B, Table B.1) for the evaluation of two parameters in the expansion coefficient calculation, leaving 536 entries in the large Data set III (see Appendix B, Table B.2). Both parameters addressed the two problems stated above, namely the size difference between ligand molecules and the differences in their shapes. Note, for the small Data set II, the ligand conformations as found in the X-ray structure were used to optimize the parameters for the ideal case of having the ligand in its correct conformation for the fitting procedure. For the large Data set III, ideal coordinates for the ligand, as found in the HIC-Up database (Kleywegt, 2007), were used to screen the density clusters, resembling the real case scenario of not knowing the true conformation of the ligand in the crystal structure prior to the fitting process.



**Figure 5.3: Electron density at various contour thresholds.**

Depicted as black coloured dots are the electron densities of an ATP in the bacterial ATPase 1ii0 at different contour thresholds ( $\sigma$ ). Increasing  $\sigma$  causes the density to shrink until  $\sigma = 3$ , at which point the density loses its integrity.

## 5.4 Results

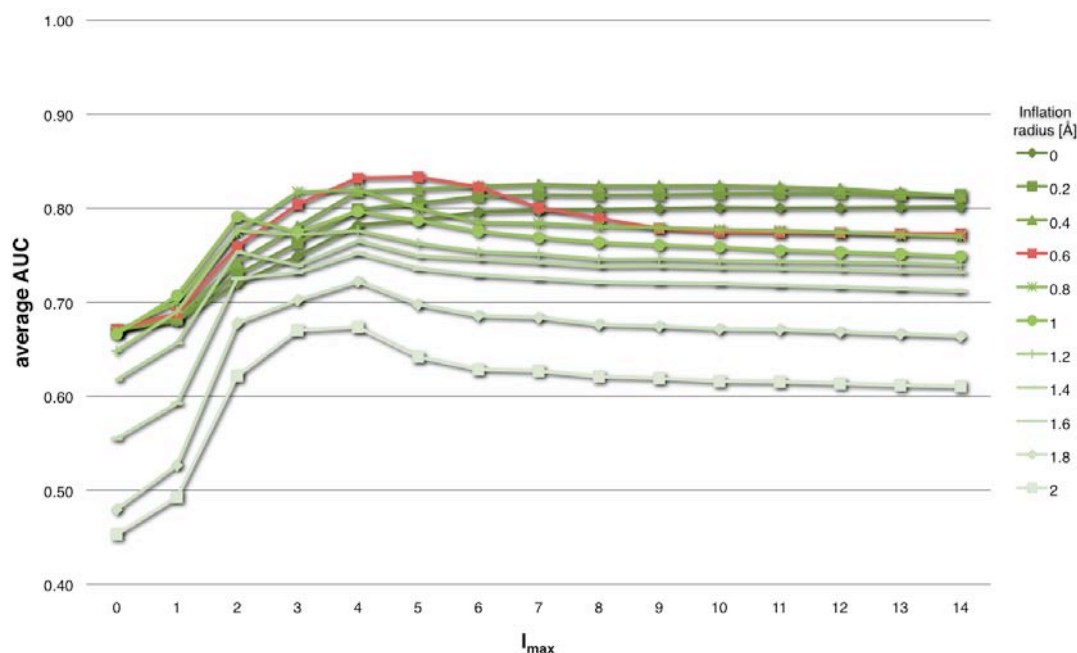
The following scenario was tested to assess the performance of the spherical harmonic shape descriptor for the automated ligand recognition. Given a difference electron density map of a protein structure with a known bound ligand present, can the spherical harmonic shape descriptor identify all correct density clusters of the ligand from the mass of clusters in the difference map by ranking them highest?

### 5.4.1 Parameter assessment

Data set II with 12 test cases showed that the electron density clusters of small molecules were often smaller in volume and distinct in their shape details when compared to the volumetric representation of the molecule itself (see Figure 5.2). This geometrical difference is caused primarily by the preference of the filtering program to select higher contour thresholds that lead to smaller sized density clusters. Nevertheless, both characteristics had to be considered in the expansion coefficient calculation before any test on the performance of the shape descriptor could be conducted. Two parameters in the expansion coefficient calculation were therefore tested:

1. The step-wise increase of the zeroth order coefficient by a factor of 0.708 to a maximum of 7.08. According to Figure 3.5 and the observed 3.54 : 1 ratio between surface RMSD and coefficient distance, this simulates a gradual increase in the size of the density cluster by inflation steps from 0.2 Å to a maximum of 2.0 Å.
2. Variation of the degree of termination  $l_{\max}$  from 0 to 14 in the expansion of the radial function with spherical harmonics. Lower termination degrees should capture gross shape, but not detailed features.

Figure 5.4 shows the overall performance of the shape descriptor at different inflation radii and degrees of expansion termini. The best classification with an average AUC value of 0.833 was achieved with an inflation radius of 0.6 Å and  $l_{\max} = 5$ , followed by 0.6 Å and  $l_{\max} = 4$  with an AUC = 0.832. However due to a lower standard deviation while having a comparable AUC



**Figure 5.4: Spherical harmonics performance at changing parameters.**

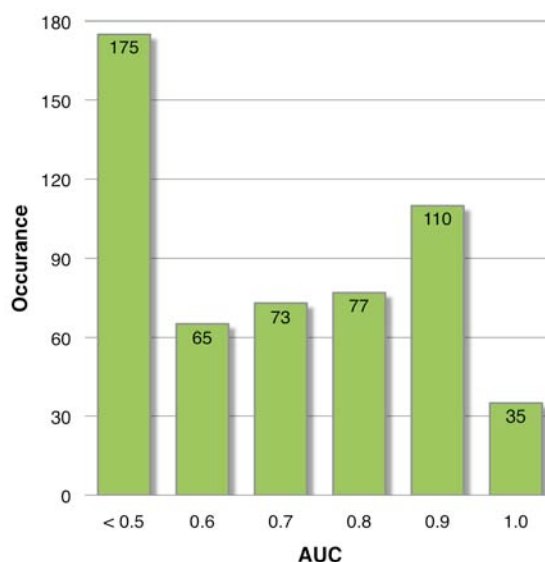
Average classification performance of the spherical harmonics shape descriptor for various coefficient expansion termini  $l_{\max}$  and inflation radii. The Area Under the Curve (AUC) were calculated for each parameter pair and averaged over Data set III. For the recognition of density clusters the parameters were set to  $l_{\max} = 6$  and inflation radius = 0.6 (red curve).

value of 0.822 (data not shown) it was decided to set the inflation radius to 0.6 Å and the expansion terminus to  $l_{\max} = 6$ .

## 5.4.2 Ligand recognition with spherical harmonics

Applying the parameters determined above to Data set III, the spherical harmonic shape descriptor achieves an average AUC value of 0.69 with an average standard deviation of 0.26 units. Figure 5.5 shows the distribution of AUC values obtained for all 536 ligand molecules in Data set III. Despite the absence of a conformation generator in CleftXplorer the histogram shows that the spherical harmonics shape descriptor performs reasonably well.

In 35 cases, the spherical harmonic shape descriptor ranked all true density clusters at top rank positions achieving an AUC of 1.0. Most of these cases were density clusters of either rigid molecules or ligands that had a distinct size and shape difference to the remaining ligands in the X-ray structure. In 145 cases an AUC value higher than 0.9 was achieved. However for about a third of Data set III the shape descriptor performed no better than random with AUC smaller than 0.5. For four density maps, the AUC value was even 0.0, implying that all true density clusters were ranked at end of the ranking list. A close look at those cases revealed extreme situations in the experimental observed diffraction data that shall be



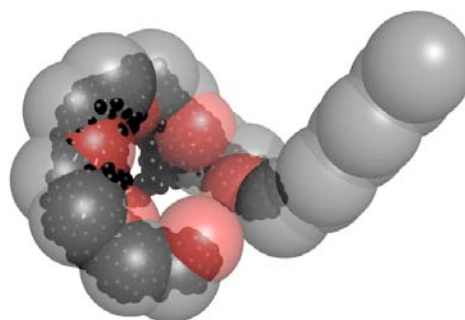
**Figure 5.5: Spherical harmonics performance on Data set III.**

Distribution of AUC values in the large data set based on the comparison of ligand and density clusters with the spherical harmonics shape descriptor.



discussed next.

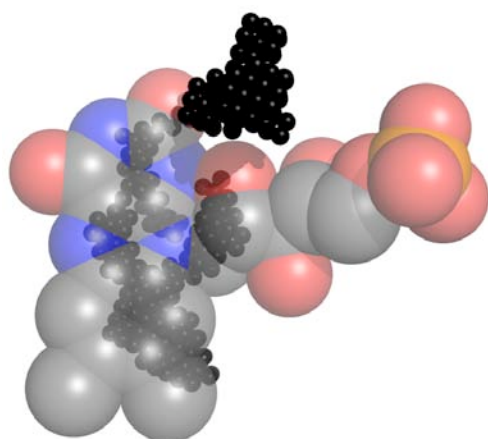
The case of the dehydrogenase 1tv5 and its ligand N8E (see Figure 5.6) shows that flexible ligand molecules bound partially to the protein generate electron density only for those parts, which are rigidly bound to the protein. Mobile moieties, in particular those protruding into the solvent do not appear in the density map, preventing global shape descriptors the recognition of the complete ligand molecule in the partial density.



**Figure 5.6: Limitations of spherical harmonics: ligand flexibility.**

The ligand N8E (grey and red coloured spheres) in the dehydrogenase (PDB Id: 1tv5) only produces density (black coloured dots) with its ring formed portion. The linear extension on the right hand side is flexible, highly solvated and unbound.

Another problem arose due to poor phases of structure factors (see section 2.2.1.2) in the oxidase 1y30 and its ligand FMN. The poor phases caused the density to lose entirely the



**Figure 5.7: Limitations of spherical harmonics: poor phases.**

Poor phases lead to hardly interpretable difference map of the oxidase 1y30 and its ligand FMN hinder a good match between the ligand (varicoloured) and its density cluster (black coloured dots).

shape signature of the ligand (see Figure 5.7). In such situations, a shape descriptor is unlikely to successfully perform any sensible prediction. The dodecaethylene glycol 12P in the aminooxidase 2b9x shows another limitation of the spherical harmonics shape descriptor, namely its restriction to describe only star-like shaped surfaces (see section 3.4.3.3). The glycol conformation as found in the aminooxidase is clearly non-star shaped, that CleftXplorer can just approximate by the outermost shell of the surface. The error that this approximation produces causes the density

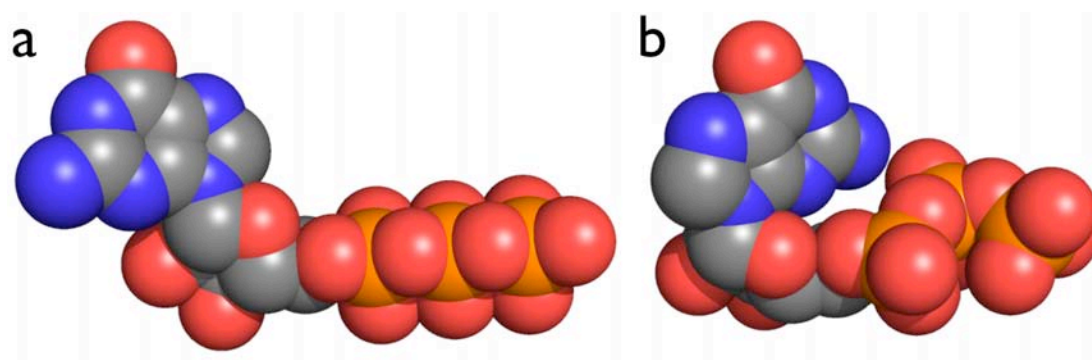


**Figure 5.8: Limitations of spherical harmonics: non star-like shapes.**

Dodecaethylene glycol (a) and its density clusters (b) as found in the aminooxidase (PDB-Id: 2b9x). The non-star-like conformation of the glycol molecule produces expansion coefficients (shown as reconstructed mesh shapes calculated at  $l_{\max} = 6$ ) that are similar to those of the FAD density cluster (c). Density clusters are depicted as black coloured dots in (b) and (c).

cluster to resemble the density of a neighbouring FAD molecule (see Figure 5.8), that eventually due to its larger size matches the large glycol molecule with lower coefficient distance.

The last case that exhibits an AUC = 0.0 was found for GTP's density cluster in the protein structure of polyhedrin 2oh7. The poor performance of the shape descriptor in this particular example was caused by the conformational difference between the ideal coordinates that were used to scan the density clusters and the conformation as found in the X-ray structure (see Figure 5.9). The already technically difficult situation was made worse by poor phases (see above example of FMN in 1y30). Altogether, the elongated conformation of the ideal coordinates of GTP was confused with the density cluster of a neighbouring ATP molecule that existed in a similar conformation.

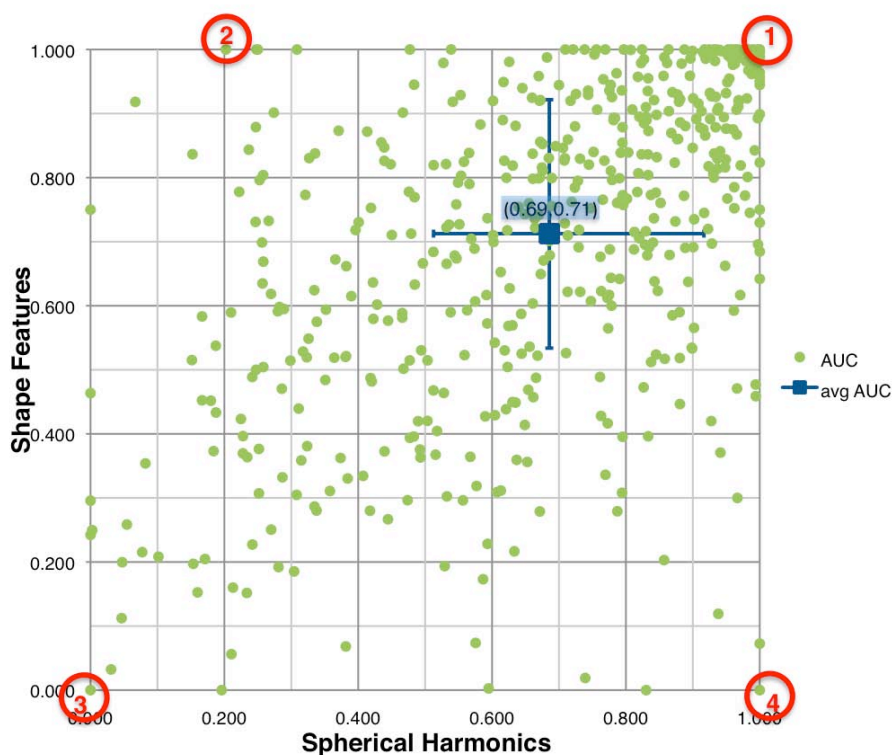


**Figure 5.9: Limitations of spherical harmonics: false ligand conformation.**

The conformational difference between the ideal coordinates (a) of GTP in the polyhedrin (PDB-Id: 2oh7) and the coordinates found in the X-ray structure (b) complicate the prediction of GTP's true density clusters in the difference map.

### 5.4.3 Spherical harmonics vs. geometric features

Having demonstrated the principal functionality of the spherical harmonic shape descriptor for the automated ligand recognition, the next step was its comparison to the geometric features that were already implemented in ARP/wARP v7.0. Not only does the comparison clarify which of the shape representations is superior over the other in predicting true density clusters, but more importantly whether the scores of the shape descriptor and shape features are correlated or not. The latter could be important for the integration of both shape representations into a single exclusive similarity score (see section 5.5)



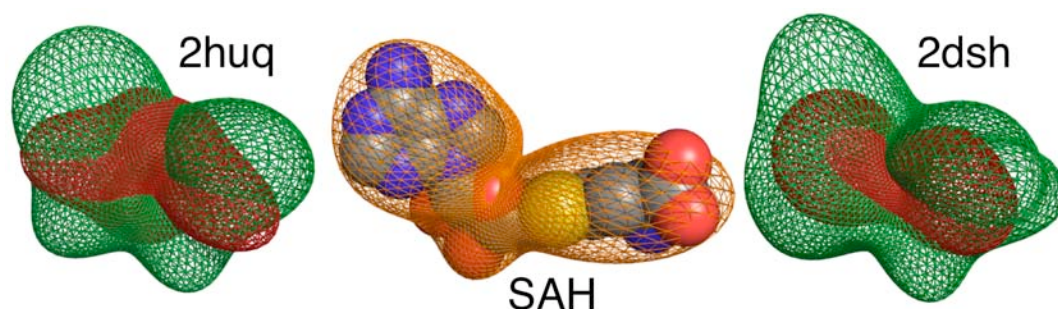
**Figure 5.10: Scatter plot of spherical harmonics vs. geometric features performances.**

Comparison of the performances of spherical harmonics shape descriptor and geometric features as measured by AUC values. The average AUC value of both shape representations is plotted in blue with the first and third quartile shown as error bars. Circled dots in red colour are discussed in detail in the main text.

### 5.4.3.1 Performance comparison

Figure 5.10 shows a scatter plot of the achieved AUC values for both spherical harmonics shape descriptor and geometric features. The average AUC value is shown as a blue square with first and third quartile depicted as error bars. On average both shape representations performed comparably well and only a closer look revealed a slight superiority of the geometric features with an average AUC of 0.71, over the shape descriptor with an average AUC = 0.69. The quartiles of the AUC values were also similar. However, for single cases large differences in the performances could be observed. In the following four extreme examples circled in red in Figure 5.10 will be discussed.

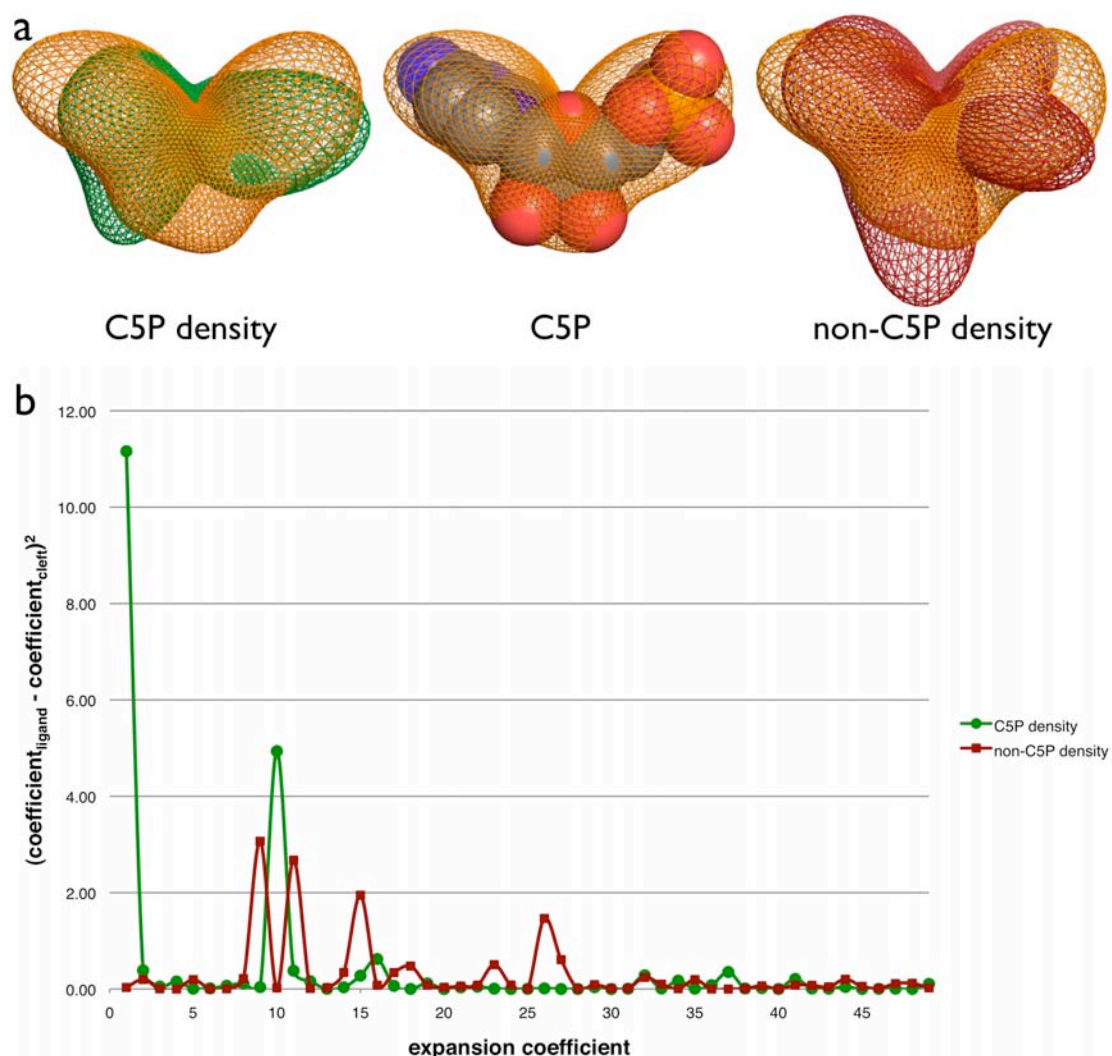
The first circle shows those cases for which both shape representations ranked all true density clusters at the top positions. 17 such cases were found in Data set III, among them five different density maps of the archaic diphthine synthetase with the substrate analogue S-adenosyl-homocysteine SAH. All five X-ray structures feature a distinct size difference between the true density clusters of the substrate analogue and the remaining false density clusters making it simple for both shape representations to predict all true density clusters (see Figure 5.11).



**Figure 5.11: Good spherical harmonics vs. good geometric features performance.**

Shape descriptors for ligand fitting work well in cases in which the true density clusters (green) have a distinct size and shape to the remaining false density clusters (red) in the difference map, such as for S-adenosyl-homocysteine (SAH) (varicoloured spheres) in the diphthine synthetases 2huq and 2dsh. The green and red coloured meshes represent the reconstructed shapes ( $l_{\max} = 6$ ) of the most similar true and false density clusters respectively and are superimposed for visual comparison reasons. The reconstructed shape of the ligand is shown as an orange coloured mesh in the centre of the figure.

The second circle in Figure 5.10 depicts the example of a cytidine-monophosphate C5P found in the *Escherichia coli* RNase 2ix0, for which the spherical harmonic shape descriptor performed poorly with an AUC = 0.2, whilst the geometric features predicted perfectly with an AUC = 1.0. An investigation into the ranking list of the shape descriptor revealed the closest



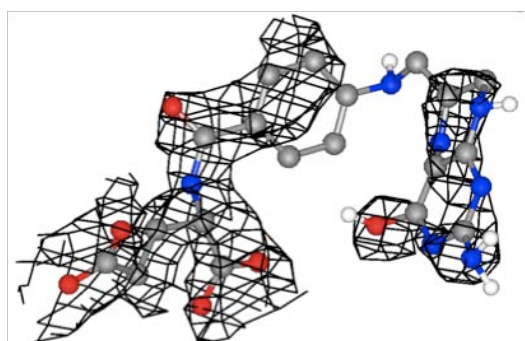
**Figure 5.12: Bad spherical harmonics vs. good geometric features performance.**

(a) At the first glance it seems that the most similar true density cluster of the C5P (green coloured mesh) in the RNase (PDB-Id: 2ix0) is more similar to C5P itself (varicoloured spheres and orange coloured mesh) than the most similar false density cluster (red coloured mesh). The meshes represent reconstructed shapes ( $l_{\text{max}} = 6$ ) calculated from the associated expansion coefficients. For visual comparison, the ligand shapes are superimposed with the density cluster shapes. The electron densities are shown in black coloured dots. (b) Plotting the square differences between ligand and density cluster coefficients shows that the first coefficient, i.e. the overall size of the false density cluster matches almost the ligand size, whilst the true density cluster separates itself size-wise.

true density cluster at rank nine with a coefficient distance of 4.09, whereas the top ranked false density cluster scored a coefficient distance of 3.75. A comparison of the reconstructed shape of C5P and the most similar true and false density cluster suggests a higher similarity for the true cluster (see Figure 5.12a). However an inspection of the distances between each ligand and density cluster coefficient reveals almost no distance between the zeroth order coefficient of the ligand and the false density cluster, indicating that both have a similar overall size (see Figure 5.12b). The true density cluster, despite having on average a smaller distance for all other coefficients, is not able to compensate the large difference in the overall size, thus scoring worse than the false density cluster.

The origin of the false density is unclear. The X-ray structure reveals a 50 Å distance to the C5P ligand binding site. However, its location in a ribonuclease II domain would suggest it to be the density of a RNA degradation product that was overlooked by the authors. Unfortunately, homologous structures in the PDB do not show any small molecule bound at the same location, giving no clues as to the nature of the density. With a working conformation generator for small molecules in CleftXplorer and the *E-coli* metabolome (Nobeli, *et al.*, 2003) it would be interesting to analyse which metabolite best fits into the density.

The third circle in Figure 5.10 illustrated the failure of both shape representations to predict the true density cluster by the case of the thymidylate synthase 1lcb and its coenzyme analogue dihydrofolic acid DHF. Both representations found all true density clusters of DHF to be least similar to the ligand, achieving an AUC without any

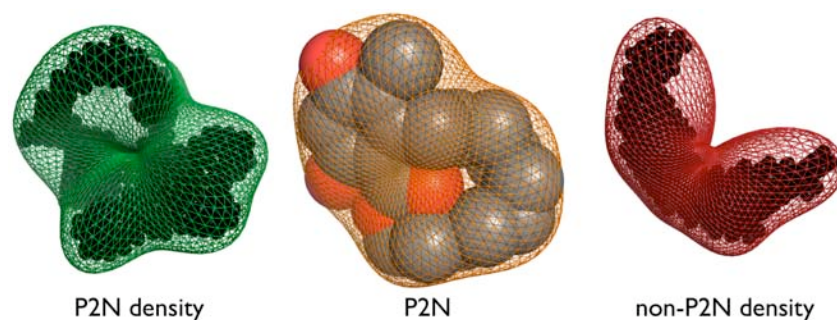


**Figure 5.13: Bad spherical harmonics vs. bad geometric features performance.**

Dihydrofolic acid DHF in thymidylate synthase (PDB-Id: 1lcb) lacks a continuous electron density cluster above the contour threshold of  $\sigma = 1$ , making it impossible for the geometric features and the shape descriptor to match the whole ligand to the partial density clusters. The density is shown at a contour threshold of  $\sigma = 1.7$  as black coloured wireframe.

score. An examination of the electron density around the ligand shows that the density is split into two distinct partial densities above a contour threshold of  $\sigma = 1$  (see Figure 5.13). Such discontinuous densities are often caused by poor phases from wrong or incomplete protein models. The discontinuous density makes it impossible for any global shape descriptor such as the two employed in this chapter to match the ligand to its density in this particular example at threshold levels above  $\sigma = 1$ .

The final circle highlights the case of the *saccharomyces cerevisiae* chaperone 2cgf inhibitor P2N, for which the spherical harmonic shape descriptor (AUC = 1.0) outperformed the geometric features (AUC = 0.07). A clear and easy explanation about the poor performance of the geometric features is difficult to state due to the nature of its score function (see 5.1 Introduction). Therefore, the focus will be rather on the spherical harmonics shape descriptor. Looking at the reconstructed shapes of P2N and the most similar true and false density clusters, it becomes evident that the true density cluster, despite the twist in its orientation, is indeed more similar to the ligand molecule in size and shape than the closest false density cluster (see Figure 5.14).



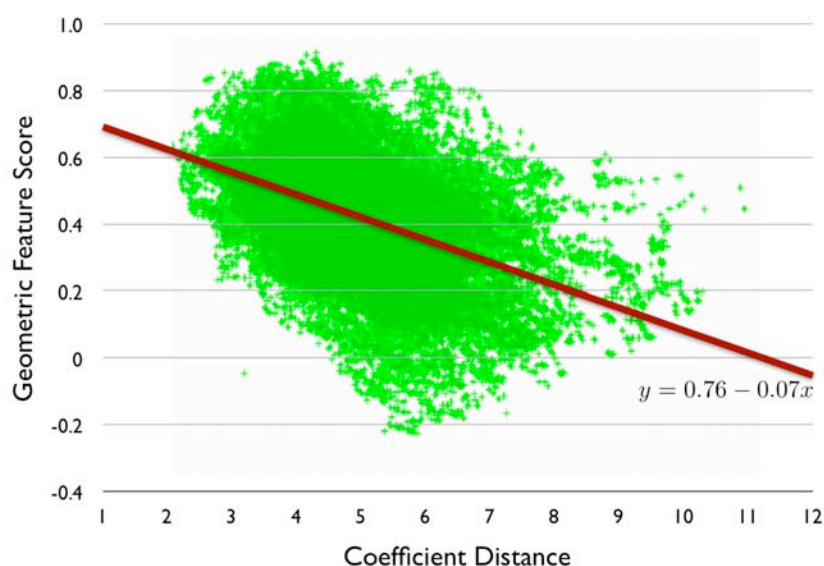
**Figure 5.14: Good spherical harmonics vs. bad geometric features performance.**

The spherical harmonic shape descriptor predicted the true density cluster (black coloured dots at the left hand side) of the inhibitor P2N (central varicoloured spheres) to the chaperone (PDB-Id: 2cgf) as the top ranked hit. In contrary did the geometric features predict a false density cluster as the most similar cluster to the ligand molecule (black coloured dots at the right hand side). The meshes around the true density cluster (green), false density cluster (red) and the ligand (orange) are reconstructed shapes calculated from expansion coefficients ( $l_{\max} = 6$ ).

### 5.4.3.2 Coefficient distance vs. geometric feature similarity

Both shape representations were comparable in predicting the true density clusters within difference density maps, with an average AUC value of around 0.70. A scatter plot (see Figure 5.15) of the feature similarities and the coefficient distances shows however that both shape representations are not similar in their scores and rather independent with modest correlation of  $R^2 = 0.2$ . The measurable correlation is negative due to the inverse relationship between the similarity metric of the geometric features and the distance metric of the spherical harmonics shape descriptor (see Figure 5.15).

The distinct similarity scores of both shape representations have an impact on the ranking list positions of the density clusters. Figure 5.16 shows the difference in the ranking positions for true density clusters in Data set III. The histogram reveals that 44.3% of all true density clusters hold a comparable ranking position with a difference of  $\pm 10\%$  relative to the complete ranking list. The majority of the true density clusters however differ by more than 10%

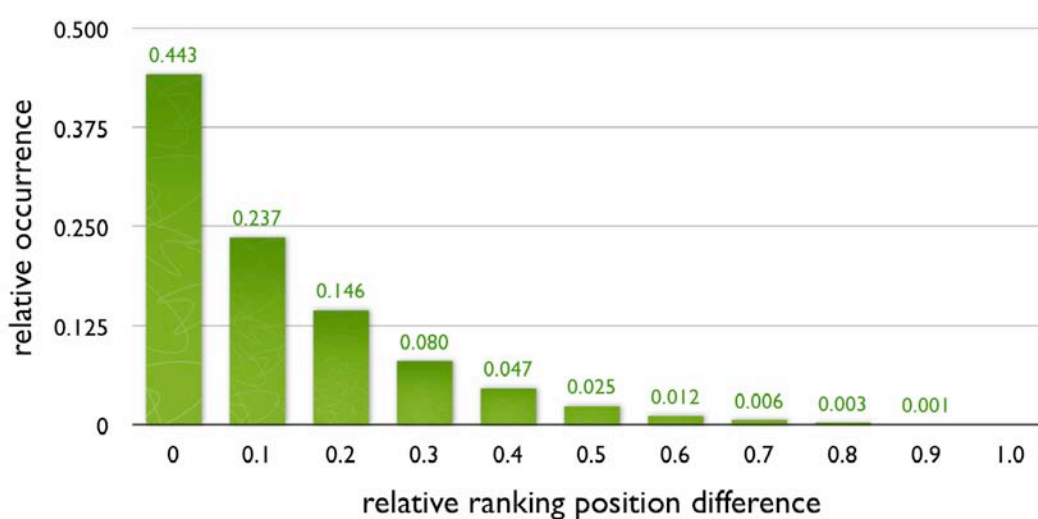


**Figure 5.15: Scatter plot of spherical harmonics vs. geometric feature scores.**

Scatter plot showing the modest correlation between the geometric feature similarity and the expansion coefficient distance between ligand molecule and true density clusters (green coloured points). The linear regression graph is given as a red coloured line together with its linear equation.



indicating a complementarity in the scoring and ranking of density clusters by both shape representations. The average Spearman rank correlation coefficient that measures the correlation between any two sequences of ranking positions is found for Data set III to be  $R' = 0.64$ . The significant deviation from the perfect positive correlation value of  $R' = 1.0$  supports the general tendency of both shape descriptors to assign all density clusters similar rank positions, however with a large number of exceptions for which the rank position for the same density cluster differs considerably among both shape representations.



**Figure 5.16: Histogram on complementary ranks between both shape representations.**

Presented are the ranking position differences for true density clusters relative to the total number of density clusters for each protein in Data set III. The ranking positions were calculated and compared between geometric feature similarity and spherical harmonic coefficient distances.

## 5.5 Discussion

The goal of any automated ligand fitting procedure in protein model building is to locate the ligand's density within a difference map and place the ligand in the correct conformation into the density without any human intervention. The spherical harmonics shape descriptor is well suited to pursue the first part of the goal as shown in the results section. The second part

however is less straightforward. Any rigid shape representation such as the geometric features implemented in the current version of ARP/wARP or the spherical harmonic shape descriptor require the calculation of shape characteristics for each ligand conformation. A systematic sampling of the entire conformational space of a ligand however can be too time consuming to be employed as a screening tool for difference electron density maps or databases. In general, the number of conformers  $N$  for a given ligand is bounded to a maximum of

$$N = \left( \frac{2\pi}{\delta} \right)^T, \quad (5.1)$$

where  $\delta$  is the rotation increment in radian and  $T$  is the number of torsion angles. In practice however, the number is less, as some conformations are impossible due to steric constraints. Given this equation, the number of conformers grows exponentially with every additional torsion bond, becoming impracticable for highly flexible molecules. An alternative to a systematic sampling is the advanced random sampling of the conformational space using a genetic algorithm or a Monte Carlo sampling technique (Leach, 2001b), combined with clustering of all conformers according to their shape similarity (Leach and Gillet, 2003a). Each cluster would be represented by a single conformation and only cluster representatives would be compared to the difference maps. After the most similar clusters are determined, a second comparison circle could be executed to determine the most similar conformations within a cluster.

However, even if the query ligand exists in the same conformation as found in the X-ray structure it is not guaranteed that all true density clusters will be found. For example, for the assessment of the expansion coefficient parameters in section 4.4.1, the ligand conformations as found in the X-ray structure were used. But rather than accomplishing a perfect prediction with maximum performance, an AUC of 0.83 was measured. In addition, one should not forget that the results presented here are the best-case scenario. The large majority of publically available electron density maps were excluded from the data set prior to

the analysis (see section 5.3) mostly due to their insufficient quality, which inevitable suggests a worsening in the performance for the majority of solved X-ray crystal structures.

The main cause for the prediction deficiency is noise that originates from erroneous phase estimates, partial atomic occupancies, non star-shaped density clusters or high local disorders of ligand molecules. The first error source can be corrected with the introduction of the ligand into the refinement process of the protein model reducing thereby the difference between experimentally observed and calculated structure factors (see section 2.2.1.2). The second error source is due to the spherical harmonics being a global shape descriptor, i.e. it assesses the overall similarity between two shapes similar to the root mean square deviation measure for a set of atomic coordinates. Although a global shape descriptor is unable to carry out a partial shape matching which is needed for the second problem, one could exploit the fragmental nature of most small molecules and truncate the ligand into several fragments. An ATP molecule, for example, could be split into an adenine, glucose, triphosphate as well as adenosine and glucose-triphosphate moiety. The spherical harmonics could then be applied to all moieties allowing it to match partial densities within difference maps. A second limitation of spherical harmonic functions is their restriction to work on star-like shapes as only star-shaped entities are defined by a single valued radial function. Highly flexible molecules however might adopt non-star shaped conformations. Several attempts have been made to work around this limitation such as the conformal mapping of closed arbitrary shaped surfaces to a unit sphere (Li and Hartley, 2007), or the application of the spherical harmonics to shells of various radii termed radial spherical extent function (Kazhdan, et al., 2003) that would cover the whole density cluster. A more elegant solution is the application of Zernike moments (Mak, et al., 2007) to describe a multi-valued radial function  $R(r)$  allowing the straight shape description of a non-star shaped density cluster. For the last problem no simple solutions exists. Flexible structural elements such as highly solvated small molecules or loop regions do not produce any strong electron density that can be measured and will always present a challenge for any automated ligand fitting software.

Despite these problems, an improvement in the prediction performance of the shape descriptor could be achieved by combining the expansion coefficients with the geometric features that are already implemented in ARP/wARP. After all, as shown in section 5.4.3.2, both shape representations score and rank complementary, i.e. distinct from each other, the similarity of density clusters. The complementarity would allow one shape representation to correct a false-positive/negative prediction of the other shape representation. However the linear combination due to the difference in the statistical nature of both scores is non-trivial (Leach and Gillet, 2003b). The expansion coefficient distance measures the distance between two shapes using a Euclidean distance metric, whereas the geometric feature similarity measures the similarity between two shapes inducing both scores to be negatively correlated. A further problem arises due to the difference in the range values of both scores. The Euclidean metric ranges from 0 to  $\infty$  with 0 being the highest similarity score, whilst the range of the geometric feature similarity values is bounded between +1 to -1 with +1 being the highest similarity score. In order to combine both scores a normalization would be required that maps every expansion coefficient distance and geometric feature similarity to the range of 0 and 1. For the bounded geometric feature similarity, such normalization is straightforward. For the unbounded coefficient distance there are two options: either choose a relative normalization constant that changes within different data sets and depends on the maximum coefficient distance found in the data set, or select an arbitrary coefficient distance as an absolute normalization constant and set all higher distances to the normalization constant. Based on the coefficient distance bins in Figure 3.6, a good candidate for an absolute normalization constant could be the coefficient distance of 10. A united score  $D$  could then be calculated with:

$$D = w_{SH} SH_{norm} + w_{GF} (1 - GF_{norm}) \quad , \quad (5.2)$$

where  $w$  are weights for both scores,  $SH_{norm}$  are the normalised expansion coefficient distances and  $GF_{norm}$  are the normalised geometric feature similarity. The success of such a combination approach remains to be tested.

## 5.6 Conclusion

Here I have presented a new fast screening methodology for the automated recognition of ligand molecules from electron density maps. The methodology is based on the comparison of expansion coefficients of spherical harmonic functions that describe the shape of density clusters (Figure 5.3) and ligand molecules. The performance of the shape descriptor was tested on a large data set of difference electron density maps (Figure 5.5) and found comparable performance to the well-established geometric features in ARP/wARP v.7.0 (Figure 5.15, Figure 5.16). However, in contrast to the geometric features, the spherical harmonic functions are a genuine shape descriptor and as such have several advantages. The speed and accuracy of spherical harmonic shape descriptors permits an effortless implementation of different screening scenarios, such as screening a specific or even all density clusters in a difference map against a large database of small molecules. The prediction performance of the shape descriptor was found to be highest for density resolutions better than 2.5 Å containing a rigid ligand with full occupancy for which good phase estimates exist (Figure 5.6 to Figure 5.14). Implemented in an existing ligand fitting software such as the one, which is part of ARP/wARP package, the spherical harmonic shape descriptor can function as a powerful fast pre-filtering and recognition method for candidate density clusters.

# Chapter 6

## Final Remarks

The work presented in this thesis challenges the concept that molecular recognition requires perfect molecular complementarity. In Chapter 3 and Chapter 4 I have shown that the shape complementarity and complementarity in physicochemical properties between small molecules and their protein receptors varies to a large extent. These results raise new questions about the nature of molecular interactions, such as:

1. To what extent is non-complementarity tolerated between proteins and ligands?
2. Is there a correlation between the binding affinity and the degree of molecular complementarity?
3. What is the role of solvent in mediating a ligand to a non-complementary binding site?
4. How does the flexibility of the binding site residues and the ligand affect molecular recognition?

### 6.1.1 Caveats

This work is limited in several respects. It might be important to distinguish the ligands in Data set I according to their functional role, i.e. whether a ligand is a substrate, product, cofactor etc. This distinction is important, as reaction products are likely to be the least complementary to their binding sites to ease their detachment from the protein. Cofactor binding sites are expected to have evolved to high complementarity to their cognate ligand molecule as in many protein families (e.g. short-chain dehydrogenases/reductases), the

cofactor remains the same throughout the family (NAD(P)(H)), whereas substrate molecules show large variation (alcohols, sugars, steroids, aromatic compounds) (Persson, et al., 2003). In my analyses, I have not distinguished between any types of ligands. All members of a ligand set were treated equal.

Furthermore, the spherical harmonics shape descriptor presented in this thesis, requires a star-shaped object. Non-star shaped objects are recast to resemble a star-shape, which in some cases can cause significant deviations from the real shape. As listed in the Discussion to Chapter 5, there are various approaches to circumvent this problem, which however have not been tested yet.

Finally, ionisable amino acids in all protein structures were protonated according to their estimated  $pK_a$  value with the empirical method of PROPKA. More accurate  $pK_a$  calculations based on Poisson-Boltzmann electrostatics (Bashford and Karplus, 1990; Miteva, *et al.*, 2005; Nielsen and McCammon, 2003; Nielsen and Vriend, 2001; Yang, *et al.*, 1993) could change estimated protonation states on some ionisable amino acids that are found in the binding site and induce changes in the calculated molecular electrostatic potentials. However, these changes are expected to be small as PROPKA was tested to reproduce experimental  $pK_a$  values with an average root mean square deviation of 0.89 (Li, et al., 2005). A factor that is neglected in most electrostatic calculations is the polarisability of single atoms within highly charged local environments. Current methodologies treat all atoms in a protein as fixed partial charges. However it was shown that the incorporation of polarisability in the electrostatic calculations is crucial and can change previously measured repulsive forces to attractive forces between protein and ligand molecules (Kundu and Gupta-Bhaya, 2004).

## 6.1.2 Future developments

Possible future developments of this project could be to implement a web interface to CleftXplorer for public use. A database could be constructed that holds and presents shape and physicochemical properties for all binding pockets in the PDB. Users could then submit their structure to the database, define the binding site and compare their binding pocket against binding sites in the PDB using spherical harmonic expansion coefficients. In addition, tools could be provided to compare homologous binding sites or binding sites with the same ligand in a similar manner as presented in this thesis. The database could be of great interest to pharmacologists, but also to anyone who seeks to understand the physicochemical nature of the function of their protein.

The next step from the scientific point of view to Chapter 3 and Chapter 4 would be the experimental determination of accurate binding affinity data for each protein-ligand complex in the data set. The computational predictions of binding affinities is still limited in accuracy (Gilson and Zhou, 2007). The binding free energy calculations conducted on Data set I (see Table 4.3) were lacking any correlation with the observed variation in the physicochemical properties, making it impossible to explain the origin of the variation. Accurate affinity data from experiments could change this outcome. It can be expected however that the biological necessity of intermolecular interactions may well require low affinity complexes to be favoured.

The automated recognition of ligands in electron density maps with spherical harmonic shape descriptors currently lacks accuracy. Using the unrefined electron density alone for ligand recognition seems insufficient and demands further information to be added into the electron density fitting procedure in the form of constraints. These constraints could come from NMR spectroscopy experiments, where nuclear spin transfer effects, also called Nuclear Overhauser Effects (NOE), between the protein and ligands or chemical shift changes of the protein upon ligand binding are tested (Meyer and Peters, 2003). Computational constraints



could be placed by virtual screening methods on the preliminary protein model excluding all ligand candidates that do not fit into the binding site.

### 6.1.3 Function prediction

The initial aim behind the development of CleftXplorer was its application for protein function prediction by comparing a query binding site, from a protein of unknown function, against a database of annotated binding sites or potential ligand molecules. The analysis and results in Chapter 3 made it clear that CleftXplorer's performance was acceptable only if the binding pocket could be accurately defined in the Interact or the Ligand Cleft Model. However, for proteins of unknown function, neither the exact location nor the precise identity of all binding site residues is known. Estimating the identity of binding site residues based on their evolutionary conservation resulted in large cleft models with adjacent substrate and cofactor binding pockets that were merged together into a single Conserved Cleft Model. For the spherical harmonics shape descriptor to work as a function prediction method, the large Conserved Cleft Models must be divided into smaller volumes such that each volume gives an accurate representation of the binding pocket. In collaboration with Dr. Kazuto Yamazaki a framework was developed for the spherical harmonic functions to be employed as local shape descriptors in protein-ligand induced-fit docking (Yamazaki, et al., 2009). Within the framework, a cleft model of a binding site is divided into a diverse set of Voronoi partitions. All adjacent partitions are enumerated and recombined to give *subsite candidates*. Subsites that have on average a repulsive van der Waals potential and are smaller than a minimum size are disregarded, while the remaining subsite candidates are described with the spherical harmonic shape descriptor and compared against expansion coefficients of chemical compounds. For function prediction purposes, the same framework could be applied to obtain an accurate cleft model of a binding pocket from within a Conserved Cleft Model. Important for such a function prediction method would be the incorporation of a probabilistic term in the similarity search that accounts for the variation in the molecular complementarity as observed

in this thesis. A workaround to the probabilistic description could be the utilization of '3D consensus binding profiles'. Such profiles would illustrate the average shape of all binding sites that bind the same ligand and represent the physicochemical environment that a ligand on average experiences in various proteins. A similarity search would then involve the comparison of the geometrical and physicochemical properties of a query binding site against a set of 3D consensus binding profiles.

### **6.1.4 Final conclusion**

Discovering the principles of molecular recognition is vital if computational biology is to become a predictive discipline that is able to model and simulate cellular processes in living cells. Without a comprehensive knowledge in molecular recognition, computational biology will remain a monitoring science unable to reliably predict molecular interactions in living cells. The two main challenges in molecular interaction prediction remain the flexibility of protein binding partners and the inability of scoring functions to distinguish true from false positive predictions (Janin and Wodak, 2007; Sousa, et al., 2006).

Protein dynamics and motions in X-ray structures are usually only visible as a lack of 'clarity' caused by the averaging process over many molecules. Molecular dynamics simulations attempt to overcome this obstacle by simulating motions in proteins using the X-ray structure as the starting point for their calculation. The steadily growing computer power, development of faster algorithms and better physicochemical parameterization in recent years have improved dynamic simulations (Dodson and Verma, 2006). Soon, larger molecular dynamic simulations will be possible and hopefully allow a deeper investigation of the importance of protein dynamics in molecular recognition (Karplus and McCammon, 2002).

But most likely, the explicit simulation of water molecules in and around proteins will have the biggest impact on our comprehension of molecular recognition. Molecules are solvated in

water and their interaction occurs in water. For many years water was necessarily omitted in molecular docking and mapping applications as their *in silico* simulation was computationally expensive. It was hoped that shape and physicochemical complementarity would be sufficient to drive molecular interactions. But many crystal structures of proteins show conserved water molecules at binding interfaces or next to binding sites and suggest an active role for water molecules in the protein-ligand complex. Especially for molecular parts that interact via hydrophobic interactions, water acts as a 'molecular glue' and induces the interaction of protein and ligand molecules. The first methodologies that simulated hydration effects on protein structures considered water as a continuum but had in general limited success. A second generation of simulation software treated water molecules explicitly but did not reach the expected accuracy especially due to the immense computational cost that dynamic simulations require. The growing computer power will eventually help in this field to provide simulations of hydration effects under physical conditions (Levy and Onuchic, 2006).

As long as the general mechanisms of molecular recognition and binding are little understood, successful predictions of molecular interactions will remain rare. However, once we achieve a comprehensive understanding of the fundamental processes in molecular binding, the *de novo* design of enzymes, i.e. the alternation of the enzymatic function, will be within reach. Other than inorganic catalysts, enzymes catalyze their reactions under mild conditions with high specificity and rate enhancements. This unique property makes enzymes attractive for many industrial processes although often they do not catalyze the required chemical reactions. Methods like rational-design and directed evolution in protein engineering have shown to be very useful in producing desired functionality in enzymes. As the factors for protein integrity namely, hydrogen bonds and hydrophobic effects, are well understood, many enzymes have been successfully altered to stabilize the structural integrity against harmful chemicals, or extreme temperature and pH conditions. Comparable results could not be obtained for altering the catalytic machinery of enzymes (Bolon, et al., 2002). Only few enzymes so far have been successfully altered like the modification of an inert ribose-binding protein into a highly active triose-phosphate-isomerase (Dwyer, et al., 2004).

Maybe, the challenges that we face in computational biology at the beginning of the 21<sup>st</sup> century are still very difficult, but I believe that advances in theory, algorithms and computer power will eventually lead to the golden goal of computational biology, namely the *in silico* simulations of a living cell in atomic detail.

# Appendix A

## Data set I

Table A.1: Data set of 100 binding sites being non-homologous in 9 ligand sets.

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
1	<i>AMP</i>	12as	A	Asparagine synthetase	6.3.1.1	3.30.930.10	AMP	X	2	–
2		1amu_1	A	Gramicidin synthetase	5.1.1.11	2.30.38.10 3.40.50.980	AMP	A	551	–
3		1c0a	A	Aspartyl t-RNA synthetase	6.1.1.12	3.30.1360.30	AMP	E	800	–
4		1ct9_1	A	Asparagine synthetase	6.3.5.4	3.40.50.620	AMP	A	1100	–
5		1jp4	A	Bisphosphate nucleotidase	3.1.3.7	3.40.190.80	AMP	B	601	–
6		1kht	B	Adenylate kinase	2.7.4.3	3.40.50.300	AMP	D	2193	–
7		1qb8	A	Adenine phosphoribosyltransferase	2.4.2.7	3.40.50.2020	AMP	C	300	–

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
8		1tb7	B	cAMP-specific-cyclic phosphodiesterase	3.1.4.17	1.10.1300.10	AMP	C	401	_
9		8gpb	A	Glycogen phosphorylase	2.4.1.1	3.40.50.2000	AMP	B	930	_
10	ATP	1a0i	_	ATP-dependent DNA ligase	6.5.1.1	3.30.470.30 3.30.1490.70	ATP	_	1	_
11		1a49_1	A	Pyruvate kinase	2.7.1.40	3.20.20.60	ATP	A	535	_
12		1ayl	A	Phosphoenolpyruvate carboxykinase	4.1.1.49	2.170.8.10 3.90.228.20	ATP	A	541	_
13		1b8a	A	Aspartyl-tRNA synthetase	6.1.1.12	3.30.930.10	ATP	C	500	_
14		1dv2	A	Biotin carboxylase	6.3.4.14	3.30.470.20 3.30.1490.20	ATP	C	1000	_
15		1dy3	A	Pyrophosphokinase	2.7.6.3	3.30.70.560	ATP	A	200	_
16		1e2q	A	Thymidylate kinase	2.7.4.9	3.40.50.300	ATP	A	302	_
17		1e8x	A	Phosphatidylinositol kinase	2.7.1.153	1.10.1070.11 3.30.1010.10	ATP	A	2000	_
18		1esq	A	Hydroxyethylthiazole kinase	2.7.1.50	3.40.1190.20	ATP	D	300	_
19		1gn8	B	Phosphopantetheine adenylyltransferase	2.7.7.3	3.40.50.620	ATP	B	600	_
20		1kvk	A	Mevalonate kinase	2.7.1.36	3.30.230.10	ATP	C	535	_
21		1o9t	A	Adenosylmethionine synthetase	2.5.1.6	3.30.300.10	ATP	B	1397	_
22		1rdq	E	cAMP-dependent protein kinase	2.7.1.37	1.10.510.10 3.30.200.20	ATP	A	600	B
23		1tid	A	Anti-sigma F factor	2.7.1.37	3.30.565.10	ATP	E	200	_

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
24	<i>FAD</i>	1cqx	A	Flavoheomprotein	1.14.12.1	2.40.30.10	FAD	A	405	_
					7	3.40.50.80				
25		1e8g	B	Vanillyl-alcohol oxidase	1.1.3.38	3.30.43.10	FAD	B	600	_
						3.30.465.20				
26		1evi	B	D-amino acid oxidase	1.4.3.3	3.30.9.10	FAD	C	353	_
						3.40.50.720				
27		1h69_1	A	NAD(P)H dehydrogenase	1.6.99.2	3.40.50.360	FAD	A	1274	_
28		1hsk	A	Acetylenolpyruvoylglucosamine reductase	1.1.1.158	3.30.43.10	FAD	D	401	_
						3.30.465.10				
29		1jqj	A	Short chain acyl-CoA dehydrogenase	1.3.99.2	1.20.140.10	FAD	E	399	_
						2.40.110.10				
30		1jr8	B	Oxidoreductase	1.8.3.?	1.20.120.310	FAD	C	334	_
31		1k87	A	Proline dehydrogenase	1.5.99.8	3.20.20.220	FAD	C	2001	_
32		1pox	A	Pyruvate oxidase mutant	1.2.3.3	3.40.50.1220	FAD	A	612	_
						3.40.50.970				
33		3grs	A	Glutathione reductase	1.8.1.7	3.50.50.60	FAD	A	479	_
34	<i>FMN</i>	1dnl	A	Pyridoxine-phosphate oxidase	1.4.3.5	2.30.110.10	FMN	C	250	_
35		1f5v	A	Oxidoreductase	1.?.?.?	3.40.109.10	FMN	C	360	_
36		1ja1_1	A	NADPH-cytochrome reductase	1.6.2.4	3.40.50.360	FMN	A	1751	_
37		1mvl	A	Lyase	4.1.1.36	3.40.50.1950	FMN	D	1001	_
38		1p4c	A	Mandelate dehydrogenase	1.1.3.15	3.20.20.70	FMN	E	490	_

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
39		1p4m	A	Transferase	2.7.1.26	2.40.30.30	FMN	B	401	_
40	<i>Glucose</i>	1bdg	A	Hexokinase	2.7.1.1	3.30.420.40 3.40.367.20	GLC	A	501	_
41		1cq1	A	Quinoprotein glucose dehydrogenase	1.1.5.2	2.120.10.30	GLC	C	3	_
42		1k1w	A	Transferase	2.4.1.25	3.20.20.? 1.20.?.? 2.70.98.?	GLC	C	653	_
43		1nf5_2	C	Transferase	?.?.?.?	1.10.530.10 3.90.550.10	GLC	D	527	_
44		2gbp	_	Periplasmic binding protein	?.?.?.?	3.40.50.2300	GLC	_	310	_
45	<i>Heme</i>	1d0c	A	Endothelial nitric oxide synthase	1.14.13.3 9	3.90.340.10	HEM	A	500	_
46		1d7c	A	Cellobiose dehydrogenase	1.1.99.18	2.60.40.1210	HEM	A	401	_
47		1dk0	A	Heme-binding protein	?.?.?.?	3.30.1500.10	HEM	A	200	_
48		1eqg	A	Prostaglandin synthase	1.14.99.1	1.10.640.10	HEM	A	601	_
49		1ew0	A	Transferase	2.7.3.?	3.30.450.20	HEM	A	501	_
50		1gwe	A	Catalase	1.11.1.6	2.40.180.10	HEM	A	504	_
51		1iqc_1	A	Heme peroxidase	1.11.1.5	1.10.760.10	HEM	A	401	_
52		1naz	E	Oxygen transport	?.?.?.?	1.10.490.10	HEM	E	200	_
53		1np4	B	Nitrophorin	?.?.?.?	2.40.128.20	HEM	B	185	_
54		1po5	A	Cytochrome	1.14.14.1	1.10.630.10	HEM	A	500	_



No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
55		1pp9	C	Oxidoreductase	?.?.?.?	1.20.810.10	HEM	C	501	_
56		1qhu	A	Binding protein hemopexin	?.?.?.?	2.110.10.10	HEM	A	500	_
57		1qla	C	Oxidoreductase	?.?.?.?	1.20.950.10	HEM	G	1	_
58		1qpa	B	Lignin peroxidase	1.11.1.14	1.10.420.10 1.10.520.10	HEM	B	350	_
59		1sox	A	Sulfite oxidase	1.8.3.1	3.10.120.10	HEM	A	502	_
60		2cpo	_	Oxidoreductase	1.11.1.10	1.10.489.10	HEM	_	396	_
61	NAD	1ej2	B	Nicotinamide adenyltransferase	2.7.7.1	3.40.50.620	NAD	H	1339	_
62		1hex	A	Isopropylmalate dehydrogenase	1.1.1.85	3.40.718.10	NAD	A	400	A
63		1ib0	A	NADH-cytochrome reductase	1.6.2.2	3.40.50.80	NAD	B	1994	_
64		1jq5	A	Glycerol dehydrogenase	1.1.1.6	1.20.1090.10 3.40.50.1970	NAD	I	401	_
65		1mew	A	Monophosphate dehydrogenase	1.1.1.205	3.20.20.70	NAD	E	987	_
66		1mi3_1	A	Oxidoreductase	1.1.1.21	3.20.20.100	NAD	A	1350	_
67		1o04_1	A	Aldehyde dehydrogenase	1.2.1.3	3.40.309.10 3.40.605.10	NAD	A	6501	_
68		1og3	A	T-cell ADP-ribosyltransferase	2.4.2.31	2.30.100.10	NAD	A	1227	_
69		1qax	A	Methylglutaryl-coenzyme reductase	1.1.1.88	3.30.70.420 3.90.770.10	NAD	G	1001	_
70		1rlz	A	Deoxyhypusine synthase	2.5.1.46	3.40.910.10	NAD	H	700	_
71		1s7g	B	NAD-dependent deacetylase	3.5.1.?	3.40.50.1220	NAD	F	701	_

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
72		1t2d	A	Lactate dehydrogenase	1.1.1.27	3.40.50.720 3.90.110.10	NAD	E	316	_
73		1tox_1	A	Diphtheria toxin	2.4.2.36	3.90.175.10	NAD	A	536	_
74		2a5f	B	Protein transport	2.4.2.36	3.90.210.10	NAD	C	1536	_
75		2npx	A	NADH peroxidase	1.11.1.1	3.50.50.60	NAD	A	818	_
76	<i>Phosphate</i>	1a6q	_	Phosphatase	3.1.3.16	3.60.40.10	PO4	_	701	_
77		1b8o	C	Purine nucleoside phosphorylase	2.4.2.1	3.40.50.1580	PO4	F	599	_
78		1brw	A	Pyrimidine nucleoside phosphorylase	2.4.2.2	3.40.1030.10	PO4	C	2001	_
79		1cqj_1	B	Succinyl-CoA synthetase	6.2.1.5	3.30.1490.20	PO4	B	904	_
80		1d1q	B	Tyrosine phosphatase	3.1.3.48	3.40.50.270	PO4	C	402	_
81		1dak	A	Dethiobiotin synthetase	6.3.3.3	3.40.50.300	PO4	C	803	_
82		1e9g	A	Inorganic pyrophosphatase	3.6.1.1	3.90.80.10	PO4	A	3001	A
83		1ejd	C	Enolpyruvyltransferase	2.5.1.7	3.65.10.10	PO4	F	2431	_
84		1euc	A	Succinyl-CoA synthetase	6.2.1.4	3.40.50.261	PO4	C	224	_
85		1ew2	A	Phosphatase	3.1.3.1	3.40.720.10	PO4	C	1005	_
86		1fbt	B	Bisphosphatase	3.1.3.46	3.40.50.1240	PO4	C	100	_
87		1gyp	A	Glyceraldehyde-phosphate dehydrogenase	1.2.1.12	3.30.360.10	PO4	A	359	_
88		1h6l	A	Phytase	3.1.3.8	2.120.10.20	PO4	A	501	_
89		1ho5_1	B	Nucleotidase	3.1.3.5	3.60.21.20	PO4	B	2603	_
90		1l5w	A	Maltodextrin phosphorylase	2.4.1.1	3.40.50.2000	PO4	D	998	_
91		1l7m_1	A	Phosphoserine phosphatase	3.1.3.3	3.40.50.1000	PO4	A	720	_

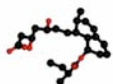
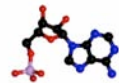
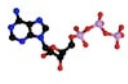
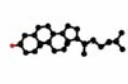
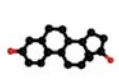
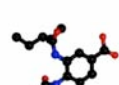
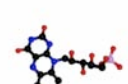

No	Ligand Set	PQS-Id	Chain-Id	Protein	EC Code	CATH Code	Ligand	Ligand Chain-Id	Ligand Residue Number	Ligand Altern Loc
92		1lby	A	Bisphosphatase	3.1.3.25	3.30.540.10 3.40.190.80	PO4	C	293	_
93		1lyv	A	Protein-tyrosine phosphatase	3.1.3.48	3.90.190.10	PO4	B	1000	_
94		1qf5	A	Adenylosuccinate synthetase	6.3.4.4	3.40.440.10	PO4	C	2	_
95		1tco	A	Serine-threonine phosphatase	3.1.3.16	3.60.21.10	PO4	D	507	_
96	<i>Steroid</i>	1e3r	B	Isomerase	5.3.3.1	3.10.450.50	AND	B	801	_
97		1fds	A	Hydroxysteroid-dehydrogenase	1.1.1.62	3.40.50.720	EST	A	350	_
98		1j99	A	Alcohol sulfotransferase	2.8.2.2	3.40.50.300	AND	B	401	A
99		1lhu	A	Sex hormone-binding globulin	?.?.?.?	2.60.120.200	EST	G	301	_
100		1qkt	A	Estradiol receptor	?.?.?.?	1.10.565.10	EST	C	600	_

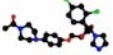
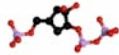
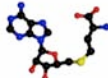
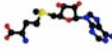
\_ is a placeholder for unlabelled chains and alternative locations. ? is a placeholder for unavailable information.

# Appendix B

## Data set II

Table B.1: Small data set of 12 ligand molecules within difference electron density maps.

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
1	114 	1hw8	2.10	220	169
2	AMP 	1mf0	2.50	141	67
3	ATP 	1ii0	2.40	295	105
4	CLR 	2rh1	2.40	102	28
5	EST 	1iol	2.30	63	27
6	FDI 	1b9s	2.50	48	20
7	FMN 	1oo5	2.50	63	41
8	HYF 	1m13	2.15	18	12

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
9	KTN 	1jin	2.30	80	63
10	PCP 	1a95	2.00	205	125
11	SAH 	2dsi	2.20	80	22
12	SAM 	1h1d	2.00	154	84

The table shows from left to right the ligand's PDB three letter code, the ligand in ball-stick representation as found in PDBsum (Laskowski, et al., 2005), the PDB-Id of the protein structure from which the ligand was extracted, the resolution of the protein's X-ray data, the total number of density ( $\rho$ ) clusters after fragmentation tree filtering, the total number of redundant density ( $\rho$ ) clusters found at the same location of the ligand in the difference electron density map.

## Data set III

Table B.2: Large data set of 536 ligand molecules within difference electron density maps.

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
1	GTT	11gs	2.30	19	30
2	GDP	1a4r	2.50	82	75
3	GNH	1a4r	2.50	75	82
4	PCP	1a96	2.00	114	113
5	TOL	1ah3	2.30	33	83
6	NPE	1aj7	2.10	8	14
7	BOG	1aua	2.50	53	19
8	FKP	1azs	2.30	49	88

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
9	GSP	1azs	2.30	88	49
10	UFP	1b02	2.50	8	32
11	ATP	1b39	2.10	10	6
12	RA2	1b9v	2.35	54	32
13	SAH	1bc5	2.20	74	50
14	MNO	1bmq	2.50	74	28
15	PIC	1bzj	2.25	62	26
16	PQQ	1c9u	2.20	65	18
17	MAL	1cdg	2.00	93	38
18	MAL	1cgv	2.50	173	44
19	101	1cs4	2.50	49	173
20	FOK	1cs4	2.50	62	164
21	GSP	1cs4	2.50	88	132
22	FOK	1cul	2.40	46	129
23	GSP	1cul	2.40	80	90
24	MAL	1cxe	2.10	162	45
25	MTX	1d1g	2.10	73	98
26	DGP	1del	2.20	24	38
27	ADP	1djn	2.20	176	208
28	BMS	1dkf	2.50	88	33
29	OLA	1dkf	2.50	52	81
30	MHF	1dnp	2.30	96	140
31	MTX	1dre	2.60	73	37
32	GNT	1dx6	2.30	85	65
33	HUX	1e66	2.10	88	42
34	ATP	1e8x	2.20	68	123
35	ATP	1ee1	2.06	42	93
36	E20	1eve	2.50	78	90
37	ADP	1f48	2.30	112	87
38	AMP	1fa9	2.40	47	76
39	BRL	1fm6	2.10	65	62
40	REA	1fm6	2.10	66	76
41	GSB	1fro	2.20	67	44
42	PNN	1fxv	2.25	11	26
43	GDP	1gim	2.50	41	48
44	IMP	1gim	2.50	52	51
45	GDP	1gin	2.80	81	25
46	IMP	1gin	2.80	36	81
47	DH2	1gp5	2.20	44	70
48	DQH	1gp5	2.20	51	63
49	LAT	1gwv	2.50	66	119
50	UDP	1gwv	2.50	114	62
51	BIA	1h1d	2.00	61	84
52	E10	1h22	2.15	42	119
53	E12	1h23	2.15	49	83
54	GDP	1h2t	2.10	68	32

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
55	AIC	1h8s	2.40	12	27
56	GDP	1ha3	2.00	161	126
57	AOE	1hj1	2.30	20	28
58	SAH	1hnn	2.40	159	128
59	FMN	1huv	2.15	47	30
60	ADP	1hw8	2.10	58	197
61	D16	1i00	2.50	36	43
62	UMP	1i00	2.50	85	56
63	ADP	1ihu	2.15	120	95
64	ADP	1ii0	2.40	81	259
65	ENA	1isi	2.10	107	86
66	NMN	1isj	2.30	55	160
67	DCU	1j07	2.35	33	105
68	ATP	1j21	2.20	265	112
69	WRA	1j3i	2.33	90	101
70	UMP	1j3k	2.10	144	179
71	OLA	1j78	2.31	31	59
72	739	1jcq	2.30	87	123
73	FPP	1jcq	2.30	88	143
74	SUC	1jcq	2.30	58	171
75	EST	1jgl	2.15	39	23
76	UVC	1jh7	2.40	22	13
77	DEB	1jio	2.10	40	88
78	FMN	1jnw	2.07	41	20
79	DEQ	1jt6	2.54	48	185
80	GTT	1k0a	2.50	30	58
81	MAL	1k1y	2.40	27	56
82	FMN	1kbi	2.30	131	56
83	AMP	1kht	2.50	78	24
84	IMP	1kkf	2.60	44	43
85	YPA	1knu	2.50	14	46
86	SAH	1kyw	2.40	61	30
87	SAH	1kyz	2.20	81	28
88	AMP	1kz8	2.00	99	70
89	PFE	1kz8	2.00	69	91
90	NCN	1l4f	2.10	40	20
91	7RP	1l5l	2.00	77	46
92	7RA	1l5m	2.00	67	39
93	T80	1lbt	2.50	31	46
94	DHF	1lcb	2.50	7	7
95	TMP	1lcb	2.50	12	12
96	TMF	1lce	2.50	26	13
97	FPP	1ld7	2.00	76	143
98	SUC	1ld7	2.00	57	163
99	U66	1ld7	2.00	85	123
100	IMO	1lny	2.20	94	36

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
101	IMO	1lon	2.10	51	19
102	GER	1lv0	2.00	15	44
103	5GP	1lvg	2.10	65	48
104	ADP	1lvg	2.10	39	63
105	AMP	1mc1	2.16	47	58
106	GSH	1md3	2.03	41	48
107	GSH	1md4	2.10	81	41
108	MOA	1mei	2.20	36	88
109	XMP	1mei	2.20	88	36
110	GDP	1mez	2.40	78	34
111	GDP	1mf0	2.50	81	64
112	IDP	1mrd	2.30	66	31
113	GDP	1mre	2.30	61	33
114	TDG	1muq	2.30	39	111
115	BNE	1mzc	2.00	59	105
116	FPP	1mzc	2.00	83	113
117	SUC	1mzc	2.00	55	144
118	FMN	1n07	2.45	23	91
119	SO1	1n0u	2.12	68	28
120	GDR	1n7g	2.20	69	267
121	GDP	1nht	2.50	19	16
122	PGS	1nht	2.50	18	20
123	153	1nhu	2.00	27	18
124	ADP	1njf	2.30	46	156
125	ATG	1njf	2.30	151	40
126	ADP	1nks	2.57	85	295
127	AMP	1nks	2.57	264	119
128	MAL	1nl5	2.10	72	29
129	APC	1nus	2.20	55	50
130	NMN	1nus	2.20	79	41
131	AMP	1nv7	2.15	97	75
132	UFP	1o28	2.10	301	86
133	LMS	1obh	2.20	125	58
134	A8B	1odc	2.20	71	31
135	FMN	1ofd	2.00	141	176
136	BEL	1oon	2.49	25	50
137	MAL	1ot2	2.10	55	169
138	TAL	1oum	2.40	90	27
139	DBM	1ov6	2.40	33	19
140	P33	1oxn	2.20	37	58
141	P33	1oxq	2.30	28	84
142	ADP	1p61	2.21	62	52
143	ADP	1p72	2.10	48	37
144	GDP	1p9b	2.00	50	25
145	IMO	1p9b	2.00	47	26
146	MAL	1pez	2.32	107	55



No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
147	ADP	1pfk	2.40	198	79
148	PRX	1pg3	2.30	123	43
149	FFO	1pj7	2.10	46	73
150	MAL	1pj9	2.00	59	103
151	ANP	1pjk	2.50	10	12
152	880	1pmq	2.20	55	15
153	TPP	1pow	2.50	166	149
154	TPP	1pox	2.10	140	152
155	FMN	1ps9	2.20	62	196
156	AMP	1ptw	2.30	68	20
157	LAT	1puu	2.30	84	229
158	GSH	1px6	2.10	101	48
159	GSH	1px7	2.03	79	51
160	A3P	1q20	2.30	74	44
161	PLO	1q20	2.30	49	70
162	A3P	1q22	2.50	77	63
163	AND	1q22	2.50	75	70
164	ATP	1qhg	2.50	76	27
165	ADP	1r0y	2.55	185	56
166	MTX	1rb3	2.30	101	138
167	AFB	1re0	2.40	76	92
168	GDP	1re0	2.40	68	61
169	MTX	1rh3	2.40	67	72
170	ADP	1rk2	2.25	126	81
171	DEO	1ros	2.00	50	35
172	TYD	1rrv	2.00	113	76
173	D7P	1rs9	2.22	76	275
174	MTX	1rx3	2.20	34	52
175	DDF	1rx4	2.20	35	56
176	RIO	1s3z	2.00	57	65
177	APR	1s7g	2.30	40	152
178	ATP	1s9j	2.40	79	39
179	BBM	1s9j	2.40	41	88
180	FPP	1sa4	2.10	88	132
181	JAN	1sa4	2.10	69	119
182	SUC	1sa4	2.10	65	157
183	MTH	1sd2	2.10	51	18
184	666	1so2	2.40	177	187
185	ORX	1szz	2.15	65	178
186	PLG	1szz	2.15	76	192
187	C8E	1t16	2.60	39	20
188	B3N	1t67	2.31	20	29
189	ADP	1t6x	2.29	25	82
190	BGL	1taq	2.40	47	23
191	BOG	1tcb	2.10	68	52
192	BOG	1tcc	2.50	40	85

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
193	FMN	1tll	2.30	113	259
194	GHA	1tqu	2.03	33	20
195	ADP	1tuu	2.50	43	52
196	AMP	1tuu	2.50	50	51
197	FMN	1tv5	2.40	35	122
198	N8E	1tv5	2.40	43	144
199	SAM	1tv8	2.20	113	228
200	FRZ	1tvo	2.50	24	23
201	CB3	1tvv	2.30	13	7
202	CB3	1tw	2.50	9	26
203	SAH	1tw2	2.50	133	44
204	SAH	1tw3	2.35	100	77
205	CBS	1tw5	2.30	89	117
206	UDH	1tw5	2.30	106	88
207	PA7	1u0y	2.30	51	45
208	BAU	1u1c	2.20	245	66
209	181	1u1d	2.00	298	75
210	NEC	1u2o	2.10	75	83
211	MTX	1u70	2.50	29	59
212	MAL	1ua3	2.01	33	36
213	MLR	1ua3	2.01	30	23
214	ALH	1ung	2.30	31	36
215	IXM	1unh	2.35	58	97
216	RRC	1unl	2.20	18	68
217	A8N	1ut6	2.40	59	171
218	PF3	1utz	2.50	114	66
219	AMP	1uxn	2.30	83	24
220	AMP	1uxu	2.25	82	136
221	AMP	1uxv	2.35	81	37
222	NFG	1uyq	2.20	56	42
223	D1L	1uyr	2.50	76	64
224	SLB	1v3c	2.30	104	60
225	DAN	1v3d	2.28	124	103
226	HA1	1v48	2.20	49	28
227	MRK	1v4s	2.30	49	54
228	AMP	1v8s	2.20	10	5
229	P2S	1va6	2.10	51	43
230	BNG	1vgo	2.50	71	73
231	TES	1vpo	2.15	21	21
232	FMN	1vrq	2.20	81	163
233	ACD	1vyg	2.40	4	1
234	CB3	1vzd	2.50	10	9
235	CB3	1vze	2.30	9	26
236	UMP	1vze	2.30	47	45
237	SIA	1w1x	2.00	98	204
238	SIA	1w20	2.08	113	156

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
239	SIA	1w21	2.08	113	155
240	TS0	1w29	2.30	29	127
241	TS1	1w29	2.30	98	127
242	DN1	1w3c	2.30	24	37
243	DN2	1w3c	2.30	13	37
244	ADP	1w4b	2.30	42	11
245	GL8	1w4l	2.16	63	34
246	ADP	1w5t	2.40	38	126
247	ANP	1w5t	2.40	111	33
248	GNT	1w6r	2.05	88	42
249	EQU	1w6y	2.10	11	14
250	GNT	1w76	2.30	94	123
251	TDP	1w88	2.30	112	46
252	DAN	1wcq	2.10	191	56
253	SAH	1wng	2.10	54	24
254	ANP	1wuu	2.50	273	86
255	FRK	1wxy	2.50	18	8
256	F29	1wzy	2.50	16	11
257	FMN	1x31	2.15	88	176
258	ADP	1x3m	2.20	21	42
259	ATP	1xdp	2.50	14	9
260	SAM	1xds	2.30	104	78
261	REA	1xiu	2.50	70	27
262	ATP	1xkv	2.20	34	100
263	188	1xkw	2.00	51	22
264	OCB	1xl8	2.20	56	42
265	CIO	1xlx	2.19	47	20
266	PIL	1xm4	2.31	68	25
267	G7M	1xmm	2.50	63	125
268	M7G	1xmm	2.50	121	53
269	7DE	1y2j	2.55	17	47
270	FMN	1y30	2.20	1	4
271	GDP	1y3a	2.50	254	140
272	C0R	1y5r	3.00	108	162
273	E89	1y5x	2.10	79	28
274	CIE	1ybh	2.50	88	234
275	APC	1ybu	2.40	9	5
276	APR	1yc2	2.40	50	135
277	PQQ	1yiq	2.20	63	87
278	BOG	1yk3	2.20	48	104
279	BNG	1ymg	2.24	52	19
280	PY4	1ynu	2.25	32	95
281	ADP	1yp4	2.30	121	130
282	ADQ	1yp4	2.30	41	130
283	AMP	1yxu	2.24	60	65
284	AMP	1yz0	2.07	136	158

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
285	SAH	1yz3	2.40	158	87
286	BOG	1yz4	2.40	28	7
287	2FA	1z35	2.50	88	24
288	DAN	1z4v	2.30	88	230
289	DAN	1z4z	2.50	82	219
290	DEB	1z8q	2.00	26	67
291	CTP	1za2	2.50	89	105
292	C01	1zeo	2.50	37	19
293	2HI	1zg3	2.35	58	35
294	SAH	1zg3	2.35	26	50
295	HMK	1zga	2.35	24	42
296	SAH	1zga	2.35	17	43
297	GSH	1zgn	2.10	64	29
298	SUC	1zs2	2.16	47	30
299	PRP	1zvw	2.30	33	82
300	BI5	1zyj	2.00	28	16
301	BOG	1zyj	2.00	21	30
302	DP9	1zzt	2.14	81	289
303	AMP	2a1u	2.11	47	100
304	GTP	2a5f	2.02	75	74
305	AUP	2aaq	2.60	28	67
306	1CA	2abi	2.33	59	16
307	CBC	2abj	2.20	166	192
308	SAM	2adm	2.60	19	60
309	UDH	2aec	2.00	234	90
310	OLA	2af9	2.00	4	2
311	UDH	2ah9	2.00	241	133
312	SAH	2an3	2.20	155	52
313	SAH	2an4	2.20	160	40
314	VCA	2awh	2.00	54	112
315	ADP	2axn	2.10	88	44
316	EDT	2axn	2.10	40	146
317	MA5	2azn	2.70	106	209
318	201	2b0m	2.00	30	103
319	FMN	2b0m	2.00	57	24
320	TRE	2b1q	2.20	29	45
321	CBI	2b1r	2.20	29	28
322	VCA	2b50	2.00	41	91
323	12P	2b9x	2.22	24	49
324	ODD	2bab	2.00	44	92
325	CM5	2bdm	2.30	42	39
326	TMI	2bdm	2.30	39	56
327	R22	2be2	2.43	66	67
328	SUC	2be2	2.43	95	70
329	CTP	2be9	2.60	55	102
330	IID	2bq7	2.20	17	35

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
331	DFW	2brh	2.10	25	50
332	NAR	2brt	2.20	11	19
333	AZQ	2bxi	2.50	47	78
334	IMN	2bxk	2.40	36	129
335	IMN	2bxm	2.50	69	117
336	P1Z	2bxp	2.30	44	133
337	SIA	2c4a	2.15	94	27
338	AMP	2c5s	2.50	58	26
339	PXI	2c6h	2.35	85	149
340	GTX	2c80	2.30	70	25
341	PHR	2c9d	2.80	79	22
342	QUE	2c9z	2.10	38	57
343	UDP	2c9z	2.10	45	38
344	GSW	2ca8	2.49	60	28
345	GSW	2caq	2.00	40	37
346	OAN	2cbj	2.35	106	74
347	N8T	2cek	2.20	59	78
348	DAN	2cex	2.20	45	47
349	P2N	2cgf	2.20	48	31
350	ADP	2cgj	2.26	34	23
351	3A3	2cgu	2.50	25	21
352	RCL	2cm9	2.30	24	14
353	F11	2cmf	2.50	52	68
354	ZMR	2cml	2.15	175	177
355	GDP	2cvw	2.40	52	87
356	TTP	2cvw	2.40	85	51
357	ADP	2cvx	2.20	44	88
358	DGT	2cvx	2.20	88	32
359	TRE	2cy6	2.00	18	9
360	A3P	2d06	2.30	158	51
361	EST	2d06	2.30	77	129
362	PQQ	2d0v	2.49	71	49
363	ATP	2d1k	2.50	77	38
364	AMP	2d1q	2.30	61	28
365	ANP	2d32	2.40	281	76
366	TNR	2d3s	2.35	179	124
367	UDP	2d7i	2.50	88	38
368	HRB	2ddq	2.35	16	15
369	FMN	2dor	2.00	168	94
370	ADP	2dpy	2.40	21	26
371	SAH	2dsg	2.00	42	20
372	SAH	2dsh	2.00	54	23
373	SAH	2dv4	2.20	40	23
374	GM6	2dw1	2.50	58	49
375	L2C	2dwe	2.50	15	55
376	ADP	2dwo	2.25	66	110

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
377	JTP	2dxs	2.20	34	16
378	SAH	2e16	2.00	40	17
379	SAH	2e4r	2.20	41	52
380	MTA	2e5w	2.00	62	211
381	SAH	2e8r	2.00	56	16
382	TRE	2ebh	2.40	35	67
383	SAH	2ed3	2.50	27	18
384	SAH	2ed5	2.10	60	28
385	SAH	2eg5	2.20	89	75
386	SAH	2emu	2.20	31	23
387	HFS	2erz	2.20	19	60
388	OLA	2ev2	2.35	25	23
389	OLA	2ev4	2.28	37	18
390	BOG	2evu	2.30	19	58
391	Y12	2ew5	2.20	4	2
392	Y13	2ew6	2.20	3	9
393	TPP	2ez4	2.03	130	175
394	TDM	2ezt	2.29	117	134
395	HTL	2ezu	2.16	90	119
396	DYM	2f13	2.26	21	36
397	AMP	2f17	2.50	27	80
398	C8E	2f1c	2.30	19	50
399	DAN	2f27	2.15	150	225
400	FBP	2f48	2.11	88	33
401	20S	2fah	2.09	103	223
402	GDP	2fah	2.09	221	123
403	SAM	2fb2	2.25	101	242
404	FSI	2fhr	2.20	37	19
405	S14	2fjp	2.40	118	164
406	GSP	2fju	2.20	88	35
407	3QC	2fme	2.10	94	164
408	ADP	2fme	2.10	125	89
409	JPA	2foi	2.50	86	93
410	OLA	2ftb	2.00	10	6
411	5IG	2g1y	2.50	125	112
412	TDK	2g25	2.10	100	26
413	ACF	2g5t	2.30	63	130
414	AAF	2g63	2.00	47	130
415	SAM	2g70	2.40	144	116
416	FTS	2g71	2.20	99	174
417	SAH	2g71	2.20	167	84
418	F21	2g72	2.00	175	176
419	SAM	2g72	2.00	144	165
420	UMP	2g86	2.40	51	30
421	F83	2g8n	2.15	171	166
422	SAH	2g8n	2.15	160	167

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
423	FMN	2gah	2.00	47	122
424	885	2gc8	2.20	62	20
425	DOI	2gcq	2.00	56	29
426	GDP	2gcq	2.00	65	24
427	BDE	2glp	2.42	31	103
428	ADP	2gm1	2.30	230	136
429	796	2gu8	2.20	44	33
430	NMN	2gvq	2.20	241	71
431	SFG	2h2j	2.45	227	74
432	APR	2h59	1.90	76	72
433	B3H	2h8p	2.25	30	100
434	N5A	2hch	2.30	78	84
435	BOG	2hd0	2.28	67	44
436	DMU	2hd0	2.28	25	34
437	B3H	2hfe	2.25	36	23
438	3TP	2hha	2.35	123	189
439	ATR	2hk9	2.20	62	172
440	SAH	2hnk	2.30	38	114
441	1CN	2hoc	2.10	43	27
442	EA5	2hp1	2.08	46	56
443	G39	2ht8	2.40	13	8
444	SAH	2huq	2.20	53	73
445	SAH	2huv	2.10	66	49
446	ANP	2hw1	2.10	8	30
447	4HX	2hx4	2.15	63	219
448	3CM	2hza	2.10	14	10
449	ADP	2if8	2.40	35	40
450	1EM	2ih1	2.40	24	37
451	872	2iit	2.35	120	199
452	565	2iiv	2.15	111	283
453	MHF	2ijg	2.10	72	85
454	CTP	2im0	2.25	25	32
455	ACJ	2ivd	2.30	61	152
456	ADP	2iw3	2.40	98	25
457	QQ2	2iw6	2.30	54	17
458	C5P	2ix0	2.44	36	21
459	LAT	2iy8	2.50	35	64
460	ADP	2iyz	2.30	22	11
461	FMN	2j09	2.00	56	56
462	ACP	2j4j	2.10	281	106
463	UDP	2j65	2.20	47	13
464	XMM	2jdz	2.10	17	26
465	ADP	2jgv	2.00	46	19
466	895	2jh5	2.50	27	18
467	SIA	2jh7	2.07	62	33
468	SIA	2jhd	2.30	58	30

No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
469	ADP	2ji6	2.06	77	212
470	ADP	2ji9	2.20	100	189
471	ADP	2jib	2.20	111	226
472	TPP	2jib	2.20	174	175
473	ADP	2no9	2.15	101	78
474	521	2nta	2.10	19	14
475	738	2o2u	2.45	55	17
476	TCB	2o9r	2.30	23	67
477	PE5	2oa5	2.10	34	16
478	F83	2obf	2.30	128	140
479	SAH	2obf	2.30	137	122
480	SAM	2obv	2.05	55	30
481	U1N	2ogz	2.10	62	86
482	ATP	2oh6	2.10	13	15
483	CTP	2oh6	2.10	21	16
484	ATP	2oh7	2.45	17	12
485	GTP	2oh7	2.45	12	16
486	8IP	2ohr	2.25	39	24
487	IP7	2ohu	2.35	16	31
488	UD1	2oi5	2.25	134	94
489	UD1	2oi6	2.20	157	40
490	19A	2ojg	2.00	33	18
491	277	2oph	2.40	115	248
492	DXC	2opx	2.53	58	22
493	SAH	2owf	2.20	24	13
494	SAH	2owg	2.10	38	47
495	SAH	2oy0	2.80	63	16
496	PRX	2p2b	2.20	145	96
497	SAH	2p5c	2.40	44	25
498	SAH	2pb4	2.10	36	24
499	MGT	2px8	2.20	93	74
500	SAH	2px8	2.20	93	127
501	GTP	2pxa	2.30	74	73
502	SAH	2pxa	2.30	84	84
503	U5P	2q0f	2.40	31	52
504	ADP	2q2r	2.10	50	97
505	ATP	2q36	2.50	84	56
506	EIC	2q9s	2.30	3	10
507	BNG	2qks	2.20	19	55
508	ADP	2qrd	2.41	88	138
509	ACP	2r7k	2.10	88	60
510	AMZ	2r7k	2.10	64	88
511	AMZ	2r7l	2.10	71	64
512	ATP	2r7l	2.10	64	68
513	AMP	2r7m	2.30	120	54
514	ADP	2r7n	2.40	69	70



No	Ligand	PDB-Id	Res [Å]	# Total $\rho$ -clusters	# Assoc. $\rho$ -clusters
515	ATP	2r86	2.50	144	111
516	C8E	2sqc	2.00	73	57
517	PLS	2trs	2.04	84	44
518	ADP	2ukd	2.20	66	72
519	C5P	2ukd	2.20	80	61
520	AD0	2uvf	2.10	18	72
521	ADP	2uyi	2.10	134	89
522	ADP	2uym	2.11	125	113
523	CDM	2v2z	2.25	128	44
524	SIA	2v73	2.20	38	18
525	ADP	2vb6	2.30	87	26
526	FMN	2vbv	2.40	38	40
527	2SA	2vd6	2.00	142	39
528	AMP	2vd6	2.00	69	166
529	ADP	2z0h	2.10	91	59
530	TYD	2z0h	2.10	91	67
531	GDP	2z1m	2.00	75	32
532	FMN	3b6j	2.05	143	120
533	MTA	3b7p	2.00	255	101
534	IM2	3bfc	2.20	84	105
535	ADP	4pfk	2.40	54	23
536	GTT	7gss	2.20	30	17

The table shows from left to right the ligand's PDB three letter code, the PDB-Id of the protein structure from which the ligand was extracted, the resolution of the protein's X-ray data, the total number of density ( $\rho$ ) clusters after fragmentation tree filtering, the total number of redundant density ( $\rho$ ) clusters found at the same location of the ligand in the difference electron density map. The ligand molecules are alphabetically sorted according to the PDB-Id of their protein structure. If two ligands originate from the same PDB file, they are additionally sorted according to their PDB three letter code.

## References

Aishima, J., Russel, D.S., Guibas, L.J., Adams, P.D. and Brunger, A.T. (2005) "Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface", *Acta crystallographica*, **61**, 1354-1363.

Albe, K.R., Butler, M.H. and Wright, B.E. (1990) "Cellular concentrations of enzymes and their substrates", *Journal of theoretical biology*, **143**, 163-195.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994a) "Ion Concentrations Can Be Measured with Intracellular Electrodes". In, *Molecular biology of the cell*. Garland Publishing, New York, 181-182.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994b) *Molecular biology of the cell*. Garland Publishing, New York.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994c) "The Breakdown of an Organic Molecule Takes Place in a Sequence of Enzyme-catalysed Reaction". In, *Molecular biology of the cell*. Garland Publishing, New York, 63-64.

An, J., Totrov, M. and Abagyan, R. (2005) "Pocketome via comprehensive identification and classification of ligand binding envelopes", *Mol Cell Proteomics*, **4**, 752-761.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S. (2004) "UniProt: the Universal Protein Knowledgebase", *Nucleic Acids Res.*, **32**, 115.

Armstrong, R.N., Kondo, H., Granot, J., Kaiser, E.T. and Mildvan, A.S. (1979) "Magnetic resonance and kinetic studies of the manganese(II) ion and substrate complexes of the catalytic subunit of adenosine 3',5'-monophosphate dependent protein kinase from bovine heart", *Biochemistry*, **18**, 1230-1238.

Auer, M. (2000) "Three-dimensional electron cryo-microscopy as a powerful structural tool in molecular medicine", *J Mol Med*, **78**, 191-202.

Babine, R.E. and Bender, S.L. (1997) "Molecular Recognition of Protein-Ligand Complexes: Applications to Drug Design", *Chemical reviews*, **97**, 1359-1472.

Bairoch, A. (2000) "The ENZYME database in 2000", *Nucleic Acids Res*, **28**, 304-305.

Baker, D. and Sali, A. (2001) "Protein structure prediction and structural genomics", *Science*, **294**, 93-96.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J. and McCammon, J.A. (2001) "Electrostatics of nanosystems: application to microtubules and the ribosome", *Proc Natl Acad Sci U S A*, **98**, 10037-10041.

Ball, P. (2008) "Water as an active constituent in cell biology", *Chemical reviews*, **108**, 74-108.

- Ballester, P.J. and Richards, W.G. (2007) "Ultrafast shape recognition to search compound databases for similar molecular shapes", *J Comput Chem*, **28**, 1711-1723.
- Barber, C.B., Dobkin, D.P. and Huhdanpaa, H. (1996) "The Quickhull algorithm for convex hulls", *ACM Transactions on Mathematical Software*, **22**, 469-483.
- Barker, J.A. and Thornton, J.M. (2003) "An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis", *Bioinformatics*, **19**, 1644-1649.
- Barrett, A.J. (1997) "Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997)", *European journal of biochemistry / FEBS*, **250**, 1-6.
- Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M. (2002) "Analysis of catalytic residues in enzyme active sites", *J Mol Biol*, **324**, 105-121.
- Barton, G.J. (1992) "Detecting Structural Similarity from Protein Sequence Comparison". In Dodson, E., Gover, S. and Wolf, W. (eds), *Molecular Replacement, Proceedings of the CCP4 Study Weekend, 31 January-1 February 1992*. Daresbury: Science and Engineering Research Council, 1-8.
- Bashford, D. and Karplus, M. (1990) "pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model", *Biochemistry*, **29**, 10219-10225.
- Bashton, M., Nobeli, I. and Thornton, J.M. (2006) "Cognate ligand domain mapping for enzymes", *J Mol Biol*, **364**, 836-852.
- Basu, G., Sivanesan, D., Kawabata, T. and Go, N. (2004) "Electrostatic potential of nucleotide-free protein is sufficient for discrimination between adenine and guanine-specific binding sites", *J Mol Biol*, **342**, 1053-1066.
- Bate, P. and Warwicker, J. (2004) "Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods", *J Mol Biol*, **340**, 263-276.
- Bell, C.E. and Eisenberg, D. (1996) "Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide", *Biochemistry*, **35**, 1137-1149.
- Benkovic, S.J. and Hammes-Schiffer, S. (2003) "A perspective on enzyme catalysis", *Science*, **301**, 1196-1202.
- Bergner, A. and Günther, J. (2004) "4.2 Structural Biology of Binding Sites". In Kubinyi, H. and Müller, G. (eds), *Chemogenomics in Drug Discovery*. WILEY-VCH Verlag, Weinheim, 101-102.
- Berman, H., Henrick, K. and Nakamura, H. (2003) "Announcing the worldwide Protein Data Bank", *Nature structural biology*, **10**, 980.
- Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007) "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data", *Nucleic Acids Res*, **35**, D301-303.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) "The Protein Data Bank", *Nucleic Acids Res*, **28**, 235-242.
- Besl, P.J. and McKay, N.D. (1992) "A Method for Registration of 3-D Shapes", *IEEE Transactions on PAMI*, **14**, 239-256.

- Bilwes, A.M., Quezada, C.M., Croal, L.R., Crane, B.R. and Simon, M.I. (2001) "Nucleotide binding by the histidine kinase CheA", *Nature structural biology*, **8**, 353-360.
- Binkowski, T.A., Adamian, L. and Liang, J. (2003a) "Inferring functional relationships of proteins from local sequence and spatial surface patterns", *J. Mol. Biol.*, **332**, 505-526.
- Binkowski, T.A. and Joachimiak, A. (2008) "Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites", *BMC Struct Biol*, **8**, 45.
- Binkowski, T.A., Naghibzadeh, S. and Liang, J. (2003b) "CASTp: Computed Atlas of Surface Topography of proteins", *Nucleic Acids Res*, **31**, 3352-3355.
- Birktoft, J.J., Rhodes, G. and Banaszak, L.J. (1989) "Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-Å resolution", *Biochemistry*, **28**, 6065-6081.
- Blundell, T.L. and Johnson, L.N. (1976) *Protein crystallography*. Academic Press, New York ; London.
- Blundell, T.L. and Mizuguchi, K. (2000) "Structural genomics: an overview", *Prog Biophys Mol Biol*, **73**, 289-295.
- Blundell, T.L., Sibanda, B.L., Sternberg, M.J. and Thornton, J.M. (1987) "Knowledge-based prediction of protein structures and the design of novel molecules", *Nature*, **326**, 347-352.
- Bock, M.E., Garutti, C. and Guerra, C. (2007) "Discovery of similar regions on protein surfaces", *J Comput Biol*, **14**, 285-299.
- Boehm, H.J. and Klebe, G. (1996) "What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs?", *Angewandte Chemie International Edition in English*, **35**, 2588-2614.
- Bogan, M.J., Benner, W.H., Boutet, S., Rohner, U., Frank, M., Barty, A., Seibert, M.M., Maia, F., Marchesini, S., Bajt, S., Woods, B., Riot, V., Hau-Riege, S.P., Svenda, M., Marklund, E., Spiller, E., Hajdu, J. and Chapman, H.N. (2008) "Single particle X-ray diffractive imaging", *Nano Lett*, **8**, 310-316.
- Bolon, D.N., Voigt, C.A. and Mayo, S.L. (2002) "De novo design of biocatalysts", *Current opinion in chemical biology*, **6**, 125-129.
- Brakoulias, A. and Jackson, R.M. (2004) "Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching", *Proteins*, **56**, 250-260.
- Branden, C. and Tooze, J. (1999) *Introduction to protein structure*. Garland Pub, New York.
- Brenner, S.E. (2001) "A tour of structural genomics", *Nature reviews*, **2**, 801-809.
- Bron, C. and Kerbosch, J. (1973) "Finding all cliques of an undirected graph", *Communications of the ACM*, **16**, 575-577.
- Brucoleri, R.E., Novotny, J. and Davis, M.E. (1997) "Finite difference Poisson-Boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing", *J. Comput. Chem.*, **18**, 268-276.
- Brunger, A.T. (1992) "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures", *Nature*, **355**, 472-475.

- Brunger, A.T., Kuriyan, J. and Karplus, M. (1987) "Crystallographic R Factor Refinement by Molecular Dynamics", *Science*, **235**, 458-460.
- Brylinski, M., Konieczny, L. and Roterman, I. (2006) "Ligation site in proteins recognized in silico", *Bioinformatics*, **1**, 127-129.
- Cai, W., Shao, X. and Maigret, B. (2002a) "Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening", *J Mol Graph Model*, **20**, 313-328.
- Cai, W., Shao, X. and Maigret, B. (2002b) "Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast efficient filter for large virtual throughput screening", *J.Mol.Graphics and Modelling*, **20**.
- Campbell, S.J., Gold, N.D., Jackson, R.M. and Westhead, D.R. (2003) "Ligand binding: functional site location, similarity and docking", *Curr Opin Struct Biol*, **13**, 389-395.
- Carvin, D., Islam, S.A., Sternberg, M.J.E. and Blundell, T.L. (2006) "12.1. The preparation of heavy-atom derivatives of protein crystals for use in multiple isomorphous replacement and anomalous scattering". In Rossmann, M.G. and Arnold, E. (eds), *International Tables of Crystallography*. Kluwer Academic Publishers, Dordrecht, 247-255.
- Cauet, E., Rooman, M., Wintjens, R., Lievin, J. and Biot, C. (2005) "Histidine-aromatic interactions in proteins and protein-ligand complexes: Quantum chemical study of X-ray and model structures", *J Chem Theory Comput*, **1**, 472-483.
- Chandonia, J.M. and Brenner, S.E. (2006) "The impact of structural genomics: expectations and outcomes", *Science*, **311**, 347-351.
- Chau, P.L. and Dean, P.M. (1994a) "Electrostatic complementarity between proteins and ligands. 1. Charge disposition, dielectric and interface effects", *Journal of computer-aided molecular design*, **8**, 513-525.
- Chau, P.L. and Dean, P.M. (1994b) "Electrostatic complementarity between proteins and ligands. 2. Ligand moieties", *Journal of computer-aided molecular design*, **8**, 527-544.
- Chau, P.L. and Dean, P.M. (1994c) "Electrostatic complementarity between proteins and ligands. 3. Structural basis", *Journal of computer-aided molecular design*, **8**, 545-564.
- Chelliah, V., Chen, L., Blundell, T.L. and Lovell, S.C. (2004) "Distinguishing structural and functional restraints in evolution in order to identify interaction sites", *J Mol Biol*, **342**, 1487-1504.
- Chothia, C. and Lesk, A.M. (1986) "The relation between the divergence of sequence and structure in proteins", *The EMBO journal*, **5**, 823-826.
- Cohen, S.X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T.K., Lamzin, V.S., Murshudov, G.N. and Perrakis, A. (2008) "ARP/wARP and molecular replacement: the next generation", *Acta crystallographica*, **64**, 49-60.
- Coleman, R.G. and Sharp, K.A. (2006) "Travel depth, a new shape descriptor for macromolecules: Application to ligand binding", *Journal of Molecular Biology*, **362**, 441-458.
- Connolly, M.L. (1983a) "Analytical Molecular-Surface Calculation", *Journal of Applied Crystallography*, **16**, 548-558.
- Connolly, M.L. (1983b) "Solvent-accessible surfaces of proteins and nucleic acids", *Science*, **221**, 709-713.

- Copley, S.D. (2003) "Enzymes with extra talents: moonlighting functions and catalytic promiscuity", *Current opinion in chemical biology*, **7**, 265-272.
- Cosgrove, D.A., Bayada, D.M. and Johnson, A.P. (2000) "A novel method of aligning molecules by local surface shape similarity", *Journal of computer-aided molecular design*, **14**, 573-591.
- Cowtan, K. (2006) "The Buccaneer software for automated model building. 1. Tracing protein chains", *Acta crystallographica*, **62**, 1002-1011.
- Davis, A.M. and Teague, S.J. (1999) "Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis", *Angew. Chem. Int. Ed*, **38**, 736-749.
- DeLano, W.L. (2002) "The PyMOL Molecular Graphics System". DeLano Scientific, San Carlos, CA.
- Denessiouk, K.A., Rantanen, V.V. and Johnson, M.S. (2001) "Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins", *Proteins*, **44**, 282-291.
- DePristo, M.A., de Bakker, P.I. and Blundell, T.L. (2004) "Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography", *Structure*, **12**, 831-838.
- Diller, D.J., Pohl, E., Redinbo, M.R., Hovey, B.T. and Hol, W.G. (1999) "A rapid method for positioning small flexible molecules, nucleic acids, and large protein fragments in experimental electron density maps", *Proteins*, **36**, 512-525.
- Dlugosz, M. and Trylska, J. (2008) "Electrostatic similarity of proteins: application of three dimensional spherical harmonic decomposition", *J Chem Phys*, **129**, 015103.
- Dodson, G. and Verma, C.S. (2006) "Protein flexibility: its role in structure and mechanism revealed by molecular simulations", *Cell Mol Life Sci*, **63**, 207-219.
- Dolinsky, T.J., Nielsen, J.E., McCammon, J.A. and Baker, N.A. (2004) "PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations", *Nucleic Acids Research*, **32**, W665-W667.
- Domingues, F.S., Rahnenfuhrer, J. and Lengauer, T. (2004) "Automated clustering of ensembles of alternative models in protein structure databases", *Protein Eng Des Sel*, **17**, 537-543.
- Dunbrack, R.L.J. (2002) "Homology Modeling in Biology and Medicine". In Lengauer, T. (ed), *Bioinformatics : from genomes to drugs, Methods and principles in medicinal chemistry*. Wiley-VCH, Weinheim ; [Great Britain], 145-149.
- Duncan, B.S. and Olson, A.J. (1993) "Approximation and characterization of molecular surfaces", *Biopolymers*, **33**, 219-229.
- Dwyer, M.A., Looger, L.L. and Hellinga, H.W. (2004) "Computational design of a biologically active enzyme", *Science*, **304**, 1967-1971.
- Eisenberg, D. and McLachlan, A.D. (1986) "Solvation energy in protein folding and binding", *Nature*, **319**, 199-203.
- Ellis, R.J. (2001) "Macromolecular crowding: an important but neglected aspect of the intracellular environment", *Curr Opin Struct Biol*, **11**, 114-119.
- Evrard, G.X., Langer, G.G., Perrakis, A. and Lamzin, V.S. (2007) "Assessment of automatic ligand building in ARP/wARP", *Acta crystallographica*, **63**, 108-117.

- Exner, T.E., Keil, M. and Brickmann, J. (2002) "Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory", *J Comput Chem*, **23**, 1176-1187.
- Fan, H. and Mark, A.E. (2003) "Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: a molecular dynamics simulation study", *Proteins*, **53**, 111-120.
- Fauchère, J.L. and Pliska, V. (1983) "Hydrophobic Parameters- $\Pi$  of Amino-Acid Side-Chains from the Partitioning of N-Acetyl-Amino-Acid Amides", *European Journal of Medicinal Chemistry*, **18**, 369-375.
- Fauchère, J.L., Quarendon, P. and Kaetterer, L. (1988) "Estimating and Representing Hydrophobicity Potential", *Journal of molecular graphics*, **6**, 203-&.
- Fersht, A. (1984) "The three-dimensional structure of enzymes". In, *Enzyme Structure and Mechanism*. W.H. Freeman and Company, New York, 1-6.
- Fersht, A. (1999) "1. The hydrophobic bond". In, *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. W.H. Freeman, New York ; Basingstoke, 332-333.
- Fersht, A.R. (1974) "Catalysis, binding and enzyme-substrate complementarity", *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character*, **187**, 397-407.
- Feynman, R.P., Leighton, B.L. and Sands, M.L. (1989a) "6-1 Equations of the electrostatic potential". In, *Feynman lectures on physics*. Addison-Wesley, Massachusetts, 6-1.
- Feynman, R.P., Leighton, B.L. and Sands, M.L. (1989b) *Feynman lectures on physics*. Addison-Wesley, Massachusetts.
- Fischer, E. (1894) "Einfluss der Configuration auf die Wirkung der Enzyme", *Ber. Dtsch. Chem. Ges*, **27**, 2985-2993.
- Frank, J. (2002) "Single-particle imaging of macromolecules by cryo-electron microscopy", *Annual review of biophysics and biomolecular structure*, **31**, 303-319.
- Gabb, H.A., Jackson, R.M. and Sternberg, M.J. (1997) "Modelling protein docking using shape complementarity, electrostatics and biochemical information", *J Mol Biol*, **272**, 106-120.
- Garini, Y., Vermolen, B.J. and Young, I.T. (2005) "From micro to nano: recent advances in high-resolution microscopy", *Curr Opin Biotechnol*, **16**, 3-12.
- Garman, E. (1999) "Cool data: quantity AND quality", *Acta crystallographica*, **55**, 1641-1653.
- Gerczei, T., Asboth, B. and Naray-Szabo, G. (1999) "Conservative electrostatic potential patterns at enzyme active sites: The anion-cation-anion triad", *J. Chem. Inf. Comput. Sci.*, **39**, 310-315.
- Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J. (1998) "Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods", *Journal of Physical Chemistry A*, **102**, 3762-3772.
- Gilson, M.K. (2000) "Introduction to Continuum Electrostatics, with Molecular Applications", *Biophysics Textbooks online*.
- Gilson, M.K., Rashin, A., Fine, R. and Honig, B. (1985) "On the calculation of electrostatic interactions in proteins", *Journal of Molecular Biology*, **184**, 503.

- Gilson, M.K. and Zhou, H.X. (2007) "Calculation of protein-ligand binding affinities", *Annual review of biophysics and biomolecular structure*, **36**, 21-42.
- Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. (2006) "A method for localizing ligand binding pockets in protein structures", *Proteins*, **62**, 479-488.
- Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information", *Bioinformatics*, **19**, 163-164.
- Gold, N.D. and Jackson, R.M. (2006a) "Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships", *J Mol Biol.*, **355**, 1112-1124.
- Gold, N.D. and Jackson, R.M. (2006b) "SitesBase: a database for structure-based protein-ligand binding site comparisons", *Nucleic Acids Res*, **34**, D231-234.
- Golovin, A., Oldfield, T.J., Tate, J.G., Velankar, S., Barton, G.J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J.M.C., John, M., Keller, P.A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tromm, S., Vranken, W. and Henrick, K. (2004) "MSD database and MSD database services", *E-MSD: an integrated data Nucleic Acids Research*, **32 (Database issue)**, 211.
- Gonzalez, B., Pajares, M.A., Hermoso, J.A., Guillerm, D., Guillerm, G. and Sanz-Aparicio, J. (2003) "Crystal structures of methionine adenosyltransferase complexed with substrates and products reveal the methionine-ATP recognition and give insights into the catalytic mechanism", *J Mol Biol*, **331**, 407-416.
- Goodford, P.J. (1985) "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules", *Journal of medicinal chemistry*, **28**, 849-857.
- Grant, J.A., Gallardo, M.A. and Pickup, B.T. (1996) "A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape", *J. Comput. Chem.*, **17**, 1653-1666.
- Grant, J.A. and Pickup, B.T. (1995) "A Gaussian Description of Molecular Shape", *J. Phys. Chem.*, **99**, 3503-3510.
- Gruber, J., Zawaira, A., Saunders, R., Barrett, C.P. and Noble, M.E. (2007) "Computational analyses of the surface properties of protein-protein interfaces", *Acta crystallographica*, **63**, 50-57.
- Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E. (2006) "The Blue Obelisk—interoperability in chemical informatics", *J. Chem. Inf. Model*, **46**, 991-998.
- Gunasekaran, K. and Nussinov, R. (2007) "How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding", *J Mol Biol*, **365**, 257-273.
- Gunner, M.R. and Honig, B. (1991) "Electrostatic control of midpoint potentials in the cytochrome subunit of the Rhodospseudomonas viridis reaction center", *Proc Natl Acad Sci U S A*, **88**, 9151-9155.
- Gutteridge, A. and Thornton, J. (2005) "Conformational changes observed in enzyme crystal structures upon substrate binding", *J Mol Biol*, **346**, 21-28.



- Halgren, T.A. (1992) "Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and vdW Parameters", *Journal of the American Chemical Society*, **114**, 7827-7843.
- Halgren, T.A. (1996) "Merck molecular force field .2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions", *J. Comput. Chem.*, **17**, 520-552.
- Hanley, J.A. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, **143**, 29-36.
- Hanley, J.A. (1983) "A method of comparing the areas under receiver operating characteristic curves derived from the same cases", *Radiology*, **148**, 839-843.
- Hardin, R.H. and Sloane, N.J.A. (1996) "McLaren's improved snub cube and other new spherical designs in three dimensions", *Discrete & Computational Geometry*, **15**, 429-441.
- Heiden, W., Moeckel, G. and Brickmann, J. (1993) "A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces", *Journal of computer-aided molecular design*, **7**, 503-514.
- Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M. and Berman, H.M. (2008) "Remediation of the protein data bank archive", *Nucleic Acids Res*, **36**, D426-433.
- Henrick, K. and Thornton, J.M. (1998) "PQS: a protein quaternary structure file server", *Trends Biochem Sci*, **23**, 358-361.
- Hirsch, A.K., Fischer, F.R. and Diederich, F. (2007) "Phosphate recognition in structural biology", *Angewandte Chemie (International ed)*, **46**, 338-352.
- Hofbauer, C., Lohninger, H. and Aszodi, A. (2004) "SURFCOMP: a novel graph-based approach to molecular surface comparison", *J Chem Inf Comput Sci*, **44**, 837-847.
- Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'Boyle, N.M., Murray-Rust, P., Thornton, J.M. and Mitchell, J.B. (2005) "MACiE: a database of enzyme reaction mechanisms", *Bioinformatics*, **21**, 4315-4316.
- Honig, B. and Nicholls, A. (1995) "Classical electrostatics in biology and chemistry", *Science*, **268**, 1144.
- Hope, H. (1990) "Crystallography of biological macromolecules at ultra-low temperature", *Annu Rev Biophys Biophys Chem*, **19**, 107-126.
- Hunter, C.A., Lawson, K.R., Perkins, J. and Urch, C.J. (2001) "Aromatic interactions", *J Chem Soc Perk T 2*, 651-669.
- Israelachvili, J. (1991a) *Intermolecular and surface forces*. Academic Press, London.
- Israelachvili, J.N. (1991b) "1.5 Modern View of the Origin of Intermolecular Forces". In, *Intermolecular and surface forces*. Academic, London, 11.
- Israelachvili, J.N. (1991c) "2.1 Interaction Energies of Molecules in Free Space and in a Medium". In, *Intermolecular and surface forces*. Academic, London, 16-20.
- Israelachvili, J.N. (1991d) "2.6 Classification of Forces". In, *Intermolecular and surface forces*. Academic, London, 27-28.

- Israelachvili, J.N. (1991e) "8. Special Interactions: Hydrogen-Bonding, Hydrophobic and Hydrophilic Interactions". In, *Intermolecular and surface forces*. Academic, London, 122-136.
- Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y. and Ramani, K. (2005) "Three-dimensional shape searching: state-of-the-art review and future trends", *Comput.-Aided Des.*, **37**, 509-530.
- Jacobson, M.P., Friesner, R.A., Xiang, Z. and Honig, B. (2002) "On the role of the crystal environment in determining protein side-chain conformations", *J Mol Biol*, **320**, 597-608.
- Jain, A.N. (2006) "Scoring functions for protein-ligand docking", *Current protein & peptide science*, **7**, 407-420.
- James, L.C., Roversi, P. and Tawfik, D.S. (2003) "Antibody multispecificity mediated by conformational diversity", *Science*, **299**, 1362-1367.
- James, L.C. and Tawfik, D.S. (2003) "Conformational diversity and protein evolution--a 60-year-old hypothesis revisited", *Trends Biochem Sci*, **28**, 361-368.
- Janin, J. and Wodak, S. (2007) "The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007", *Structure*, **15**, 755-759.
- Jiang, F., Lin, W. and Rao, Z. (2002) "SOFTDOCK: understanding of molecular recognition through a systematic docking study", *Protein engineering*, **15**, 257-263.
- Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (1997) "Development and validation of a genetic algorithm for flexible docking", *J Mol Biol*, **267**, 727-748.
- Jones, S. and Thornton, J.M. (1996) "Principles of protein-protein interactions", *Proc Natl Acad Sci U S A*, **93**, 13-20.
- Jones, T.A., Zou, J.Y., Cowan, S.W. and Kjeldgaard, M. (1991) "Improved methods for building protein models in electron density maps and the location of errors in these models", *Acta Crystallogr A*, **47 ( Pt 2)**, 110-119.
- Kahraman, A., Morris, R.J., Laskowski, R.A., Favia, A.D. and Thornton, J.M. (2009) "On the Diversity of Physicochemical Environments Experienced by Identical Ligands in Binding Pockets of Unrelated Proteins", *Proteins* (under review).
- Kahraman, A., Morris, R.J., Laskowski, R.A. and Thornton, J.M. (2007a) "Shape variation in protein binding pockets and their ligands", *J Mol Biol*, **368**, 283-301.
- Kahraman, A., Morris, R.J., Laskowski, R.A. and Thornton, J.M. (2007b) "Variation of geometrical and physicochemical properties in protein binding pockets and their ligands", *BMC Bioinformatics*, **8**, S1.
- Kahraman, A. and Thornton, J.M. (2008) "Methods to Characterize the Structure of Enzyme Binding Sites". In Schwede, T. and Peitsch, M.C. (eds), *Computational Structural Biology - Methods and Applications* -. World Scientific Publishing Co., 189-221.
- Kangas, E. and Tidor, B. (2001) "Electrostatic complementarity at ligand binding sites: Application to chorismate mutase", *J. Phys. Chem. B*, **105**, 880-888.
- Karplus, M. and McCammon, J.A. (2002) "Molecular dynamics simulations of biomolecules", *Nature structural biology*, **9**, 646-652.

- Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992) "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques", *Proc Natl Acad Sci U S A*, **89**, 2195-2199.
- Kazhdan, M., Funkhouser, T. and Rusinkiewicz, S. (2003) "Rotation invariant spherical harmonic representation of 3D shape descriptors", *Proceedings of the Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 156-164.
- Kellogg, G.E., Semus, S.F. and Abraham, D.J. (1991) "HINT: a new method of empirical hydrophobic field calculation for CoMFA", *Journal of computer-aided molecular design*, **5**, 545-552.
- Khersonsky, O., Roodveldt, C. and Tawfik, D.S. (2006) "Enzyme promiscuity: evolutionary and mechanistic aspects", *Current opinion in chemical biology*, **10**, 498-508.
- Kinoshita, K., Furui, J. and Nakamura, H. (2002) "Identification of protein functions from a molecular surface database, eF-site", *Journal of structural and functional genomics*, **2**, 9-22.
- Kinoshita, K. and Nakamura, H. (2003) "Identification of protein biochemical functions by similarity search using the molecular surface database eF-site", *Protein Sci.*, **12**, 1589-1595.
- Kitchen, D.B., Decornez, H., Furr, J.R. and Bajorath, J. (2004) "Docking and scoring in virtual screening for drug discovery: methods and applications", *Nat Rev Drug Discov*, **3**, 935-949.
- Klapper, I., Hagstrom, R., Fine, R., Sharp, K. and Honig, B. (1986) "Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification", *Proteins*, **1**, 47-59.
- Kleywegt, G.J. (1999) "Recognition of spatial motifs in protein structures", *J. Mol. Biol.*, **285**, 1887-1897.
- Kleywegt, G.J. (2007) "Crystallographic refinement of ligand complexes", *Acta crystallographica*, **63**, 94-100.
- Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A. and Jones, T.A. (2004) "The Uppsala Electron-Density Server", *Acta crystallographica*, **60**, 2240-2249.
- Kleywegt, G.J. and Jones, T.A. (1997) "Template convolution to enhance or detect structural features in macromolecular electron-density maps", *Acta crystallographica*, **53**, 179-185.
- Koehl, P. (2006) "Electrostatics calculations: latest methodological advances", *Curr. Opin. Struct. Biol.*, **16**, 142-151.
- Kortemme, T., Morozov, A.V. and Baker, D. (2003) "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes", *J Mol Biol*, **326**, 1239-1259.
- Koshland, D.E. (1958) "Application of a Theory of Enzyme Specificity to Protein Synthesis", *Proc Natl Acad Sci U S A*, **44**, 98-104.
- Kraut, D.A., Sigala, P.A., Pybus, B., Liu, C.W., Ringe, D., Petsko, G.A. and Herschlag, D. (2006) "Testing electrostatic complementarity in enzyme catalysis: Hydrogen bonding in the ketosteroid isomerase oxyanion hole", *Plos Biology*, **4**, 501-519.
- Kreyszig, E. (1999a) "4.7 Sturm-Liouville Problems. Orthogonal Functions ". In, *Advanced Engineering Mathematics*. John Wiley & Sons, Singapore, 233-239.

- Kreyszig, E. (1999b) "4.8 Orthogonal Eigenfunction Expansion". In, *Advanced Engineering Mathematics*. John Wiley & Sons, Singapore, 233-239.
- Kreyszig, E. (1999c) "A3.1 Formulas for Special Functions". In, *Advanced Engineering Mathematics*. John Wiley & Sons, Singapore, A53.
- Krissinel, E. and Henrick, K. (2007) "Inference of macromolecular assemblies from crystalline state", *J Mol Biol*, **372**, 774-797.
- Kundu, S. and Gupta-Bhaya, P. (2004) "How a repulsive charge distribution becomes attractive and stabilized by a polarizable protein dielectric", *J Mol Struct-Theochem*, **668**, 65-73.
- Kurland, C.G., Collins, L.J. and Penny, D. (2006) "Genomics and the irreducible nature of eukaryote cells", *Science*, **312**, 1011-1014.
- Lacapere, J.J., Pebay-Peyroula, E., Neumann, J.M. and Etchebest, C. (2007) "Determining membrane protein structures: still a challenge!", *Trends Biochem Sci*, **32**, 259-270.
- Lamzin, V.S. and Wilson, K.S. (1993) "Automated refinement of protein models", *Acta crystallographica*, **49**, 129-147.
- Langer, G., Cohen, S.X., Lamzin, V.S. and Perrakis, A. (2008) "Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7", *Nat Protoc*, **3**, 1171-1179.
- Larsen, T.M., Benning, M.M., Rayment, I. and Reed, G.H. (1998) "Structure of the bis(Mg<sup>2+</sup>)-ATP-oxalate complex of the rabbit muscle pyruvate kinase at 2.1 Å resolution: ATP binding over a barrel", *Biochemistry*, **37**, 6247-6255.
- Laskowski, R. (2006) "Determining Function from Structure". In Sundström, M., Norin, M. and Edwards, A. (eds), *Structural Genomics and High Throughput Structural Biology*. CRC Press, Boca Raton.
- Laskowski, R.A. (1992) "Prediction, Analysis and Determination of Protein Structure, including applications of parallel computing". *Department of Crystallography*. University of London, Birkbeck College, London, 239.
- Laskowski, R.A. (1995) "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions", *Journal of molecular graphics*, **13**, 323-330, 307-328.
- Laskowski, R.A. (2003) "Structural Quality Assurance", *Structural Bioinformatics*, 273-303.
- Laskowski, R.A., Chistyakov, V.V. and Thornton, J.M. (2005) "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids", *Nucleic Acids Res*, **33**, 266.
- Laskowski, R.A., Luscombe, N.M., Swindells, M.B. and Thornton, J.M. (1996) "Protein clefts in molecular recognition and function", *Protein Sci*, **5**, 2438-2452.
- Laurie, A.T. and Jackson, R.M. (2005) "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites", *Bioinformatics*, **21**, 1908-1916.
- Leach, A. and Gillet, V. (2003a) "2. Cluster Analysis". In, *An Introduction to Chemoinformatics*. Springer, Dordrecht, The Netherlands, 120-128.
- Leach, A. and Gillet, V. (2003b) "3. Similarity Coefficients". In, *An Introduction to Chemoinformatics*. Springer, Dordrecht, The Netherlands, 102-105.

- Leach, A.R. (2001a) "4.10.1 Dispersive Interactions". In, *Molecular modelling : principles and applications*. Prentice Hall, Harlow, 204-206.
- Leach, A.R. (2001b) "9.4 Random Search Methods". In, *Molecular modelling : principles and applications*. Prentice Hall, Harlow, 465-466.
- Leach, A.R. (2001c) "11.10.4 Methods Based upon the Poisson-Boltzmann Equation". In, *Molecular modelling : principles and applications*. Prentice Hall, Harlow, 603-606.
- Ledvina, P.S., Yao, N., Choudhary, A. and Quioco, F.A. (1996) "Negative electrostatic surface potential of protein sites specific for anionic ligands", *Proc Natl Acad Sci U S A*, **93**, 6786-6791.
- Lee, B. and Richards, F.M. (1971) "Interpretation of Protein Structures - Estimation of Static Accessibility", *Journal of Molecular Biology*, **55**, 379-&.
- Leicester, S., Finney, J. and Bywater, R. (1994) "A Quantitative Representation of Molecular-Surface Shape. 1.Theory and Development of the Method", *Journal of Mathematical Chemistry*, **16**, 315-341.
- Levitt, M. (1976) "A simplified representation of protein conformations for rapid simulation of protein folding", *J Mol Biol*, **104**, 59-107.
- Levitt, M. and Lifson, S. (1969) "Refinement of protein conformations using a macromolecular energy minimization procedure", *J Mol Biol*, **46**, 269-279.
- Levitt, M. and Perutz, M.F. (1988) "Aromatic rings act as hydrogen bond acceptors", *J Mol Biol*, **201**, 751-754.
- Levy, Y. and Onuchic, J.N. (2006) "Water mediation in protein folding and molecular recognition", *Annual review of biophysics and biomolecular structure*, **35**, 389-415.
- Li, H., Robertson, A.D. and Jensen, J.H. (2005) "Very fast empirical prediction and rationalization of protein pKa values", *Proteins*, **61**, 704-721.
- Li, H.D. and Hartley, R. (2007) "Conformal spherical representation of 3D genus-zero meshes", *Pattern Recognit.*, **40**, 2742-2753.
- Liang, J., Edelsbrunner, H. and Woodward, C. (1998) "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design", *Protein Sci*, **7**, 1884-1897.
- Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) "An evolutionary trace method defines binding surfaces common to protein families", *J Mol Biol*, **257**, 342-358.
- Lin, J.H. and Clark, T. (2005) "An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties", *J Chem Inf Model*, **45**, 1010-1016.
- Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (2001) "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", *Advanced Drug Delivery Reviews*, **46**, 3-26.
- Livesay, D.R., Jambeck, P., Rojnuckarin, A. and Subramaniam, S. (2003) "Conservation of electrostatic properties within enzyme families and superfamilies", *Biochemistry*, **42**, 3464-3473.
- Luo, R., Gilson, H.S., Potter, M.J. and Gilson, M.K. (2001) "The physical basis of nucleic acid base stacking in water", *Biophysical journal*, **80**, 140-148.

- Lyne, P.D., Lamb, M.L. and Saeh, J.C. (2006) "Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring", *Journal of medicinal chemistry*, **49**, 4805-4808.
- Mackerell, A.D., Jr. (2004) "Empirical force fields for biological macromolecules: overview and issues", *J Comput Chem*, **25**, 1584-1604.
- Mak, L., Grandison, S. and Morris, R.J. (2007) "An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison", *J Mol Graph Model*.
- Mancera, R.L. (2007) "Molecular modeling of hydration in drug design", *Current opinion in drug discovery & development*, **10**, 275-280.
- Mansfield, M.L., Covell, D.G. and Jernigan, R.L. (2002) "A new class of molecular shape descriptors. 1. Theory and properties", *J Chem Inf Comput Sci*, **42**, 259-273.
- Mao, J., Hauser, K. and Gunner, M.R. (2003) "How cytochromes with different folds control heme redox potentials", *Biochemistry*, **42**, 9829-9840.
- Mao, L., Wang, Y., Liu, Y. and Hu, X. (2004) "Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis", *J Mol Biol*, **336**, 787-807.
- Marsili, S., Chelli, R., Schettino, V. and Procacci, P. (2008) "Thermodynamics of stacking interactions in proteins", *Phys Chem Chem Phys*, **10**, 2673-2685.
- Maslen, E.N., Fox, A.G. and O'Keefe, M.A. (2006) "6.1.1. X-ray scattering". In Prince, E. (ed), *International Tables for Crystallography*. Kluwer Academic Publishers, Dordrecht, 554-590.
- Masuda, S., Murakami, K.S., Wang, S., Anders Olson, C., Donigian, J., Leon, F., Darst, S.A. and Campbell, E.A. (2004) "Crystal structures of the ADP and ATP bound forms of the Bacillus anti-sigma factor SpoIIAB in complex with the anti-anti-sigma SpoIIAA", *J Mol Biol*, **340**, 941-956.
- Mathews, B. (1966) "The determination of the position of anomalously scattering heavy atom groups in protein crystals", *Acta Cryst.*, **20**, 230-239.
- Max, N.L. and Getzoff, E.D. (1988) "Spherical Harmonic Molecular-Surfaces", *Ieee Computer Graphics and Applications*, **8**, 42-50.
- McCoy, A.J., Chandana Epa, V. and Colman, P.M. (1997) "Electrostatic complementarity at protein/protein interfaces", *J Mol Biol*, **268**, 570-584.
- McDonald, I.K. and Thornton, J.M. (1994) "Satisfying hydrogen bonding potential in proteins", *J. Mol. Biol.*, **238**.
- McMahon, T. (1973) "Size and shape in biology", *Science*, **179**, 1201-1204.
- McPherson, A. (June 2000) "Crystallization of Proteins and Protein-Ligand Complexes". In Melino, G. (ed), *Encyclopedia of Life Sciences*. Macmillan Publishers Ltd, Nature Publishing Group, London, 209-213.
- McRee, D. (1992) "XtalView: a visual protein crystallographic software system for Xll/XView", *J. Mol. Graph*, **10**, 44-47.
- Meakin, P. (1998) "2.1 Self-Similar Fractals". In, *Fractals, scaling and growth far from equilibrium*. Cambridge University Press, Cambridge, 65-69.

- Meyer, B. and Peters, T. (2003) "NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors", *Angewandte Chemie (International ed)*, **42**, 864-890.
- Mezey, P.G. (1993) *Shape in chemistry : an introduction to molecular shape and topology*. VCH, New York, NY ; Cambridge.
- Miranker, A. and Karplus, M. (1991) "Functionality maps of binding sites: a multiple copy simultaneous search method", *Proteins*, **11**, 29-34.
- Miteva, M.A., Tuffery, P. and Villoutreix, B.O. (2005) "PCE: web tools to compute protein continuum electrostatics", *Nucleic Acids Res*, **33**, W372-375.
- Mitra, K. and Frank, J. (2006) "Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps", *Annual review of biophysics and biomolecular structure*, **35**, 299-317.
- Mooij, W.T., Hartshorn, M.J., Tickle, I.J., Sharff, A.J., Verdonk, M.L. and Jhoti, H. (2006) "Automated protein-ligand crystallography for structure-based drug design", *ChemMedChem*, **1**, 827-838.
- Morris, R.J. (2006) "An evaluation of spherical designs for molecular-like surfaces", *J Mol Graph Model*, **24**, 356-361.
- Morris, R.J., Najmanovich, R.J., Kahraman, A. and Thornton, J.M. (2005) "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons", *Bioinformatics*, **21**, 2347-2355.
- Morris, R.J., Perrakis, A. and Lamzin, V.S. (2007) "Getting a macromolecular model: model building, refinement, and validation". In Sanderson, M.R. and Skelly, J.V. (eds), *Macromolecular Crystallography, conventional and high-throughput methods*. Oxford University Press, USA, 153-168.
- Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) "Refinement of macromolecular structures by the maximum-likelihood method", *Acta crystallographica*, **53**, 240-255.
- Najmanovich, R.J., Allali-Hassani, A., Morris, R.J., Dombrovsky, L., Pan, P.W., Vedadi, M., Plotnikov, A.N., Edwards, A., Arrowsmith, C. and Thornton, J.M. (2007) "Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family", *Bioinformatics*, **23**, e104-109.
- Nakamura, H., Komatsu, K., Nakagawa, S. and Umeyama, H. (1985a) "Visualization of Electrostatic Recognition by Enzymes for Their Ligands and Cofactors", *Journal of molecular graphics*, **3**, 2-11.
- Nakamura, H., Komatsu, K. and Umeyama, H. (1985b) "Electrostatic Complementarities between Guest Ligands and Host Enzymes", *Journal of the Physical Society of Japan*, **54**, 3257-3260.
- Naray-Szabo, G. (1989) "Electrostatic complementarity in molecular associations", *Journal of molecular graphics*, **7**, 76-81, 98.
- Naray-Szabo, G. and Nagy, P. (1989) "Electrostatic complementarity in molecular aggregates. 9: Protein-ligand complexes", *International Journal of Quantum Chemistry*, **35**, 215-221.
- Navaza, J. (2006) "13.2. Rotation functions". In Rossmann, M. and Arnold, E. (eds), *International Tables for Crystallography*: . Kluwer Academic Publishers, Dordrecht, 269-274.

- Nayal, M. and Honig, B. (2006) "On the nature of cavities on protein surfaces: application to the identification of drug-binding sites", *Proteins*, **63**, 892-906.
- Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. and Hajdu, J. (2000) "Potential for biomolecular imaging with femtosecond X-ray pulses", *Nature*, **406**, 752-757.
- Nicholls, A., Sharp, K.A. and Honig, B. (1991) "Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons", *Proteins*, **11**, 281-296.
- Nicklaus, M.C., Wang, S., Driscoll, J.S. and Milne, G.W. (1995) "Conformational changes of small molecules binding to proteins", *Bioorg Med Chem*, **3**, 411-428.
- Nielsen, J.E. and McCammon, J.A. (2003) "Calculating pKa values in enzyme active sites", *Protein Science*, **12**, 1894-1901.
- Nielsen, J.E. and Vriend, G. (2001) "Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK(a) calculations", *Proteins*, **43**, 403-412.
- Nobeli, I., Ponstingl, H., Krissinel, E.B. and Thornton, J.M. (2003) "A structure-based anatomy of the E.coli metabolome", *J Mol Biol*, **334**, 697-719.
- Nooren, I.M. and Thornton, J.M. (2003) "Diversity of protein-protein interactions", *The EMBO journal*, **22**, 3486-3492.
- Norel, R., Petrey, D., Wolfson, H.J. and Nussinov, R. (1999a) "Examination of shape complementarity in docking of unbound proteins", *Proteins*, **36**, 307-317.
- Norel, R., Wolfson, H.J. and Nussinov, R. (1999b) "Small molecule recognition: solid angles surface representation and molecular shape complementarity", *Comb Chem High Throughput Screen*, **2**, 223-237.
- Nussinov, R. and Wolfson, H.J. (1991) "Efficient Detection of Three - Dimensional Motifs In Biological Macromolecules by Computer Vision Techniques", *Proceedings of the National Academy of Sciences, U.S.A.*, **88**, 10495-10499.
- O'neal, C.J., Jobling, M.G., Holmes, R.K. and Hol, W.G.J. (2005) "Structural basis for the activation of cholera toxin by human ARF6-GTP", *Science*, **309**, 1093-1096.
- Oldfield, T.J. (2001) "X-LIGAND: an application for the automated addition of flexible ligands into electron density", *Acta crystallographica*, **57**, 696-705.
- Oldfield, T.J. and Hubbard, R.E. (1994) "Analysis of C alpha geometry in protein structures", *Proteins*, **18**, 324-337.
- Ondrechen, M.J., Clifton, J.G. and Ringe, D. (2001) "THEMATICS: a simple computational predictor of enzyme function from structure", *Proc Natl Acad Sci U S A*, **98**, 12473-12478.
- Oppermann, U., Filling, C., Hult, M., Shafqat, N., Wu, X., Lindh, M., Shafqat, J., Nordling, E., Kallberg, Y., Persson, B. and Jornvall, H. (2003) "Short-chain dehydrogenases/reductases (SDR): the 2002 update", *Chem Biol Interact*, **143-144**, 247-253.
- Palmer, R.A. and Niwa, H. (2003) "X-ray crystallographic studies of protein-ligand interactions", *Biochem Soc Trans*, **31**, 973-979.
- Pearl, F.M.G., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J.M. and Orengo, C.A. (2003) "The CATH database: an extended protein family resource for structural and functional genomics", *Nucleic Acids Research*, **31**, 452-455.



- Perez-Nueno, V.I., Ritchie, D.W., Borrell, J.I. and Teixido, J. (2008) "Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket", *J Chem Inf Model*.
- Persson, B., Kallberg, Y., Oppermann, U. and Jornvall, H. (2003) "Coenzyme-based functional assignments of short-chain dehydrogenases/reductases (SDRs)", *Chem Biol Interact*, **143-144**, 271-278.
- Perutz, M., Rossmann, M., Cullis, A., Muirhead, H., Will, G. and North, A. (1960) "Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis", *Nature*, **185**, 416-422.
- Pickering, S.J., Bulpitt, A.J., Efford, N., Gold, N.D. and Westhead, D.R. (2001) "AI-based algorithms for protein surface comparisons", *Comput Chem*, **26**, 79-84.
- Pliska, V.K. (2001) "Substrate Binding to Enzymes". In, *Encyclopedia of Life Sciences*. Macmillan Publishers Ltd, Nature Publishing Group, London.
- Pliska, V.K. (2002) "Substrate Binding to Enzymes". In Atkins, D.e.a. (ed), *Encyclopedia of Life Sciences*. Macmillan Publishers Ltd, Nature Publishing Group, London, 594-603.
- Ponder, J.W. and Case, D.A. (2003) "Force fields for protein simulations", *Advances in protein chemistry*, **66**, 27-85.
- Porter, G.T., Bartlett, G.J. and Thornton, J.M. (2004) "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data", *Nucl. Acids. Res.*, **32**, 129.
- Poupon, A. (2004) "Voronoi and Voronoi-related tessellations in studies of protein structure and interaction", *Curr Opin Struct Biol*, **14**, 233-241.
- Ravelli, R.B. and Garman, E.F. (2006) "Radiation damage in macromolecular cryocrystallography", *Curr Opin Struct Biol*, **16**, 624-629.
- Reedy, C.J. and Gibney, B.R. (2004) "Heme protein assemblies", *Chemical reviews*, **104**, 617-649.
- Rhodes, G. (2000) *Crystallography made crystal clear*. Academic Press San Diego.
- Richards, F.M. (1974) "Interpretation of Protein Structures - Total Volume, Group Volume Distributions and Packing Density", *Journal of Molecular Biology*, **82**, 1-14.
- Richards, F.M. (1977) "Areas, Volumes, Packing, and Protein-Structure", *Annual Review of Biophysics and Bioengineering*, **6**, 151-176.
- Ringe, D. (1995) "What makes a binding site a binding site?", *Curr Opin Struct Biol*, **5**, 825-829.
- Ritchie, D.W. (1998) "Parametric Protein Shape Recognition". Departments of Computing Science and Molecular & Cell Biology, University of Aberdeen.
- Ritchie, D.W. (2003) "Evaluation of Protein Docking Predictions Using Hex 3.1 in CAPRI Rounds 1 and 2", *PROTEINS: Struct. Funct. Genet.*, **52**, 98-106.
- Ritchie, D.W. (2005) "High Order Analytic Translation Matrix Elements For Real Space Six-Dimensional Polar Fourier Correlations", *J. Appl. Cryst.*, **38**, 808-818.

- Ritchie, D.W. and Kemp, G.J.L. (1999) "Fast Computation, Rotation, and Comparison of Low Resolution Spherical Harmonic Molecular Surfaces", *J. Comp. Chem.*, **20**, 383-395.
- Ritchie, D.W. and Kemp, G.J.L. (2000) "Protein Docking Using Spherical Polar Fourier Correlations", *PROTEINS: Struct. Funct. Genet.*, **39**, 178-194.
- Roberts, S.A. and Montfort, W.R. (2007) "Haem Proteins". In, *Encyclopedia of Life Sciences*. Macmillan Publishers Ltd, Nature Publishing Group, London, 1-7.
- Rosen, M., Lin, S.L., Wolfson, H. and Nussinov, R. (1998) "Molecular shape comparisons in searches for active sites and functional similarity", *Protein engineering*, **11**, 263-277.
- Rossmann, M.G. (1990) "The molecular replacement method", *Acta Crystallogr A*, **46 ( Pt 2)**, 73-82.
- Schmitt, E., Moulinier, L., Fujiwara, S., Imanaka, T., Thierry, J.C. and Moras, D. (1998) "Crystal structure of aspartyl-tRNA synthetase from *Pyrococcus kodakaraensis* KOD: archaeon specificity and catalytic mechanism of adenylate formation", *The EMBO journal*, **17**, 5227-5237.
- Schmitt, S., Kuhn, D. and Klebe, G. (2002) "A new method to detect related function among proteins independent of sequence and fold homology", *J Mol Biol.*, **323**, 387-406.
- Schneider, H.J. (2008) "Ligand binding to nucleic acids and proteins: Does selectivity increase with strength?", *Eur J Med Chem*, **43**, 2307-2315.
- Schneider, S., Marles-Wright, J., Sharp, K.H. and Paoli, M. (2007) "Diversity and conservation of interactions for binding heme in b-type heme proteins", *Natural product reports*, **24**, 621-630.
- Shen, J. and Wendoloski, J. (1996) "Electrostatic binding energy calculation using the finite difference solution to the linearized Poisson-Boltzmann equation: Assessment of its accuracy", *J. Comput. Chem.*, **17**, 350-357.
- Shoichet, B.K. and Kuntz, I.D. (1991) "Protein docking and complementarity", *J Mol Biol*, **221**, 327-346.
- Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter, K. and Vajda, S. (2003) "Identification of substrate binding sites in enzymes by computational solvent mapping", *J Mol Biol*, **332**, 1095-1113.
- Sitkoff, D., Sharp, K.A. and Honig, B. (1994) "Accurate calculation of hydration free-energies using macroscopic solvent models", *J. Phys. Chem.*, **98**, 1978-1988.
- Sly, W.S. and Hu, P.Y. (1995) "Human carbonic anhydrases and carbonic anhydrase deficiencies", *Annu Rev Biochem*, **64**, 375-401.
- Smith, P.E. and Tanner, J.J. (2000) "Conformations of nicotinamide adenine dinucleotide (NAD(+)) in various environments", *J Mol Recognit*, **13**, 27-34.
- Smith, R.D., Hu, L., Falkner, J.A., Benson, M.L., Nerothin, J.P. and Carlson, H.A. (2006) "Exploring protein-ligand recognition with Binding MOAD", *J Mol Graph Model*, **24**, 414-425.
- Sommer, I., Muller, O., Domingues, F.S., Sander, O., Weickert, J. and Lengauer, T. (2007) "Moment invariants as shape recognition technique for comparing protein binding sites", *Bioinformatics*, **23**, 3139-3146.

Sotriffer, C. and Klebe, G. (2002) "Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design", *Farmaco*, **57**, 243-251.

Sousa, S.F., Fernandes, P.A. and Ramos, M.J. (2006) "Protein-ligand docking: current status and future challenges", *Proteins*, **65**, 15-26.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. and Willighagen, E. (2003) "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics", *J Chem Inf Comput Sci*, **43**, 493-500.

Stockwell, G.R. (2005) "Structural Diversity of Biological Ligands and their Binding Sites in Proteins". *Biomolecular Structure and Modelling Unit Department of Biochemistry and Molecular Biology University College London, London*, 347.

Stockwell, G.R. and Thornton, J.M. (2006) "Conformational diversity of ligands bound to proteins", *J Mol Biol*, **356**, 928-944.

Ten Eyck, L.F. and Watenpaugh, K.D. (2006) "18.1. Introduction to refinement". In Rossmann, M. and Arnold, E. (eds), *International Tables for Crystallography*. Kluwer Academic Publishers, Dordrecht, 369-374.

Terwilliger, T.C. (2003) "Automated main-chain model building by template matching and iterative fragment extension", *Acta crystallographica*, **59**, 38-44.

Testa, B., Carrupt, P.A., Gaillard, P., Billois, F. and Weber, P. (1996) "Lipophilicity in molecular modeling", *Pharm Res*, **13**, 335-343.

Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D.C. and Poda, G.I. (2006) "Can we estimate the accuracy of ADME-Tox predictions?", *Drug discovery today*, **11**, 700-707.

Thompson, D. and Simonson, T. (2006) "Molecular dynamics simulations show that bound Mg<sup>2+</sup> contributes to amino acid and aminoacyl adenylate binding specificity in aspartyl-tRNA synthetase through long range electrostatic interactions", *J Biol Chem*, **281**, 23792-23803.

Thompson, D.W. and Bonner, J.T. (1992) *On Growth and Form*. Cambridge University Press Cambridge.

Toth, G., R, F.M. and Lovas, S. (2001) "Stabilization of local structures by pi-CH and aromatic-backbone amide interactions involving prolyl and aromatic residues", *Protein engineering*, **14**, 543-547.

Traut, T.W. (1994) "Dissociation of enzyme oligomers: a mechanism for allosteric regulation", *Crit Rev Biochem Mol Biol*, **29**, 125-163.

Tsai, C.-J., Norel, R., Wolfson, H.J., Maizel, J.V. and Nussinov, R. (2001) "Protein-Ligand Interactions: Energetic Contributions and Shape Complementarity". In, *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd: Chichester.

Tsai, C.-J., Norel, R., Wolfson, H.J., Maizel, J.V. and Nussinov, R. (2002) "Protein-Ligand Interactions: Energetic Contributions and Shape Complementarity". In Atkins, D.e.a. (ed), *Encyclopedia of Life Sciences*. Macmillan Publishers Ltd, Nature Publishing Group, London, 505-512.

Tsuchiya, Y., Kinoshita, K. and Nakamura, H. (2004) "Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces", *Proteins*, **55**, 885-894.

Tsuchiya, Y., Kinoshita, K. and Nakamura, H. (2006) "Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity", *Protein Eng Des Sel*, **19**, 421-429.

Tsuzuki, S. and Fujii, A. (2008) "Nature and physical origin of CH/ $\pi$  interaction: significant difference from conventional hydrogen bonds", *Phys Chem Chem Phys*, **10**, 2584-2594.

Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. and Tanabe, K. (2000a) "Origin of the attraction and directionality of the NH/ $\pi$  interaction: Comparison with OH/ $\pi$  and CH/ $\pi$  interactions", *Journal of the American Chemical Society*, **122**, 11450-11458.

Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. and Tanabe, K. (2000b) "The magnitude of the CH/ $\pi$  interaction between benzene and some model hydrocarbons", *Journal of the American Chemical Society*, **122**, 3746-3753.

Uversky, V.N., Gillespie, J.R., Millett, I.S., Khodyakova, A.V., Vasilenko, R.N., Vasiliev, A.M., Rodionov, I.L., Kozlovskaya, G.D., Dolgikh, D.A., Fink, A.L., Doniach, S., Permyakov, E.A. and Abramov, V.M. (2000) "Zn(2+)-mediated structure formation and compaction of the "natively unfolded" human prothymosin alpha", *Biochemical and biophysical research communications*, **267**, 663-668.

Uversky, V.N., Oldfield, C.J. and Dunker, A.K. (2005) "Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling", *Journal of Molecular Recognition*, **18**, 343-384.

Vajda, S. and Guarnieri, F. (2006) "Characterization of protein-ligand interaction sites using experimental and computational methods", *Current opinion in drug discovery & development*, **9**, 354-362.

Veltkamp, R.C. (2001) "Shape matching: similarity measures and algorithms". *Shape Modeling and Applications, SMI 2001 International Conference on.*, 188-197.

Via, A., Ferre, F., Brannetti, B. and Helmer-Citterich, M. (2000) "Protein surface similarities: a survey of methods to describe and compare protein surfaces", *Cell Mol Life Sci*, **57**, 1970-1977.

Wallace, A.C., Laskowski, R.A. and Thornton, J.M. (1996) "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases", *Protein Sci.*, **5**, 1001-1013.

Wallimann, P., Marti, T., Furer, A. and Diederich, F. (1997) "Steroids in Molecular Recognition", *Chemical reviews*, **97**, 1567-1608.

Walsh, M.A., Evans, G., Sanishvili, R., Dementieva, I. and Joachimiak, A. (1999) "MAD data collection - current trends", *Acta crystallographica*, **55**, 1726-1732.

Wang, R.X., Gao, Y. and Lai, L.H. (2000a) "Calculating partition coefficient by atom-additive method", *Perspectives in Drug Discovery and Design*, **19**, 47-66.

Wang, R.X., Gao, Y. and Lai, L.H. (2000b) "LigBuilder: A multi-purpose program for structure-based drug design", *J Mol Model*, **6**, 498-516.

Weisel, M., Proschak, E. and Schneider, G. (2007) "PocketPicker: analysis of ligand binding-sites with shape descriptors", *Chemistry Central Journal*, **1**, 7.

Weisstein, E.W. (2009a) "Spherical Design", From MathWorld--A Wolfram Web Resource, <http://mathworld.wolfram.com/SphericalDesign.html>.

- Weisstein, E.W. (2009b) "Spherical Harmonics", From MathWorld--A Wolfram Web Resource, <http://mathworld.wolfram.com/SphericalHarmonic.html>.
- Winn, M.D., Murshudov, G.N. and Papiz, M.Z. (2003) "Macromolecular TLS refinement in REFMAC at moderate resolutions", *Methods in enzymology*, **374**, 300-321.
- Wlodek, S., Skillman, A.G. and Nicholls, A. (2006) "Automated ligand placement and refinement with a combined force field and shape potential", *Acta crystallographica*, **62**, 741-749.
- Wong, C.W. (1991) "1.5 Vector differentiation of a vector field". In, *Introduction to mathematical physics : methods and concepts*. Oxford University Press, Oxford, 26-32.
- Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation", *J Mol Biol*, **285**, 1735-1747.
- Wuthrich, K. (1990) "Protein structure determination in solution by NMR spectroscopy", *J Biol Chem*, **265**, 22059-22062.
- Yamagishi, M.E., Martins, N.F., Neshich, G., Cai, W., Shao, X., Beutrait, A. and Maignet, B. (2006) "A fast surface-matching procedure for protein-ligand docking", *J Mol Model*, **12**, 965-972.
- Yamazaki, K., Kahraman, A. and Thornton, J.M. (2009) "Protein-Ligand Induced-fit Docking using Spherical Harmonics Descriptors". *Journal of Molecular Biology*. (under review).
- Yang, A.S., Gunner, M.R., Sampogna, R., Sharp, K. and Honig, B. (1993) "On the calculation of pKas in proteins", *Proteins*, **15**, 252-265.
- Yee, A., Gutmanas, A. and Arrowsmith, C.H. (2006) "Solution NMR in structural genomics", *Curr Opin Struct Biol*, **16**, 611-617.
- Zhang, D.S. and Lu, G.J. (2004) "Review of shape representation and description techniques", *Pattern Recognit.*, **37**, 1-19.
- Zheng, J., Knighton, D.R., ten Eyck, L.F., Karlsson, R., Xuong, N., Taylor, S.S. and Sowadski, J.M. (1993) "Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor", *Biochemistry*, **32**, 2154-2161.
- Zwart, P.H., Langer, G.G. and Lamzin, V.S. (2004) "Modelling bound ligands in protein crystal structures", *Acta crystallographica*, **60**, 2230-2239.