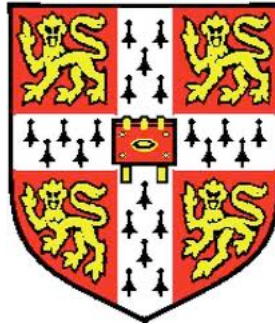


Molecular recognition from atomic interactions: insights into drug discovery



Alicia Perez Higuieruelo
Darwin College
University of Cambridge

A thesis submitted for the degree of
Doctor of Philosophy
December 2011

Declaration

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. My dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. This dissertation does not exceed the word limit set by the Biology Degree Committee.

For the Stapler

Acknowledgements

I have always believed “thank you” is a really small word. “Acknowledgements” is longer but still far away of expressing the feeling I have in writing this section. Without all of you this thesis wouldn’t have been possible. Simply. Thank you.

First and foremost I thank my supervisor Professor Sir Tom Blundell FRS, FMedSci, for his guidance and patience, for his respect and his ability to offer space to grow and for being such an amazing human being. I am also really grateful for the opportunity to continue studying molecular interactions in the lab as a postdoc. I look forward to it.

I thank Will Pitt, my industrial supervisor, former colleague, famous wizard and friend, for all these years of support, mentorship and collaboration, for running and for everything. For this thesis in particular, Chapter 4 would not have been the same without him.

I thank Colin Groom, for his encouragement and support to set up this project, for being such an inspirational individual, magnificent leader and friend.

And of course I thank the Stapler, to whom this thesis is dedicated.

I thank Nikolay Todorov for his friendship, for many thoughtful discussions over the years, for the magical connections and beautiful symmetries.

I thank Iwan de Esch for being always there and for believing in this project even before myself.

I thank my former colleagues at UCB from whom I learnt about Medicinal Chemistry, Biology and other side effects of life. It was a privilege to be part of that team.

I thank UCB for sponsoring this project when I was an UCB employee and for its continuous commitment after the closure of the site. I thank the BBSRC for funding the CASE studentship.

I thank my past and present colleagues at the Department of Biochemistry. Especially to Richard Bickerton for being such a easy person to work with, for his infinite patience helping me with PICCOLO and his brilliant

idea to use wordle to open the thesis chapters. To Adrian Schreyer for his help with CREDO and for sharing his impressive knowledge of python and databases. To Semin Lee for his kindness and patience helping me every time my scripts and connections were attacked by gremlins. To Sungsam Gong for his beautiful explanations about genes and his considerate behaviour with everybody. To Cynthia Lampert Moore for looking after all of us from all angles. To Graham Eliff for his tireless support and welcoming explanations about how everything works. To Christine Thulborn, Irene Kightley and Niki Miller for their constant availability and helping me out when tricky things appear. To Rinaldo Walter Montalvao for the thermodynamics course and the generosity of his knowledge. To Marko Hyvonen for sharing his PPI interest and sending relevant papers to enhance TIMBAL. To Noha Abdel-Rahman for her care and friendship. To Abel Moreno, Alex Litvinenko, Alison Baylay, Andy Ng, Anja Winter, Anna Sigurdardottir, Beata Blaszczyk, Bernardo Ochoa, Cat Donaldson, Clova Thompson, Dima Chirgadze, Harry Jubb, Heide Kirschenlohr, Janet Mellor, Jawon Song, Jim Metcalfe, Joe Maman, John Lester, Laura Perez Cano, Lauren Coulson, Leo Hernani Silvestre, Lorenzo Palmieri, Luca Pellegrini, Lynn Sibanda, Marcio, Martin Moncrieffe, May Marsh, Michal Blaszczyk, Monique Gangloff, Qian Wu, Ravi Nookala, Sachin Surade, Sara Lejon, Seema Patel, Sue Leach, Takashi Ochi, Victor Bolanos Garcia, Wi Duangrudee Tanramluk, Xue Pei and Zsuzsanna Ament for making the daily basis really pleasurable.

I thank Darwin College for being such a great College. Especially to Sally Williams for her kindness and help turning the spiral in the best direction. Grazie a Francesco Capponi per la sua amicizia e per farmi sentire a casa dal giorno uno. To the fantastic porters Derek Scott, June Cobden and Cliff Pennick. And last but forever to the Darwin College Boat Club where I have been so so happy, to Rebecca Rancourt and John Martin, The Bumps and the river Cam.

I thank the best crew ever in and outside the boat, Barbara Nikolaidou, Carol Baltar, Jessie Hohmann, Marloes Bagijn and Olivia Lewis. For all the

power10s and yanked finishes, for Savino's and because as Jessie said we'll be always a team!

I thank Aleix Altimiras for being so brave and for finding Carol. To both of them for their party and for giving such a great honour.

Gràcies al Pere de Antolin per acompanyar-me sempre.

Gràcies a la meva rosa blanca Dolo Fernandez i Floriach per compartir tots els moments; els bons i els no tan bons i per ser sempre un punt de referència crucial.

I thank Paloma Diaz Fernandez and Ryan Bentley for Los Miercoles and for all the things without name.

Grazie alla mia sosia Sabina Ferrauto, al Djavi Prats i l'Anna Laura Huguet per ser-hi sempre i per les millors vacances ever!

I thank Dorica Naylor for her care and strength. Grazie a Paola Spada per credere sempre ai miei studi ed essere una dona così forte e bella. Gracias a Esperanza Audrey Fernandez por saber compartir y ser feliz. Gracias a la Tata y al Vito.

Gracias a mis padres, sin ellos nada de todo esto hubiera ocurrido. Gracias a mi madre por ser fuerte y dulce, por su incansable apoyo y por ser la mejor Yaya del mundo!

Gràcies a la meva preciosa filla Clara, per el seu somriure i per fer de la meva vida un lloc meravellós. Life is the best present.

I thank David Stunning Bettinson for his solid support, for his smile, for all the important small things that make life with him just dreamland. T'estimo David.

All 3D representations of protein complexes have been done with the free version of Warren DeLano's PyMol (<http://www.pymol.org/>). The chapter openings figures have been done with Jonathan Feinberg's wordle (<http://www.wordle.net/>)

Abstract

The failure of the pharmaceutical industry to increase the delivery of new drugs into the market is driving a re-assessment of practices and methods in drug discovery and development. In particular alternative strategies are being pursued to find therapeutics that are more selective, including small molecules that target protein-protein interactions. However, success depends on improving our understanding of the recognition of small molecules by interfaces in order to develop better methods for maximising their affinity and selectivity, whilst trying to confer an appropriate therapeutic profile.

This thesis starts with the description of the creation of TIMBAL, a database that holds small molecules disrupting protein-protein interactions. The thesis then focuses on the analysis of these molecules and their interactions in a medicinal chemistry and structural biology context. TIMBAL molecules are profiled against other sets of molecules (drugs, drug-like and screening compounds) in terms of molecular properties. Using the structural databases in the Blundell group, the atomic detail of the interaction patterns of TIMBAL molecules with their protein targets are compared with other molecules interacting with proteins, comprising natural molecules, small peptides, synthetic small molecules (including drug-like and drugs) and other proteins. The structural features and composition of the binding sites of these complexes are also analysed. Keeping in mind that current drug candidates are somewhat too lipophilic to succeed, these interaction profiles are defined in terms of polar and apolar contacts, with the aim of migrating natural patterns into the design of new therapeutics.

Nomenclature

Acronyms and definitions

2P2I	Database of structures of protein-protein complexes with known inhibitors
3D	Three Dimensional
AASA	Apolar Solvent Accessible Surface Area
ACD-SC	Available Chemical Directory for Screening Compounds
ADMET	Absorption Distribution Metabolism Excretion and Toxicity
ADP	Adenosine Diphosphate
AMP	Adenosine Monophosphate
AND	3-Beta-Hydroxy-5-Androsten-17-one
ASA	Solvent Accessible Surface Area
ATP	Adenosine Triphosphate
B-catenin	Beta Catenin
BAD	Bcl-2 Associated Death promoter
Bcl-2	B-cell lymphoma 2
Bcl-XL	B-cell lymphoma-extra large
BEI	Binding Efficiency Index
BIPA	Database of protein-nucleic acid atomic interactions
CATH	Class Architecture Topology Homologous Superfamily
CCDC	Cambridge Crystallographic Data Centre
CD80	T-lymphocyte activation antigen CD80
CDR	Complementary Determining Region
ChEMBL	Database of bioactive drug-like small molecules
CMR1	Esport Receptor CMR1
cMyc	Myc proto-oncogene protein
COA	Coenzyme A
CREDO	Database of protein-ligand atomic interactions
CSD	Cambridge Structural Database
CTGFA	Combinatorial Target-Guided Fragment Assembly
CTLA4	Cytotoxic T-lymphocyte protein 4
distFC	Distance From Centre
DL	Drug Like
DMPK	Distribution Metabolism and Pharmacokinetics
DNA	Deoxyribonucleic Acid
DOS	Diversity Oriented Synthesis
E1	Replication protein E1
E2	Regulatory protein E2
EBI	European Bioinformatics Institute
ELISA	Enzyme-Linked Immunosorbent Assay
ELM	Eukaryotic Linear Motif Database

ER	Estrogen Receptor
ESP	Electrostatic Potential
EST	Estradiol
ESX	Epithelial-specific transcription factor
FAD	Flavin Adenine Dinucleotide
FCPI	Flow Cytometry Protein Interaction assay
FDA	The Food and Drug Administration
FMN	Flavin Mononucleotide
FPA	Fluorescence Polarization Assay
FtsZ	Cell Division protein FtsZ
GFRP	GTP-cyclohydrolase I Feedback Regulatory Protein
GLC	Alpha-D-Glucose
GTP	Guanosine-5'-Triphosphate
GTPCHI	GTP-cyclohydrolase I
GV	Gap Volume
HB	Hydrogen Bond
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HBPLUS	Software to calculate hydrogen bonds in proteins
HEM	Heme
HIV-1	Human Immunodeficiency Virus I
HMDB	The Human Metabolome Database
HSP90	Heat Shock Protein 90
HSV-Pol	Herpes Simplex Virus DNA polymerase catalytic subunit Pol
HTS	High Throughput Screening
HYP	Hydroxyproline
IgE-FC	Immunoglobulin-E bound to FC receptor
IGF-1	Insulin-like Growth Factor I
IL2	Interleukin 2
IL2Ra	Interleukin 2 Receptor Alpha subunit
Inh	Inhibitor
iNOS	Nitric Oxide Synthase, inducible
IRAK-4	Interleukin-1 Receptor Associated Kinase 4
ITC	Isothermal Titration Calorimetry
IUPAC	International Union of Pure and Applied Chemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes
LBDD	Ligand-Based Drug Design
LC	Long Chain
LE	Ligand Efficiency
LLE	Ligand Lipophilicity Efficiency
LM	Linear Motif
logP	Octanol-Water Partition Coefficient
Max	Myc-Associated factor X
mc	Main chain atoms
MDDR	MDL Drug Data Report
MDL	Molecular Design Limited

MDM2	Murine Double Minute
MGEx	Pure natural products list from AnalytiCon Discovery
MK2	MAPK-activated Protein Kinase-2
MLR	Multiple Linear Regression
MLY	Methyllysine
mmCIF	Macromolecular Crystallographic Information Files
MSE	Selenomethionine
MW	Molecular Weight
MySQL	Open source database engine
NAD	Nicotinamide Adenine Dinucleotide
NAP	Nicotinamide Adenine Dinucleotide Phosphate
NCC	Neighbouring Chemical Compounds
NES	Nuclear Export Signal
NME	New Molecular Entity
NMR	Nuclear Magnetic Resonance
NP	Natural Product
NPDDS	Nanoparticle Drug Delivery System
NR-LBD	Nuclear Receptor Ligand Binding Domain
nsSNP	Non-synonymous Single Nucleotide Polymorphisms
OEChem	OpenEye Chemical toolkit
OLS	Ordinary Least Squares
p53	Protein 53
PASA	Polar Solvent Accessible Surface Area
PCA	Principal Component Analysis
PCY	Pactamycin
PDB	Protein Data Bank
PDBBind	Database of experimentally measured affinity data for PDB entries
PICCOLO	Database of protein-protein atomic interactions
PISA	Protein Interactions Surfaces and Assemblies
PO4	Phosphate
PP	Protein-Protein
PPI	Protein-Protein Interactions
PPI-Net	Protein-Protein Interactions Network
PQS	Protein Quaternary Structure
PSA	Polar Surface Area
PTPP	Phosphotyrosine Protein Phosphatase
R&D	Research and Development
RGS4	Regulator of G-protein Signaling Protein 4
RNA	Ribonucleic Acid
S100B	S100 calcium binding protein B
SAH	S-Adenosyl-L-Homocysteine
SAM	S-Adenosylmethionine
SAR	Structure Activity Relationship
SBDD	Structure-Based Drug Design
Sc	Shape Complementarity

SC	Short Chain
sc	Side chain atoms
SCOP	Structural Classification of Proteins
SEI	Surface Efficiency Index
SLiM	Short Linear Motif
SLM	Short Linear Motif
SM	Small Molecule
Smac	Second Mitochondria-derived Activator of Caspases
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SPR	Surface Plasmon Resonance
Sur-2	Ras-linked subunit
Tcf4	Transcription Factor 4
TIMBAL	Database of small molecule inhibitors of protein-protein interactions
TNF	Tumor Necrosis Factor
ToxT	Virulence transcriptional activator
UL42	DNA-binding protein UL42
UniProt	Universal Protein Resource
vdw	Van der Waals
VS	Virtual Screening
XIAP	X-linked Inhibitor of Apoptosis Protein
ZipA	Cell Division protein ZipA

Chemical structures

The true protonation state of a molecule depends upon its environment and experimental conditions. Therefore, chemical structures are represented in the neutral form (unless they are permanent ions such as NAD) by a single tautomer. Hydrogens bound to heteroatoms are drawn explicitly.

Contents

DECLARATION	3
ACKNOWLEDGEMENTS	7
ABSTRACT	11
NOMENCLATURE.....	13
<i>Acronyms and definitions</i>	13
<i>Chemical structures</i>	16
CONTENTS.....	17
LIST OF FIGURES.....	23
CHAPTER 1	37
INTRODUCTION	
1.1 DRUG DISCOVERY.....	38
1.1.1 <i>Decline in drug discovery productivity</i>	38
1.1.2 <i>Medicinal chemistry practises</i>	39
1.1.3 <i>New targets: protein-protein interactions</i>	41
1.1.3.1 Challenging undruggability	43
Hot spots	44
Site adaptability.....	46
1.2 MOLECULAR RECOGNITION FROM ATOMIC INTERACTIONS	47
1.2.1 <i>Non-covalent forces</i>	48
1.2.1.1 van der Waals attractions	48
1.2.1.2 Hydrogen bonds.....	49
1.2.1.3 Weak hydrogen bonds.....	50
1.2.1.4 Ionic interactions.....	51
1.2.1.5 Hydrophobic interactions.....	52
1.2.1.6 Aromatic interactions.....	53
1.2.1.7 pi-cation interactions.....	54
1.2.1.8 Halogen bonds	54
1.2.1.9 Sulphur interactions.....	55
1.2.2 <i>Structural characteristics of protein-protein complexes</i>	55
1.2.2.1 Functional protein complexes.....	56
Specific vs. crystallographic complexes	57
1.2.2.2 Constituents and lifetime of protein complexes	58
1.2.2.3 Descriptors and topology of protein-protein interfaces	59
Size	60
Shape	60
Packing	60

Electrostatic interactions	61
Amino acid composition	61
Pairing preferences	62
1.2.3 Structural characteristics of protein-small molecule complexes	63
1.2.3.1 Classification of small molecules bound to proteins	64
1.2.3.1.1 Natural molecules	64
1.2.3.1.2 Synthetic small molecules	65
1.2.3.2 Peptide binding sites	65
1.2.3.3 Nucleotide and natural molecule binding sites	66
1.2.3.4 Synthetic molecule binding sites	69
1.2.3.5 Comparisons between different types of small molecules binding sites	72
1.3 AIMS OF THIS THESIS	74
CHAPTER 2	77
CREATION AND ANALYSIS OF THE TIMBAL DATABASE	
2.1 INTRODUCTION	78
2.1.1 Protein-Protein interactions (PPI) as drug targets	78
2.1.2 Survey of literature reviews of small molecules inhibitors of PPI	79
2.1.3 Need for collecting existing data and derive knowledge from it	83
2.2 METHODS	84
2.2.1 Creation of a database of small molecule inhibitors of PPI: TIMBAL	84
2.2.2 Profile and analysis of TIMBAL	86
2.2.2.1 TIMBAL profile	86
2.2.2.2 Pharmacophoric analysis of the interface	87
2.3 RESULTS AND DISCUSSION	89
2.3.1 TIMBAL database	89
2.3.2 Profile and analysis of TIMBAL	92
2.3.2.1 TIMBAL profile	92
2.3.2.2 Pharmacophoric analysis of the interface	102
2.3.2.2.1 Chemical functionality	102
2.3.2.2.2 Atomic contacts	104
2.3.2.2.3 Buried surface area	107
2.3.2.2.4 Ligand efficiency	107
2.4 CONCLUSIONS	111
CHAPTER 3	113
COMPARISON OF CREDO AND PICCOLO	
3.1 INTRODUCTION	114
3.1.1 CREDO	114
3.1.2 PICCOLO	115

3.2 METHODS.....	117
3.2.1 PICCOLO-CREDO intersection	117
3.3 RESULTS AND DISCUSSION	118
3.3.1 Stored distances	119
3.3.2 Atomic radii.....	119
3.3.3 Ionic, pi-cation, hydrophobic and aromatic contacts.....	120
3.3.4 Hydrogen bonds	121
3.3.5 Creation of simple contact definitions.....	123
3.4 CONCLUSIONS.....	127
CHAPTER 4	129
STRUCTURAL INTERACTION PROFILES OF PROTEIN-PROTEIN AND PROTEIN-SMALL MOLECULES	
4.1 INTRODUCTION	130
4.2 METHODS.....	132
4.2.1 General considerations and filtering.....	132
4.2.1.1 Interactions outside of each database scope.....	132
4.2.1.2 Crystallographic interactions.....	135
4.2.1.3 Ligands to remove.....	137
4.2.1.4 Identifiers.....	137
4.2.1.5 Redundancy removal.....	137
4.2.2 Contact definitions.....	138
4.2.3 Subset definitions.....	139
4.2.3.1 Small molecule protein-protein interactions inhibitors.....	139
4.2.3.2 Natural molecules.....	139
4.2.3.3 Small peptides.....	139
4.2.3.4 Drug-like molecules.....	140
4.2.3.5 Approved and oral drugs	140
4.2.3.6 Obligate and transient dimers	140
4.2.3.7 Quaternary interfaces.....	141
4.2.4 Data representation	141
4.2.4.1 Scissors plots.....	141
4.2.4.2 Multiple linear regression.....	143
4.2.4.3 Distribution polar versus sum of contacts.....	144
4.2.4.4 Bar charts of the polar/sumContacts ratio binned by sum of contacts	145
4.2.4.5 Contour plots.....	145
4.2.4.6 Molecular properties.....	145
4.2.4.7 Bar charts of matched and unmatched atoms	145
4.2.4.8 Buried surface area calculation.....	146
4.2.5 Hann's complexity model	146

4.3 RESULTS	149
4.3.1 <i>Data sets</i>	149
4.3.1.1 Small molecule protein-protein interaction inhibitors.....	151
4.3.1.2 Natural molecules.....	153
4.3.1.3 Small peptides.....	155
4.3.1.4 Drug-like molecules.....	156
4.3.1.5 Approved and oral drugs	157
4.3.1.6 Protein-protein sets.....	159
4.3.1.7 Resolution dependency.....	160
4.3.2 <i>Polarity of the interactions</i>	160
4.3.3 <i>More polar interactions in natural subsets</i>	162
4.3.4 <i>Atomic composition and molecular flexibility</i>	171
4.3.5 <i>Matched and unmatched atoms at the binding interfaces</i>	173
4.3.6 <i>Drug-like complexes. Property versus interaction profile</i>	176
4.3.7 <i>Drug-like complexes. Affinity versus interaction profile</i>	178
4.3.8 <i>Menagerie of small molecules for the same target</i>	182
4.3.8.1 Estrogen receptor	183
4.3.8.2 HIV-1 Reverse transcriptase.....	184
4.3.8.3 HIV-1 Protease – retropepsin.....	184
4.3.8.4 Thrombin	186
4.3.8.5 Approved and oral drugs	186
4.3.9 <i>Small molecule inhibitors of PPI</i>	189
4.3.10 <i>Natural molecules and small peptides</i>	191
4.4 DISCUSSION	195
4.5 CONCLUSIONS.....	199
CHAPTER 5	201
STRUCTURAL FEATURES OF BINDING SITES IN PROTEIN-PROTEIN AND PROTEIN-SMALL MOLECULE COMPLEXES	
5.1 INTRODUCTION	202
5.1.1 <i>Other studies classifying interfaces and cavities</i>	202
5.1.1.1 Pocket detection and druggable interfaces.....	202
5.1.1.2 Protein-protein interfaces	204
5.1.1.3 Protein-protein interfaces inhibited by small molecules.....	205
5.2 METHODS.....	208
5.2.1 <i>Subsets definitions</i>	208
5.2.2 <i>Definition of binding interfaces and binding pockets</i>	208
5.2.3 <i>Residue propensity plots</i>	210
5.2.4 <i>Depth of the protein atoms at the interface</i>	211

<i>5.2.5 Size of the protein in protein-protein complexes.....</i>	212
<i>5.2.6 Statistical treatment.....</i>	213
5.3 RESULTS AND DISCUSSION	214
<i>5.3.1 Pocket detection algorithms.....</i>	214
<i>5.3.2 Residue propensity.....</i>	218
5.3.2.1 Charged, polar and hydrophobic	224
5.3.2.1.1 Protein-protein complexes inhibited by small molecules	229
5.3.2.2 Small, medium and bulky	236
5.3.2.3 Constrained, free, rigid, medium, flexible and aromatic.....	237
<i>5.3.3 Proportion of main chain atoms at the binding interfaces.....</i>	239
<i>5.3.4 Proportion of polar atoms at the binding interface.....</i>	245
<i>5.3.5 Depth of protein atoms at the binding interface.....</i>	248
5.3.5.1 Depth of the protein atoms at the interface versus chain length.....	250
<i>5.3.6 Density of contacts at the binding interface.....</i>	253
5.4 CONCLUSIONS.....	258
CHAPTER 6	261
CONCLUSIONS	
6.1 PROTEIN-PROTEIN INTERACTIONS AS DRUG TARGETS.....	262
6.2 MOLECULAR RECOGNITION, SYNTHETIC VERSUS NATURAL MOLECULES	263
6.3 CONCLUDING REMARKS	264
BIBLIOGRAPHY	265

List of figures

- Figure 1.1. Stages of the drug discovery process. Reprinted from (Lombardino *et al.* 2004). 39
- Figure 1.2. Concept of undruggable surfaces. a: Protein (green) with a cavity evolved to recognise an endogenous ligand (grey) with multiple interactions converging in a small volume. b: Classical drug target where the drug molecule (magenta) occupies the volume maximising interactions, as most of its surface is in contact with the protein target. c: Protein-protein complex (green and grey) with a large surface with spread interactions. d: Small drug molecule (magenta) cannot engage many interactions as the absence of grooves translates into small contact areas. Blue arrows represent hydrogen bonds and yellow patches represent hydrophobic contacts. Reprinted from (Whitty *et al.* 2006)..... 43
- Figure 1.3. Lennard-Jones potential (V) for Argon dimer as function of the interatomic distance (r). The minimum of potential corresponds to ϵ and potential is equal to zero when the distance is σ 49
- Figure 1.4. Relative geometries of side chain charged groups in proteins. Colour coded per ion-pair type: Salt bridges (blue, side chain centroids and O-N pairs from Glu/Asp-Arg/Lys/His are within 4Å), N-O bridges (green, only O-N pairs are within 4Å, but not the side chain centroids) and longer range ion pairs (red, centroids and N-O pairs more than 4Å apart). Geometry of the ion-pairs is represented by the distance between centroids of the charged groups (radii of the polar plot) and by the relative angular orientation between side chains (angle of the polar plot measured as the angle between the vectors formed by the C-alpha and the side chain centroid of each residue). Most ion pairs with distances $\leq 5\text{Å}$ are stabilizing of the structure and destabilizing for longer distances. See original paper for details. Reprinted from (Kumar *et al.* 2002)..... 52
- Figure 1.5. Definition of types of protein-protein interactions by function of their binding affinity (Y axis) and the localisation of the protomers (X axis). In red the factors that affect transient interactions. * denotes large conformational changes that usually occur with the association. Reprinted from (Nooren *et al.* 2003). 59
- Figure 1.6. Chemical structures of the set of molecules studied by Kahraman *et al.* (Kahraman *et al.* 2007; Kahraman *et al.* 2010). The labels show the three-letter code of the molecule in the HET entry and the number in parenthesis denotes the number of instances used in the Kahraman study..... 67
- Figure 2.1. Number of publications per year (normalised by the total number of publications per year) containing in the title “protein-protein interaction”. The colour code is as follows: blue, only PPI in the title; red, ppi and small molecule (SM) in the title; orange, ppi and inhibitor (inh) in the title; yellow, all the above in the title. Searches have been done in PubMed..... 79

Figure 2.2. Complete schema of TIMBAL database. Chemical structures are held as SMILES (Simplified Molecular Input Line Entry System), generated with the Accord functionality within Excel. These sets of tables have been defined to normalise TIMBAL and avoid redundancy.	85
Figure 2.3. Distribution of molecular properties for the different sets of molecules described in the main text. See section 2.2.1. Colour coded: dark blue (PDB ligands), grid dark blue (PDB ligands drug-like subset), yellow (Drugs from MDDR), cyan (Screening compounds), pink (TIMBAL, small molecule inhibitors of protein-protein interactions). MW: Molecular weight; alogP: Calculated logarithm of the partition coefficient; NRings: Number of rings; RotBonds: Rotatable bonds.	93
Figure 2.4. Distribution of molecular properties for the different sets of molecules described in the main text. See section 2.2.1. Colour coded: dark blue (PDB ligands), grid dark blue (PDB ligands drug-like subset), yellow (Drugs from MDDR), cyan (Screening compounds), pink (TIMBAL, small molecule inhibitors of protein-protein interactions). PSA: Polar surface area; HBA: Number of hydrogen bond acceptors; HBD: Number of hydrogen bond donors.	94
Figure 2.5. Distribution of the Molecular Weight (MW) of the TIMBAL molecules colour coded by target. Only targets with more than one molecule are plotted.	97
Figure 2.6. Distribution of the calculated logarithm of the partition coefficient (alogP) of the TIMBAL molecules colour coded by target. Only targets with more than one molecules are plotted.	98
Figure 2.7. Three-dimensional projection of the principal components of the molecular properties for the different sets of molecules.	100
Figure 2.8. Distribution of the distances to the arithmetic centre of the PCA space for each set of molecules. TIMBAL molecules represented by dots for clarity. The mean of this distance is 2.18 with a standard deviation of 1.49. Table 2.4 shows the percentage of molecules in each bin.	101
Figure 2.9. Range of Ligand Efficiency, LE (X axis) of the TIMBAL molecules separated by target.	108
Figure 2.10. Average of the molecular properties for TIMBAL molecules binned by LE. Blue: Average of the sum of hydrogen bond donors and acceptors. Red: Average of rotatable bonds. Yellow: Average of alogP. Black: Average of number of atoms.	110
Figure 3.1. Scatter plots of the comparison of PICCOLO and CREDO contacts. In all nine plots, X-axes are for CREDO contacts and Y-axes for PICCOLO contacts. Each scatter plot is for one of the common contact types in both databases, from top left to bottom right: covalent, van der Waals, van der Waals clash, hydrogen bond, ionic, pi-cation, hydrophobic and proximal. Proximal is defined as when the two atoms are less than or equal to 6.05Å apart, the maximal distance of a water-mediated hydrogen bond. The red line in each plot denotes the slope that is given when the two databases give identical results.	118

- Figure 3.2. Geometric criteria for hydrogen bonds used in HBPLUS, adapted from figure 1 in (McDonald *et al.* 1994). D is the donor heavy atom. H is hydrogen, A is the acceptor heavy atom. DD is donor antecedent (an atom two covalent bonds away from the hydrogen). AA is acceptor antecedent. All three angles highlighted in the figure are required to be greater than or equal to 90 degrees to meet the hydrogen bond criterion. 122
- Figure 3.3. Scatter plots of specific contacts versus simple contacts for each database and type for the subset common to both databases. Simple polar and apolar contacts are distance cut-offs between polar-polar and apolar-apolar atom type as described in the text. Specific contacts refer to the contacts defined in CREDO and PICCOLO. Hydrogen bond, pi-cation and ionic are considered as polar contacts and hydrophobic is considered as apolar. The green line has a slope = 1 to aid visualisation. See Table 3.3 for details of the linear correlation..... 125
- Figure 3.4. Scatter plots of buried surface area upon binding and the number of atomic contacts (polar, apolar and sum of contacts) for the subset of complexes common to both databases. The sum of contacts has been calculated over all interacting chains for comparison with buried area..... 126
- Figure 4.1. Structure 1HNX (30S ribosomal subunit in complex with Pactamycin). Small molecule ligand (PCY) represented by red spheres. Ribosomal RNA in cyan cartoon, fragment of messenger RNA in orange cartoon. Protein S7 in blue cartoon with surface and Protein S11 in magenta cartoon with surface..... 133
- Figure 4.2. Binding interface between human immunoglobulin epsilon chain C (IgE-FC in cyan) and its high affinity immunoglobulin epsilon receptor alpha subunit (magenta) from PDB entry 1F6A. At this interface, electron density is also observed for five molecules of the CHAPS detergent (only steroid heads resolved, in stick representation with different colour for each CHAPS molecule). 134
- Figure 4.3. Structure 1A42, human carbonic anhydrase II complexed with brinzolamide. Zinc atom is represented by a black sphere, protein atoms by pale pink lines and brinzolamide ligand by magenta sticks..... 135
- Figure 4.4. Left: Structure 1T6J, phenylalanine ammonia-lyase with carboxycinnamic acid (magenta spheres). Right: Acriflavine resistance protein B with Ciprofloxacin. This molecule (stick representation) binds into two independent sites, the interaction with more atomic contacts is kept for the analysis..... 136
- Figure 4.5. Example of a scissor plot. X axis represents sum of contacts (as polar + apolar). Y axis represents the contacts, apolar in blue and polar in red. See text for discussion about these graphs..... 142
- Figure 4.6. Heteroscedasticity. Fan shape of the residuals for the apolar regression line of the drug-like (DL) set. 144
- Figure 4.7. PDB 1PW6, crystal structure of IL-2 bound to inhibitor SP2456. This entry was not considered for the non-redundant subset of inhibitors of protein-protein interactions

because the small molecule (in stick representation, green and yellow) interacts with itself in the crystal packing. Note these are identical molecules packed in the asymmetric unit.....	152
Figure 4.8. Examples of chemical structures of the small molecules inhibiting protein-protein complexes. Each structure is labelled with the protein complex it inhibits.....	153
Figure 4.9. Distribution of the natural small molecule subset in terms of entries per chemical structure of the small molecule bound to protein. Only higher frequency entries are labelled for clarity. Note more than half of the subset is composed of the complexes with eight different molecules: ADP, NAD, NAP, ATP, AMP, FAD, SAH and COA.	154
Figure 4.10. Examples of chemical structures in the natural molecules set. Labels correspond to the manual classification based in their structures and functions, so these molecules are categorised into natural product like, peptide like, steroid like, sugar like, lipid like, antibiotic like and nucleotide like.	155
Figure 4.11. Examples of chemical structures in the drug-like subset. Molecules are labelled with their hetID (residue) identifier from the PDB. Ligand 8PP is depicted here as an extreme example of the result of the broad filters applied to select these molecules...	157
Figure 4.12. Examples of chemical structures in the approved and oral drugs set. Labels correspond to the manual classification based on their structures, so these molecules are categorised into natural product like, peptide like, steroid like, sugar like, lipid like, antibiotic like, nucleotide like and none of the above (NOTA).	159
Figure 4.13. Resolution versus ratio of polar contacts as (polar/[polar + apolar]) for the protein-small molecule complexes (left) and for the protein-protein complexes (right). Contour levels show the density of points in the graphs, where red denotes high density and pale blue low density.	160
Figure 4.14. Scatter plot of buried surface area upon binding and the number of atomic contacts (polar and apolar) the small molecules made. Points are from all small molecule sets: drug-like, approved drugs, oral drugs, protein-protein interaction inhibitors, natural molecules and small peptides.	161
Figure 4.15. Scissors plots for the non-redundant-by-complex (table 1) sets of protein complexes. A: drug-like small molecules bound to proteins. B: Protein-protein interactions small molecule inhibitors bound to proteins. C: Small peptides bound to proteins. D: Natural small molecules bound to proteins. E: Natural small molecules without containing phosphor bound to proteins. F: Transient protein-protein dimers. G: Obligate protein-protein dimers. H: Homo protein-protein interfaces from quaternary structures. I: Hetero protein-protein interfaces from quaternary structures. Polar (red) and apolar (blue) contacts are scattered against sum of contacts. Details of the regression lines for each graph and contact type can be found in Table 4.4.	164
Figure 4.16. Normalised distributions of the ratio of polar contacts (represented by polar/[polar+apolar]), each chart compares drug-like against the others. A: drug-like versus natural small molecules with and without phosphor. B: drug-like versus approved	

- and oral drugs. C: drug-like versus small peptides, obligate and transient protein-protein dimers, homo and hetero quaternary protein-protein interfaces. D: drug-like versus PPI inhibitors. 167
- Figure 4.17. Comparisons of polar/sumContacts ratio means, binned by sum of contacts (polar+apolar), each chart compares drug-like against the others. A: drug-like versus approved and oral drugs. B: drug-like versus PPI inhibitors. C: drug-like versus small peptides. D: drug-like versus natural molecules. E: drug-like versus natural molecules without phosphor. F: drug-like versus transient protein-protein dimers. G: drug-like versus obligate protein-protein dimers. H: drug-like versus homo quaternary protein-protein interfaces. I: drug-like versus hetero quaternary protein-protein interfaces. Error bars denote the standard error of the mean. 169
- Figure 4.18. Ratio of polar/(polar+apolar) versus sum of contacts (polar+apolar). Contour levels show the density of points in the graphs, where red denotes high density and pale blue low density. The black line in all the graphs goes between 0.9 ratio to 200 sum of contacts to have the same reference to aid comparison between sets. A: drug-like small molecules bound to proteins. B: Approved and oral drugs bound to proteins. C: Small peptides bound to proteins. D: Natural small molecules bound to proteins. E: Natural small molecules without containing phosphor bound to proteins. F: Transient protein-protein dimers. G: Obligate protein-protein dimers. H: Homo protein-protein interfaces from quaternary structures. I: Hetero protein-protein interfaces from quaternary structures. For clarity, graphs for protein-protein complexes are plotted up to 600 contacts only. 170
- Figure 4.19. A: Distribution of the ratio of number of heteroatoms by number of heavy atoms for drug-like small molecules, natural molecules, natural molecules without phosphor and small peptides. B: Distribution of the ratio number of heteroatoms versus number of heavy atoms for drug-like small molecules, approved and oral drugs. C: Distribution of the ratio of number of rotatable bonds by number of heavy atoms for drug-like small molecules, natural molecules, natural molecules without phosphor and small peptides. D: Distribution of the ratio of number of rotatable bonds by number of heavy atoms for drug-like small molecules, approved and oral drugs. 173
- Figure 4.20. Mean of the percentage of buried atoms engaged in successful interactions (Matched contacts, left chart) and mean of the percentage of buried atoms without an appropriate partner in the other side of the interface (Unmatched contacts, right chart). The percentage is divided into polar (red) and apolar (blue) contribution. Each subset has two bars, one on the left for the atoms in the protein and one on the right for the atoms in the ligand or smaller protein in the case of protein complexes. Error bars denote the standard error of the mean. Subsets are ordered from left to right: Drug-like small molecules, approved Drugs, oral drugs, PPI small molecule inhibitors, natural molecules, natural molecules without phosphor, small peptides, obligate protein-protein

- dimers, transient protein-protein dimers, homo quaternary protein-protein interfaces and hetero quaternary protein-protein interfaces..... 175
- Figure 4.21. Linear correlation of the ratio of heteroatoms by number of heavy atoms versus the ratio of polar contacts by sum of contacts for the natural-product-like subset of the natural molecules set..... 176
- Figure 4.22. Ratio of polar/(polar+apolar) versus molecular weight (A), AlogP (B), buried area upon binding (C) and sum of contacts (D) for protein complexes with drug-like small molecules. Different colours denote SCOP families: Protein kinase catalytic subunit (green), nuclear receptor ligand-binding domain (blue), eukaryotic proteases (red), retroviral proteases - retropepsin (cyan), reverse transcriptase (magenta), Higher-molecular weight phosphotyrosine protein phosphatases (yellow), HSP90 N-terminal domain (black). For clarity, only SCOP families binding to more than 20 different ligands are shown. 177
- Figure 4.23. Free energy of ligand binding versus the polar ratio of contacts [polar/(polar+apolar)] for the drug-like set (yellow) and the small peptide set (blue).... 179
- Figure 4.24. Binned binding affinity (BA) data for drug-like small molecules (A and B) and for small peptides (C and D). Bars in A and C denote the average of molecular properties for each affinity bin: alogP (yellow), rotatable bonds (red), sum of hydrogen bond donors and acceptors (blue) and number of atoms (black). Bars in B and D denote the average of the ratio of polar contacts [polar/(polar+apolar)] (orange) and the average of the ratio of heteroatom content [num heteroatoms/num atoms] (cyan). Error bars are the standard error of each sample..... 180
- Figure 4.25. Free energy of ligand binding versus the number of atoms of the ligand. Drug-like set is plotted in yellow, and small peptide set in blue..... 181
- Figure 4.26. Ratio of polar/(polar+apolar) versus AlogP for four different proteins. A: Estrogen receptor from NR-LBD SCOP family. B: HIV-1 Reverse transcriptase from Ribonuclease H SCOP family. C: HIV-1 Protease from retroviral proteases SCOP family. D: Thrombin heavy chain from the eukaryotic proteases SCOP family. Colour coding refers to the subsets, which the small molecules belong to: Oral drugs (magenta), Approved drugs (cyan), Natural molecules (green), Small peptides (blue) and Drug-like (yellow)..... 183
- Figure 4.27. Bcl-XL bound with one of its putative partners (BAD, in magenta, PDB 2BZW) and with small molecule inhibitor (ABT-737, in cyan PDB 2YXJ). Only polar contacts are shown for clarity. Colour of the contacts (in dotted lines) is the same as the molecules making them. Synthetic molecule only uses a fraction of the polar contacts available for the natural counterpart..... 191
- Figure 4.28. Examples of natural molecules (magenta) and drug-like molecules (cyan) binding to the same protein target. Only polar contacts are shown for clarity. Colour of the contacts (in dotted lines) are the same as the molecules making them. LEFT: Visfatin with Nicotinamide Mononucleotide (2G96) and FK-866 (2G97). RIGHT: Phospholipase

- A2 with a tetrapeptide (2O1N) and Diclofenac (2B17). In both cases synthetic molecules only use a fraction of the polar contacts available for the natural counterparts..... 194
- Figure 5.1. Structures of human IRAK-4 bound to different small molecules. LEFT: Staurosporine, 2NRY. RIGHT: benzimidazole inhibitor, 2NRU. Cyan cartoon represents the kinase domain; residues within 4.5Å of each ligand are displayed in magenta with stick representation of the side chains. The frontal loop of the five non-contacting residues is not shown for clarity..... 209
- Figure 5.2. The concept of Rinaccess calculation. Reprinted from (Kawabata 2010). Three spherical probes are used: 3Å, 4Å and 5Å. Grid representation captures the smallest of the larger spheres that cannot access the grid point; the number represented is the radius of the sphere plus the grid resolution. Red and blue shapes represent different ligands bound in different regions of the pocket. The average of grid values per ligand gives a measure of the depth where the ligand is bound. 212
- Figure 5.3. LEFT: IRAK-4 kinase domain (2NRU) bound to benzimidazole inhibitor (not shown). The magenta region represents the residues in contact with the inhibitor; coloured spheres represent the pockets predicted by Fpocket, where each colour represents a different pocket. RIGHT: Protein-based overlay of the pocket prediction from Fpocket shown on the left with Staurosporine from 2NRY. The benzimidazole inhibitor is represented by blue sticks, Staurosporine with magenta sticks. Note the binding mode for benzimidazole inhibitor is covered by pocket 1 (red) and pocket 4 (orange). 215
- Figure 5.4. Comparison of pocket detection by Fpocket (LEFT) and ghecom (RIGHT) for the human IRAK-4 (2NRU). Ghecom gives one single large pocket in the ATP binding site (magenta cloud), whereas Fpocket gives several different ones. The small coloured clouds on the right picture are the additional pockets found by ghecom. 216
- Figure 5.5. Comparison of the volumes for the pockets identified by Fpocket (Y axis) and ghecom (X axis) programs for the small molecule data sets (Drug-like, drugs, natural molecules and small peptides). These points represent the volume of the pocket that matched the ligand bound. Red straight line represents the line of slope one to aid comparison. One quarter (23%) of the pockets have greater volume for Fpocket than ghecom. 217
- Figure 5.6. Comparison of the number of residues forming a pocket from Fpocket (left) and ghecom (right) predictions versus the number of contacting residues from the binding partner (buried residues). Red straight line represents the line of slope one to aid comparison. For Fpocket, 23% of the predicted pockets enclose fewer residues than the residues buried upon binding, on average 77% of the binding site is covered for these cases. For ghecom this proportion is less than 0.1%, and for these few cases more than the 90% of the binding site is covered by the prediction..... 217
- Figure 5.7. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (yellow) and SCOP family (orange) for the drug-like set.

Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).218

Figure 5.8. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (cyan) and SCOP family (magenta) for the approved drug set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).219

Figure 5.9. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (green) and SCOP family (light blue) for the oral drugs set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).219

Figure 5.10. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (magenta) and SCOP family (light pink) for the small molecule PPI set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).220

Figure 5.11. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (purple) and SCOP family (bright pink) for the natural molecules set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).220

Figure 5.12. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (pale pink) and SCOP family (pale green) for the natural molecules not containing phosphorus set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).221

Figure 5.13. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (blue) and SCOP family (grey) for the small peptides set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).221

Figure 5.14. Comparison of residue propensities for long chain (LC, pale pink) and short chain (SC, magenta) of the homo quaternary interfaces of the protein complexes. Bar heights represent the mean of the percentage of each residue at the interface. Error

- bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).222
- Figure 5.15. Comparison of residue propensities for long chain (LC, pale orange) and short chain (SC, green) of the hetero quaternary interfaces of the protein complexes. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).223
- Figure 5.16. Comparison of residue propensities for long chain (LC, bright green) and short chain (SC, blue) of the obligate protein dimers. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).223
- Figure 5.17. Comparison of residue propensities for long chain (LC, bright green) and short chain (SC, blue) of the transient protein dimers. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).224
- Figure 5.18. Comparison of residue propensities at the binding sites for drug-like (yellow) versus natural molecules (purple). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).225
- Figure 5.19. Comparison of residue propensities at the binding sites for natural molecules (purple) versus natural molecules without phosphorus (pale pink). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).225
- Figure 5.20. Comparison of residue propensities at the binding sites for drug-like (yellow) versus protein-protein quaternary hetero interfaces (pale pink). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).227
- Figure 5.21. Comparison of residue propensities at the binding sites for obligate dimers (bright green) versus transient dimers (dark blue). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).227
- Figure 5.22. Average proportion of charged (red), polar (orange) and hydrophobic (blue) residues at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors,

- natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero- quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only the long chain is considered.228
- Figure 5.23. Comparison of residue propensities at the binding sites for small molecule protein-protein inhibitors (magenta) versus protein-protein complexes inhibited by them (cyan). Note these subsets are small (9 and 7 complexes respectively). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).....229
- Figure 5.24. **S100B**. Upper left: 1DT7, S100B (cyan) with the C-terminal negative regulatory domain of p53 (green). Lower right: 3GK1, S100B (dark grey) with small molecule inhibitor (green). The surface covers the S100B residues that are within 4.5Å of p53. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. .230
- Figure 5.25. **IL-2**. Upper left: 1Z92, IL-2 (cyan) bound to IL-2R alpha subunit (green). Lower right: 1PY2, IL-2 (dark grey) with a Sunesis small molecule inhibitor (green). The surface covers the IL-2 residues that are within 4.5Å of the IL-2Ra. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.231
- Figure 5.26. **MDM2**. Upper left: 1YCR, MDM2 (cyan) bound to the transactivation domain of p53 (green). Lower right: 1T4E, MDM2 (dark grey) with a benzodiazepine inhibitor (green). The surface covers the MDM2 residues that are within 4.5Å of the p53. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.232
- Figure 5.27. **ZipA**. Upper left: 1F47: ZipA (cyan) bound to a fragment of FtsZ (green). Lower right: 1Y2F: ZipA (dark grey) with an aminopyrimidine inhibitor (green). The surface covers the ZipA residues that within 4.5Å of the FtsZ. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note the small molecule does not engage a single polar contact.....233
- Figure 5.28. **XIAP**. Upper left: 1G3F, BIR3 domain of XIAP (cyan) bound to an active nine-residue peptide derived from Smac (green). Lower right: 1TFT, XIAP (dark grey) with a small molecule inhibitor (green). The surface covers the XIAP residues that within 4.5Å of the Smac fragment. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.234
- Figure 5.29. **Bcl-XL**. Upper left: 2BZW, Bcl-XL (cyan) bound to BAD (green). Lower right: 2YXJ, Bcl-XL (dark grey) with the Abbott compound ABT-737 (green). The surface covers the Bcl-XL residues that within 4.5Å of BAD. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note that the small molecule only engages polar contacts at the bottom of the picture and it is bound to Bcl-XL mainly through apolar contacts.235
- Figure 5.30. **TNF**. Upper left: 1TNF, TNF alpha trimer, two chains are coloured in cyan and the third in green. Lower right: 2AZ5, two chains of the TNF trimer (dark grey) bound to

a small molecule (green) that accelerates subunit dissociation. The surface covers the residues in these chains that are within 4.5Å of the third chain. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note small molecule binds to an area where there are no interactions in the trimer.236

Figure 5.31. Average proportions of small (cyan), medium (green) and bulky (magenta) residues at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only the long chain is considered.237

Figure 5.32. Average of the proportion of constrained (yellow), free (orange), rigid (red), medium (green), flexible (cyan) and aromatic (blue) at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only long chain is considered.....239

Figure 5.33. Average of the percentage of main chain atoms for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. A and C: percentage of main chain atoms at the interface (defined as atoms within 4.5Å of the binding partner) colour coded by the proportion that are matched (magenta) or unmatched (cyan). B and D: percentage of main chain atoms from the matched atoms colour coded by polar (red) and apolar (blue). Both levels of redundancy are plotted, A and B: protein-small molecule complexes with distinct UniProt identifiers. C and D: proteins-small molecule complexes belonging with distinct SCOP families.....240

Figure 5.34. Average percentage of contacts involving main chain atoms for each molecular subset for both levels of protein redundancy: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. Colour coded by polar (red) and apolar (blue).....242

- Figure 5.35. Distribution of the natural small molecule subset (filtered for protein redundancy by distinct UniProt) in terms of entries per chemical structure of the small molecule bound to protein. Only higher frequency entries are labelled for clarity. Note that more than half of the subset is composed of the complexes with seven different molecules: ADP, NAD, FAD, NAP, ATP, AMP and SAH.243
- Figure 5.36. Scatter plot of the number of different SCOP families bound to the same small molecule versus the average of contacts involving main chain atoms that these molecules are engaging.....245
- Figure 5.37. Average percentage of protein polar atoms for each molecular subset: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. A and C: percentage of protein polar atoms at the interface (defined as atoms within 4.5Å of the binding partner) colour coded by the proportion that are matched (magenta) or unmatched (cyan). B and D: percentage of protein polar atoms from the total atoms that are matched. Both levels of redundancy are plotted, A and B: protein-small molecule complexes with distinct UniProt identifiers. C and D: proteins-small molecule complexes belonging with distinct SCOP families.246
- Figure 5.38. Distribution of the average of protein polar atoms at the binding interface for drug-like molecules at the UniProt level by molecular weight of the small molecule. The proportion of polar atoms is colour coded if they are engaged in successful interactions with the ligand (magenta) or are unmatched (cyan). Error bars denote the standard error of the mean.248
- Figure 5.39. Proportion of Rinaccess values for the atoms at the interface for each molecule set at the SCOP family redundancy level. The colour in the bars denotes the Rinaccess values: red (<2Å), orange (2-3Å), yellow (3-4Å), green (4-5Å), cyan (5-6Å), blue (6-7Å), grey (7-10Å) and black (> 10Å). A: for all atoms at the interface, B: for main chain atoms at the interface, C: for polar atoms at the interface and D: for polar main chain atoms at the interface. For protein-protein complexes, only the longest chain is considered.249
- Figure 5.40. Normalised distribution of the ratio between the lengths of short and long chain for the protein-protein complexes subsets: Obligate dimers, Transient dimers, Hetero quaternary interfaces and protein-protein (PP) complexes inhibited by small molecules (SM). Homo quaternary interfaces are not plotted, as they have virtually no difference in chain length, see Table 5.2.....251
- Figure 5.41. Proportion of Rinaccess values for the atoms at the interface of the long chain of the hetero quaternary interfaces (A) and for the long chain (with at least 100 residues in length) of the transient dimers and protein-protein complexes inhibited by small molecules (B). The colour in the bars denotes the Rinaccess values: red (<2Å), orange

(2-3Å), yellow (3-4Å), green (4-5Å), cyan (5-6Å), blue (6-7Å), grey (7-10Å) and black (> 10Å). Each bar represents different length range for the short length of the complex. 253

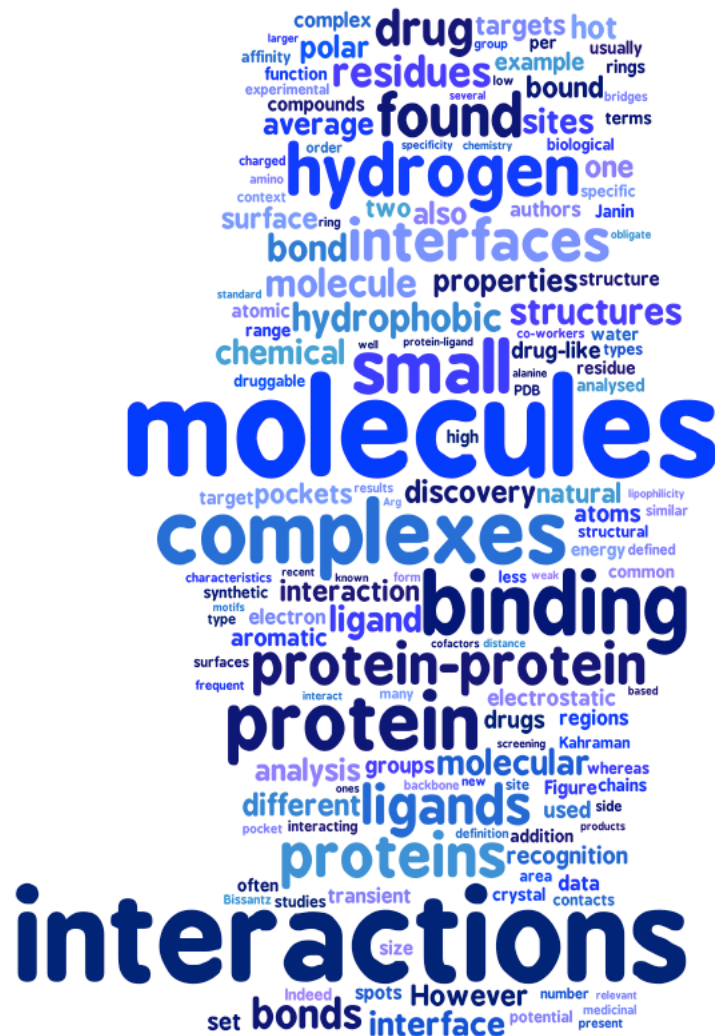
Figure 5.42. Average density of contacts per interface atom for each subset: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For small molecule complexes, both protein side (PS, pale blue) and ligand side (LS, orange) are plotted. For the PP complexes, both long chain (LC, pale blue) and short chain (SC, orange) are plotted. Error bars denote the standard error of the mean. A: density of proximal contacts (atom pairs within 4.5Å) at UniProt level of protein redundancy. B: density of successful contacts at UniProt level. C: density of proximal contacts at SCOP family level of protein redundancy. D: density of successful contacts at SCOP level.254

Figure 5.43. Examples of oral drugs binding to proteins, proximal contacts are represented by grey dotted lines. LEFT: 2HM9, dihydrofolate reductase complexed with trimethoprim, 16.3 proximal contacts per buried ligand atom. RIGHT: 3C9J, transmembrane domain of M2 protein complexed with amantadine, 4.6 proximal contacts per buried ligand atom.256

Figure 5.44. Scatter plot of the proximal contact density (the number of contacts per interacting ligand atom) versus the number of ligand atoms for the small molecule subset. The redundancy filter applied here is by distinct UniProt and distinct small molecule. Drug-like (yellow), approved drugs (cyan), oral drugs (green), natural molecules (purple), protein-protein inhibitors (magenta) and small peptides (blue). The histogram in the centre of the figure represents the molecular weight distribution.257

Chapter 1

Introduction



During my time as a molecular modeller at UCB Pharma, both my employer and I became interested in the emerging field of protein-protein interactions as drug targets. This interest is shared by many researchers in the field, as noted by the recent creation of PPI-Net, a UK network for protein-protein interactions founded by several Research Councils (<http://ppi-net.org/>). The project described in this thesis focuses on protein-protein and protein-small molecule interactions in the context of drug discovery enhanced by the structural biochemistry expertise and databases of the Department of Biochemistry.

1.1 Drug discovery

1.1.1 Decline in drug discovery productivity

The decrease in productivity in drug discovery and development (as the number of approved drugs per average R&D cost to put them in the market) is a many fold problem (Garnier 2008). It is now well documented and accepted that one of the main reasons for this decline is the poor quality of drug candidates entering into clinical trials ((Leeson *et al.* 2007; Keserü *et al.* 2009; Gleeson *et al.* 2011) and the references therein). The weakness of the current candidates can be pinned down to inadequate target selection (Paul *et al.* 2010) but also to an inappropriate profile of the chemical entities. These candidates are far too lipophilic to have good chances of success as safe drugs, as logP (octanol-water partition coefficient) correlates positively with compound promiscuity (Leeson *et al.* 2007). Lipophilicity of oral drugs is generally considered a requirement for their absorption by passive diffusion in the membranes, although there is increasing debate about the possibility of active transport. (Dobson *et al.* 2008; Sugano *et al.* 2010). Another explanation for the lipophilic trend of drug candidates is that it is a consequence of the standard medicinal chemistry practices, and these will be discussed in the next section. Several studies and opinion articles encourage medicinal chemists to keep lipophilicity as low as possible (Cooper *et al.* 2010;

Leeson *et al.* 2010; Hann 2011), as well as revising the standard medicinal chemistry settings and screening cascades that pursue maximizing affinity for single targets in isolated assays (Gleeson *et al.* 2011).

1.1.2 Medicinal chemistry practises

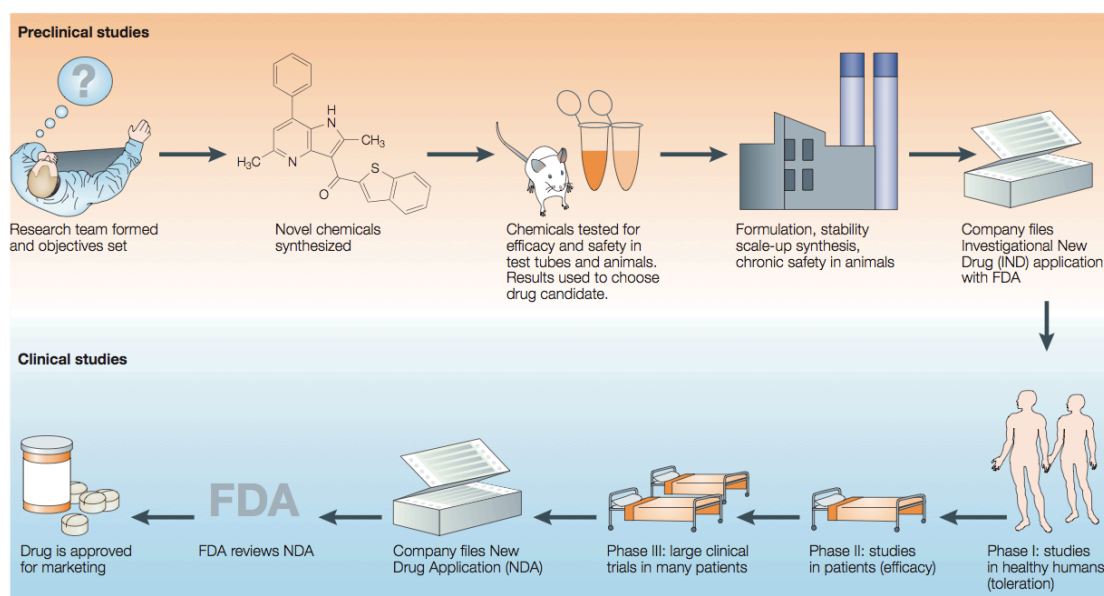


Figure 1.1. Stages of the drug discovery process. Reprinted from (Lombardino *et al.* 2004).

Keseru and Makara (Keserü *et al.* 2009) analysed the properties of hits and their follow-up leads that were published between the years 2000-2007. The authors found that fragment-based screening and natural products deliver better quality hits, in terms of low lipophilicity, than hits selected by HTS (High Throughput Screening). However, the profile of final leads was the same (high logP), whatever the starting point, highlighting the tendency of increasing potency by adding lipophilicity. Furthermore, a recent analysis by Walters *et al.* (Walters *et al.* 2011) of the molecules, published in Journal of Medicinal Chemistry between 1959 and 2009, revealed that the properties of the synthetic molecules over time, in particular lipophilicity and carbon sp³ content, have diverged from those of marketed drugs since the 80's.

Why, after so many reports analysing the properties of drugs, do we still make molecules that move away from the intended profile? Walters and co-workers argued that many of the advances applied in drug discovery are contributing to increase this tendency. For example, progress in synthetic and analytical techniques has lowered the difficulty of making and purifying bigger and more complex molecules. Development of robust scalable reactions like sp²-sp² couplings, have yielded corporate collections richer in flatter molecules and scarce in natural product-like compounds (Lovering *et al.* 2009). Improvements in formulation enable discovery projects to progress compounds with less optimised properties. However, doing so it seems we are only delaying the failure to the development phase (Hann 2011). Finally, the advances in molecular biology have led to target-based drug discovery where optimisation of the compound properties happens sequentially. Usually, primary screens are competitive binding assays of the isolated target, which in turn facilitates the increase of affinity regardless of other molecular properties; these will be optimised later in the screening cascade. Even the structure-based drug design has been partially misused, as it has encouraged medicinal chemists to target hydrophobic pockets where a burst of potency can be gained (Walters *et al.* 2011).

In the past, before the explosion of genomics and projects with defined molecular targets, medicinal chemists evolved compounds with feedback from *in vivo* primary screens (Lombardino *et al.* 2004) where most of the pharmacokinetic problems we face today were solved “on the fly” with the efficacy on animal models. Indeed, a recent analysis by Swinney *et al.* (Swinney *et al.* 2011) reported that although the widespread focus of target-based small molecule drug discovery, the majority of the first-in-class small molecule new molecular entities (NME) approved between 1999 and 2008 have been discovered by phenotypic-based approaches. Most of the self-criticisms in the field recognised that there was a “wrong turn” (Hirschler 2009) that converted the art of drug discovery into an industrial process (Garnier 2008). In particular, early stages of this process are driven by a

“perceived need for potency” (Hann 2011) and a sense of urgency to deliver leads into development. In practical terms this translates into more lipophilic candidates, as there is not much room to elaborate more risky exploratory molecules (Keserü *et al.* 2009).

To match these chemical challenges, more adventurous exploration of the chemical space is emerging, like DOS (diversity oriented synthesis) (Galloway *et al.* 2010), stapled peptides (Walensky 2004; Gavathiotis *et al.* 2008; Bird *et al.* 2010), or the rescue of NP (natural products) for drug discovery (Li *et al.* 2009; Bauer *et al.* 2010). As well as this, new technologies with microfluidics and microreactors have been developed to enable faster exploration of the biology and chemistry space of a project (Wong-Hawkes *et al.* 2007; Kang 2008). In addition, the consolidation and several successful outcomes of the fragment-based lead discovery, even for challenging targets including protein-protein interactions (Coyne *et al.* 2010), seem to have come to the rescue, at least for projects where fragment approaches can be used.

Fragment-based technologies require fragment solutions in high concentration, which in turn deliver almost exclusively polar hits (Congreve *et al.* 2008; Keserü *et al.* 2009; Ladbury *et al.* 2010). The conscious effort to optimise these hits containing hydrophobicity as much as possible is regarded as the new paradigm in drug discovery (Hann 2011). The ligand lipophilicity efficiency (LLE) index (Leeson *et al.* 2007) and other ligand efficiency indices, including polarity of molecules (Abad-Zapatero *et al.* 2010), are currently used towards this end.

1.1.3 New targets: protein-protein interactions

In parallel, researchers pursue alternative strategies to find therapeutics, one of which is a new area in drug discovery: targeting protein-protein interactions with small molecules (Wells *et al.* 2007). Multi-protein complexes orchestrate most functions in living organisms; therefore they are

attractive targets for therapeutic intervention. However, traditional drugs are taken orally and orally bioavailable drugs are usually small molecules (Lipinski *et al.* 1997). Consequently, it is usually assumed that the biological intended target must have a small pocket or cleft where our candidate drug (or lead molecule) can maximise its interactions in order to show the required high affinity.

In 2002 Hopkins and Groom coined the term “the druggable genome” (Hopkins *et al.* 2002). They defined a druggable target as a protein that is not only linked with a disease but also has a ‘beautiful’ pocket where a small drug-like molecule can bind. Classical drug targets are enzymes and receptors, usually treated as monomeric proteins with an active site for an endogenous small ligand. Moreover, the existence of these small endogenous mediators has influenced the way pharmaceutical companies classically seek hit molecules. Hit identification campaigns often rely upon competition assays and those that monitor enzymatic turnover, which can be easily scaled up for HTS, where medium or large drug-like (or lead-like) chemical libraries are screened against the biological target. In this context, protein-protein interactions have long been believed to be undruggable (Whitty *et al.* 2006). This belief has been supported by the assumption that a small molecule is unable to compete with one of the partners in a multi-protein complex, where the average surface area buried at interfaces is 2000\AA^2 with an average of 23 residues in each protomer (Janin *et al.* 2007), see Figure 1.2.

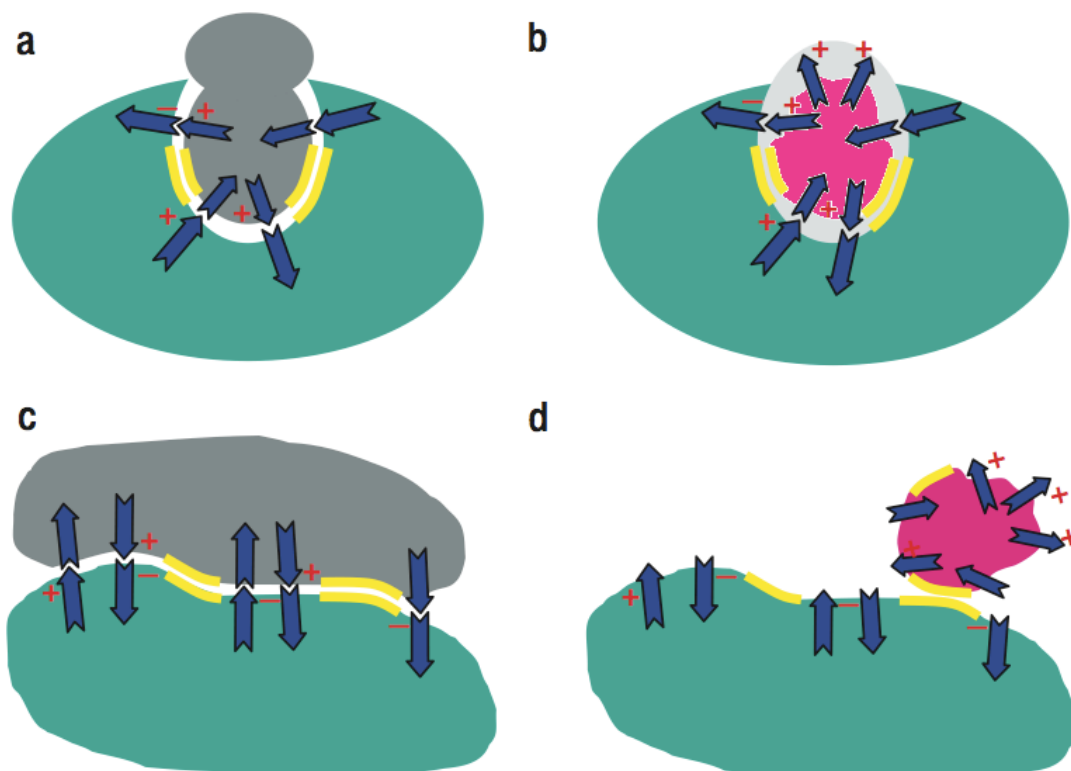


Figure 1.2. Concept of undruggable surfaces. a: Protein (green) with a cavity evolved to recognise an endogenous ligand (grey) with multiple interactions converging in a small volume. b: Classical drug target where the drug molecule (magenta) occupies the volume maximising interactions, as most of its surface is in contact with the protein target. c: Protein-protein complex (green and grey) with a large surface with spread interactions. d: Small drug molecule (magenta) cannot engage many interactions as the absence of grooves translates into small contact areas. Blue arrows represent hydrogen bonds and yellow patches represent hydrophobic contacts. Reprinted from (Whitty *et al.* 2006).

1.1.3.1 Challenging undruggability

Whitty and Kumaravel (Whitty *et al.* 2006) classified drug targets in terms of two types of risk. Biological risk accounts for the potential therapeutic effect of the modulation of the target under evaluation, and the chemical risk relates to the likelihood of finding a small molecule modulator for that target. As mentioned previously, protein-protein interactions are attractive targets for drug discovery due to their omnipresence in disease processes. In fact, particularly in the case of extra-cellular targets, antibody-based drugs are a validation of this concept (Adair *et al.* 2005). Many protein-protein interactions are considered low biological risk drug targets. The key

question is then the chemical risk for protein complexes, or in other words the probability of finding a small molecule capable of disrupting the interactions between proteins. Without considering the possibility of allosteric modulation (Christopoulos 2002), two experimental findings have lowered this chemical risk: the existence of energetic hot spots at the interfaces and site adaptability.

Hot spots

One of the most striking characteristics of the protein-protein interacting surfaces is the existence of so called "hot spots". In 1995, Clackson and Wells (Clackson *et al.* 1995), using a technique called alanine scanning mutagenesis, systematically mutated to alanine the receptor residues at the interface between the human growth hormone and its receptor and measured the energy of binding of the resulting complex mutants. In this pioneering work, the authors found that certain residues were responsible for most of the interaction energy of the complex. Many other experimental studies have proved that this is a common characteristic of almost all interfaces of the protein complexes (Reichmann *et al.* 2007). Publically accessible databases hold both experimental data for alanine scanning mutagenesis (Thorn *et al.* 2001) and computationally predicted hot spots, see for example (Guney *et al.* 2008; Segura *et al.* 2011). The accepted criteria used to define a residue as part of a hot spot is that upon its mutation to alanine, the free energy of complex binding increases by at least 2 kcal/mol. Hot spots are concentrated patches of such residues, called "hot regions" as discussed below.

Bogan and Thorn (Bogan *et al.* 1998) analysed datasets from alanine scanning mutagenesis experiments and found that all the hot spots share common characteristics. Their work led them to postulate the "O-ring" hypothesis for hot spot residues in protein-protein binding interfaces. Energetically, hot spot residues are usually clustered at the centre of the

interface and are surrounded by energetically neutral residues. The role of these neutral residues is to shield the hot spots from the solvent by creating a microenvironment around the hot spot with a lower dielectric constant, thus enhancing electrostatic interactions and reducing the desolvation cost of binding. It is no surprise then, that the most frequent hot spots residues Trp, Tyr and Arg are capable of both hydrophobic and electrostatic interactions. Bogan and Thorn also found that hot spots are self-complementary across the interfaces (Bogan *et al.* 1998).

Nussinov and co-workers (Keskin *et al.* 2005) found that hot spot residues, identified by experimental alanine scanning, tended to be evolutionarily conserved. They went on to study the organization of hot spots identified by sequence analysis and found that they are not evenly distributed in the interface as they cluster together in "hot regions". These areas are tightly packed and within a region, hot spots form networks of cooperative interactions. In contrast, the contribution to the global energy of binding is additive between hot regions. In addition, clustered hot spots in dense hot regions mean that the removal of water molecules is easier, strengthening the electrostatic interactions in a similar same way to the O-ring arrangement. Furthermore, these regions are more rigid as they are densely packed and therefore pay a lower entropy penalty upon binding, whereas non-optimal packed regions are responsible for site flexibility.

In conclusion, protein-protein interactions are locally optimised in these hot regions, whereas the rest of the interface is less specific. This fact could explain the diversity in protein binding partners often accepted at a particular interface (Keskin *et al.* 2005). Furthermore, the existence of these locally optimised regions, responsible for most of the binding energy between proteins, makes competitive small molecules more credible. Indeed, several studies report the first small molecules interfering with protein-protein interactions, and this will be the focus of the chapter 2 of this thesis.

Site adaptability

According to one of the principles of druggability described previously (Hopkins *et al.* 2002), biological targets need to present pockets or clefts in order to accommodate small molecule drugs. Interfaces of protein complexes are usually relatively flat. Nevertheless, structural evidence of flexible adaptability in these regions (for instance in IL-2 (Arkin *et al.* 2003; Thanos *et al.* 2006)), opens the prospect of the existence of more druggable protein complexes as targets. Indeed, druggability predictions are dependent on the flexibility of the target, as Brown and Hajduk showed (Hajduk *et al.* 2005; Brown *et al.* 2006).

Recent analyses of the protein-protein interfaces inhibited by small molecules have suggested that this adaptability occurs mainly through the flexibility of side chains (Fuller *et al.* 2009; Bourgeas *et al.* 2010). Although flexibility is fundamental to molecular recognition and is a key factor to consider in the quest to find small molecule drugs to modulate protein-protein interactions, it is still difficult to predict. However, increasing computing power is making longer molecular dynamic simulations feasible, and several bioinformatic tools are being used to evaluate plasticity in proteins (Gonzalez-Ruiz *et al.* 2006).

1.2 Molecular recognition from atomic interactions

In an editorial in the *Journal of Molecular Recognition*, van Regenmortel defined molecular recognition as “the non-covalent specific interaction between two or more biological molecules” (van Regenmortel 1999). The van Regenmortel perspective, however, emphasised the limitations of static structures to explain dynamic activities between biomolecules; and how molecular recognition is a “mutual adaptation” rather than a frozen lock-and-key model. Furthermore, biological interactions are cooperative (positively or negatively) and rarely additive (Williams *et al.* 2004) i.e. the final outcome is rarely the sum of their parts.

In addition, it is worth remembering the inherent limitation of crystal structures (85% of the content of the PDB (PDB Team 2011)), which are the interpretation of the experimental diffraction patterns of a crystallised sample (Davis *et al.* 2003). In turn, only a portion of molecules will be amenable to be crystallised, and if they are, the conformations in the crystal lattice might not be biologically relevant (Acharya *et al.* 2005), although it is worth noting that crystal structures are “wet”, as most crystals have 35 to 70% of their content as solvent.

In the case of small molecules, refinement methods developed for proteins are used to fit the electron density of the ligand with an accuracy that is difficult to assess (Böhm *et al.* 1996). It is clear then, that it is not possible to determine the fundamental laws of molecular recognition from the current atomic models. However, insights can be gained from the characterised structures. Going back to van Regenmortel’s definition, molecular recognition derives from the non-covalent interactions between the molecules involved. In this way, trends in these non-covalent interactions can

be elucidated between different types of molecules with the aim of identifying possible different modes of recognition.

1.2.1 Non-covalent forces

Although conformational adaptability, long-range interactions, solvation and desolvation processes are key components in the binding event, they are not discussed here. I will focus in the specific non-covalent atomic intermolecular interactions between binding partners.

1.2.1.1 van der Waals attractions

The transient polarization of the electron cloud of a nonpolar atom will induce in turn an opposite polarization in the nearby nonpolar atom, which will create a tiny attraction force between them, known as London dispersion force. Although small, these attractions sum up to a significant interaction at interfaces where two molecules are close together (Voet *et al.* 1992). These forces are distance dependent and they will become repulsive if the two entities are too close together due to steric hindrance of the electron cloud. The physical model commonly used to describe this behaviour is the Lennard-Jones 6-12 potential:

$$V(r)_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right]$$

where r is the distance between two atoms, ε and σ are two constants defined by the system. Figure 1.3 shows that the potential becomes increasingly repulsive for close distances due to the first term and attractive for an optimal range, in which atomic van der Waals radii are derived from experimental structures.

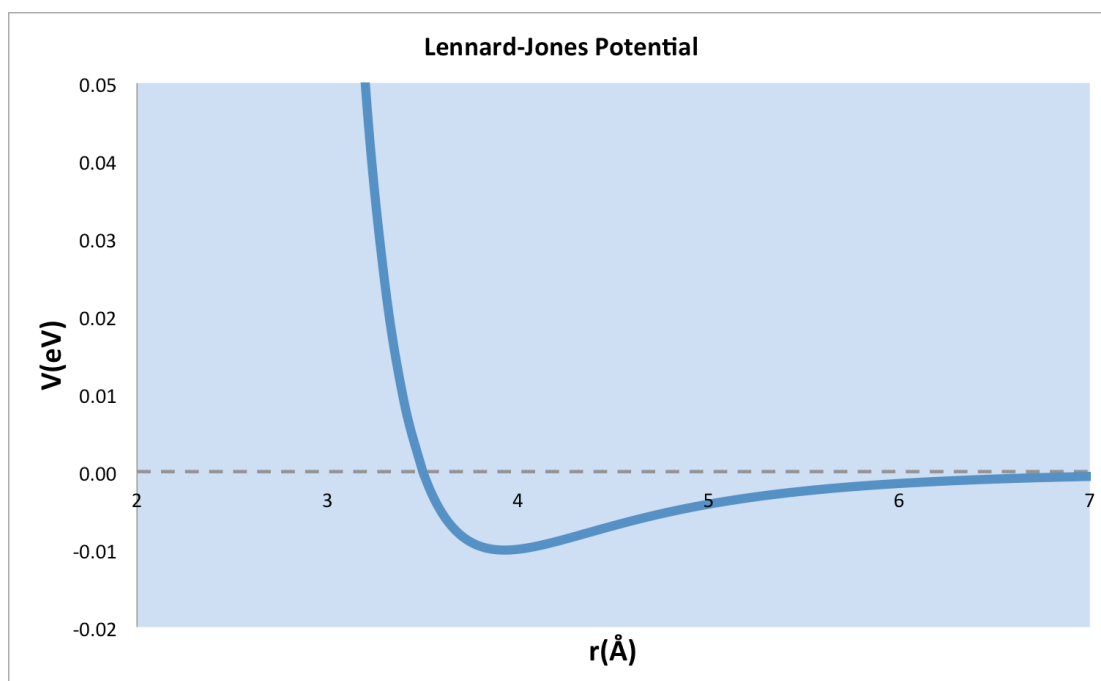


Figure 1.3. Lennard-Jones potential (V) for Argon dimer as function of the interatomic distance (r). The minimum of potential corresponds to ϵ and potential is equal to zero when the distance is σ .

1.2.1.2 Hydrogen bonds

I quote here the new definition of the hydrogen bond recommended by IUPAC (Arunan *et al.* 2011): "The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation". This definition also comprises the weak hydrogen bonds described later. Classical hydrogen bonds are highly directional polar interactions between two electronegative atoms sharing a hydrogen. The usual geometrical ranges to identify a hydrogen bond are (McDonald *et al.* 1994):

$$\text{Distance}(\text{HBD}, \text{HBA}) < 3.9 \text{ \AA}$$

$$\text{Distance}(\text{H}, \text{HBA}) < 2.5 \text{ \AA}$$

$$\text{Angle}(\text{HBD}, \text{H}, \text{HBA}) > 90^\circ$$

Hydrogen bond strengths depend on the electronegativity of the heavy atoms involved and on the environment where the atoms are located, as well as the surrounding hydrogen bond network. Due to the restricted directionality of this type of interaction, hydrogen bonds play an important role in molecular recognition giving specificity to the binding event. However, they don't usually contribute much to the free energy as the desolvation of both donor and acceptor compensate the hydrogen bond formation energy. Estimates based on ITC data and burial of polar surface, range from 4-11KJ/mol (Olsson *et al.* 2008). Regarding protein structure, the NH and CO backbone groups are usually forming hydrogen bonds (McDonald *et al.* 1994) (in fact virtually all buried NH form hydrogen bonds), and are usually positioned correctly with respect to each other, especially in high-resolution crystal structures. It is also found that higher losses in affinity occur on ligand binding when removing a hydrogen bond from a backbone NH than a backbone CO. For example, in kinase inhibitor crystal structures only one structure is reported with an orphan NH in the hinge region, whereas it is more common to have the CO unpaired (Bissantz *et al.* 2010).

1.2.1.3 Weak hydrogen bonds

As the broad IUPAC definition describes (Arunan *et al.* 2011), the ability to share a hydrogen is not limited to strong electronegative atoms (N and O). It has become apparent in recent years that weak hydrogen bonds do occur in protein structures and protein-ligand binding. Weak hydrogen bond donors are polarized C-H, C_{alpha}-H and NH in proteins, whereas weak hydrogen bond acceptors are the π orbitals of aromatic rings. In addition, interactions between CF and XH (X= N,O) and C-H engaged with O and N in aromatic heterocycles are also observed. Analysis of the CSD (The Cambridge Structural Database - The world repository of small molecule crystal structures, <http://www.ccdc.cam.ac.uk/products/csd>) and the PDB (The Protein Data Bank - An Information Portal to Biological Macromolecular Structures, <http://www.pdb.org>) shows that although interactions between

XH (X= N,O) and weak acceptors (π orbitals of aromatic rings or donor- π) are observed, they are very rare (Bissantz *et al.* 2010).

1.2.1.4 Ionic interactions

Ionic interactions are electrostatic attractions between atoms with opposite charge. In an aqueous environment, these attractive forces are attenuated by the water molecules interacting with the charge (Voet *et al.* 1992), or in other words the high dielectric constant of the water diminishes the attraction force between two opposite charges following Coulomb's law. These interactions are only distance dependent and do not have preferred geometries. Indeed, analysis of relative geometries of charged side chains for NMR ensembles of 11 non-homologous proteins show clear distance dependency for each ion-pair type (Kumar *et al.* 2002), whereas the relative orientation is spread across the whole range, see Figure 1.4.

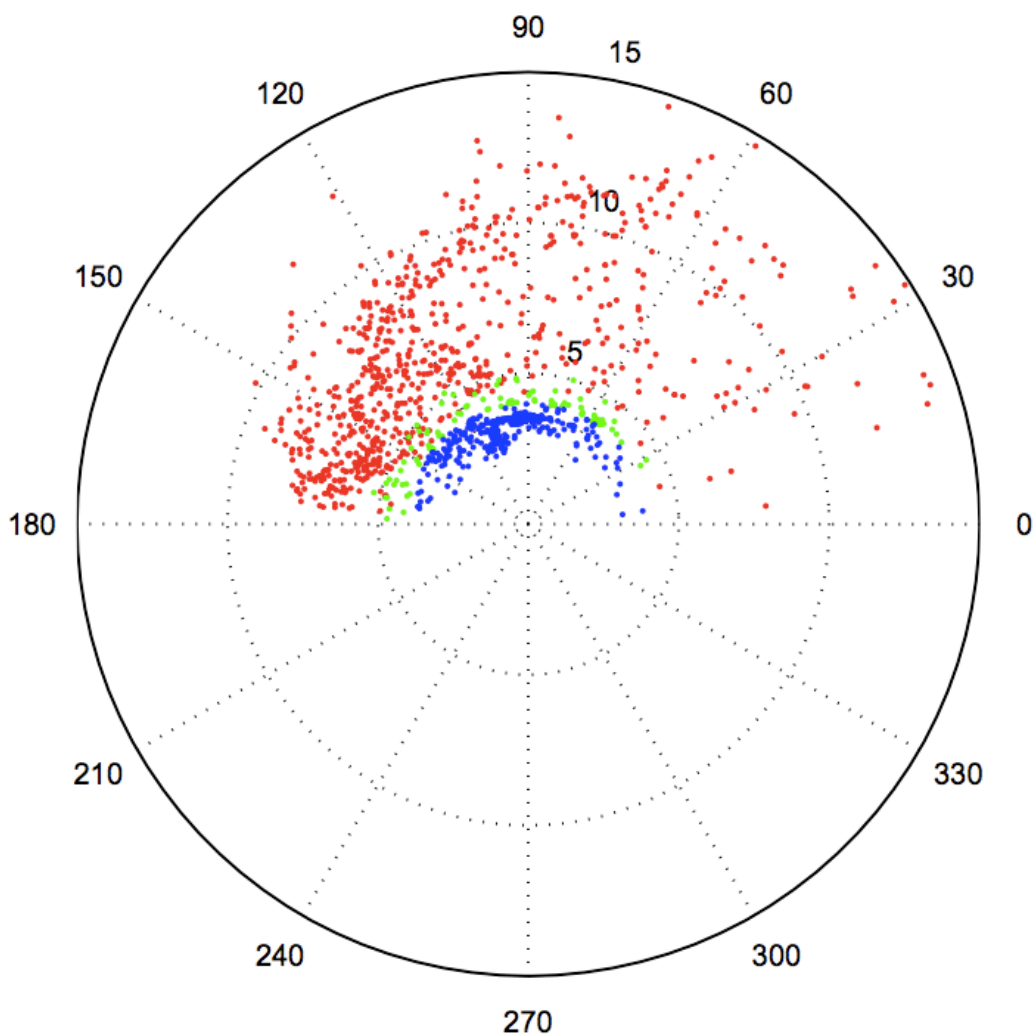


Figure 1.4. Relative geometries of side chain charged groups in proteins. Colour coded per ion-pair type: Salt bridges (blue, side chain centroids and O-N pairs from Glu/Asp-Arg/Lys/His are within 4Å), N-O bridges (green, only O-N pairs are within 4Å, but not the side chain centroids) and longer range ion pairs (red, centroids and N-O pairs more than 4Å apart). Geometry of the ion-pairs is represented by the distance between centroids of the charged groups (radii of the polar plot) and by the relative angular orientation between side chains (angle of the polar plot measured as the angle between the vectors formed by the C-alpha and the side chain centroid of each residue). Most ion pairs with distances $\leq 5\text{\AA}$ are stabilizing of the structure and destabilizing for longer distances. See original paper for details. Reprinted from (Kumar *et al.* 2002).

1.2.1.5 Hydrophobic interactions

The generic definition for hydrophobic interaction is the preference of nonpolar regions to pack closely together instead of interacting with water. According to this definition, the energy is gained by displacing water

molecules and therefore this interaction is entropy driven and not specific. However, Bissantz et al (Bissantz *et al.* 2010) discuss several examples in their review, arguing that there is more to it than solely displacement of water molecules. Indeed, protein surfaces or ligand chemical groups are not in a binary scale of polar and hydrophobic; instead they present a continuum of polarizability, where shape matching and close contacts will contribute enthalpically to the free energy of the association. Furthermore, several studies (reviewed at (Bissantz *et al.* 2010)) suggest that part of the affinity gained by filling a protein hydrophobic pocket is due to the poorly solvated state of the pocket in the apo form, where water molecules are rarely detected as they are not positionally fixed or not making many hydrogen bonds. These examples highlight the difficulty to deconvolute the binding energy into independent non-cooperative contributions.

1.2.1.6 Aromatic interactions

Aromaticity is a chemical property found extensively in natural and synthetic molecules. All five nucleotides and four of the 20 standard amino acids have an aromatic ring in their side chains. In medicinal chemistry, aromatic rings are habitual components of drug-like molecules (Pitt *et al.* 2009). Aromatic rings are planar structures with $4n+2$ ($n = 0,1,2\dots$) delocalised π electrons (Hückel rule). The delocalised π system has maximum electron density on each side of the ring and a minimum in the ring itself, which translates into a small positive partial charge in the peripheral hydrogens. Indeed, aromatic protons are significantly deshielded and present greater NMR (Nuclear Magnetic Resonance) chemical shifts than standard sp^2 hydrogens. These special shapes and electronic properties guide their interactions to specific geometries. In protein structures, there is a wide range of orientations that depend on the residue type and local environment, although typically the preferred orientation is displaced parallel stacking of the rings (off-centred) followed by edge-to-face or T-shape (Chakrabarti *et al.* 2007).

For ligands bound to proteins, the relative geometry of the aromatic rings depends on the substituents of the ring that confer specific electron properties to the conjugated system. Thus, electron poor rings (with electro withdrawing groups) interact with electron rich (with electro donating groups) in a preferred stacking geometry. Electro withdrawing groups in para or ortho position to a CH, make this hydrogen more acidic, favoring the T-shape arrangement with another ring. Regarding heteroaromatic rings, the preferred geometries follow the alignment of partial charges in the ring depending on its composition and substituents (Bissantz *et al.* 2010).

1.2.1.7 pi-cation interactions

Due to the electron density of the delocalised π system, attractive forces occur between a cation and the face of an aromatic ring. Singh and Thornton identified this type of interaction during a study of interactions between all residue types in proteins (Singh *et al.* 1992). Gallivan and Dougherty (Gallivan *et al.* 1999) further studied pi-cation interactions and found they were common in high-resolution protein structures, with Trp and Arg being the two residues with highest propensity to engage a pi-cation arrangement. Several examples have also been reviewed for this type of interaction in protein-ligand complexes (Bissantz *et al.* 2010).

1.2.1.8 Halogen bonds

Halogens are common components in drug-like molecules, they are used to fine tune electrostatic properties of aromatic rings, to fix optimal conformations adding steric impediments, to increase metabolic stability by blocking reactive positions and to modulate lipophilicity. Traditionally, halogen interactions have been considered mainly as van der Waals, hydrophobic by water displacement and shape complementarity, especially for the heavier halogens with the softer electron cloud. Recently, the importance of halogen bonds as weak but specific interactions is gaining relevance in drug design.

Heavy halogens (Cl, Br and I) bound to carbon present a small positive electrostatic potential opposite to the sigma bond (Bissantz *et al.* 2010), which interacts favourably, for example with the oxygen atom of the carbonyl backbone with specific geometry (linear C-X...O=C). Conversely, as discussed for weak hydrogen bonds, fluorine acts as hydrogen bond acceptor with polar hydrogens HX (X = N,O).

1.2.1.9 Sulphur interactions

In protein structures the sulphur of the Cys can form a covalent disulphide bond with other Cys by oxidation of both atoms. In free form, the sulphur of the Cys is a hydrogen bond donor and can form hydrogen bonds especially with the backbone carbonyl oxygen (Zhou *et al.* 2009). Sulphur atoms from Met residues have a dual behaviour as they can interact with electron rich and electron poor groups. Indeed, methionine interacts with aromatic rings through both the face (electron rich) and the edge (electron poor) of the ring (Pal *et al.* 2001). Analysis of the CSD and PDB of ligands containing sulphur, highlight the versatile interaction pattern of the sulphonyl moieties, as they can act as weak hydrogen bond acceptors and hydrophobic groups (Bissantz *et al.* 2010).

1.2.2 Structural characteristics of protein-protein complexes

In the quest to understand and predict protein-protein interactions, in particular interaction sites, the structural analysis of known complexes is key. However the diversity, both in terms of function and constituents, makes the task of finding universal rules for protein complexes an extraordinarily difficult one (Reichmann *et al.* 2007). It has also been argued that the small number of protein complexes analysed so far leads to contradictory conclusions (Ofra *et al.* 2003). It may be that a systematic distinction regarding function, constituents and lifetime of complexes is needed to reduce noise in

experimental structural data, together with an increase in the number of complexes structurally resolved.

1.2.2.1 Functional protein complexes

The huge versatility and often overlapping functions of proteins in the cell make their classification difficult. There is not a definitive consensus regarding functional classification of proteins. In general terms, they can be categorised as enzymes, hormones, receptors, antibodies, structural proteins, motor proteins, transport proteins, signalling proteins and storage proteins (Ruzheinikov 2007). Structural studies of protein-protein complexes typically split them into four general groups (Jones *et al.* 2000; Cho *et al.* 2006; Janin *et al.* 2007).

The first group comprises antibody-antigen complexes. Antibody structures contain six CDR (complementary-determining regions) that identify the protein-antigen with high specificity (Braden *et al.* 2000); these regions are highly variable yet enriched with serines and tyrosines (Livesay *et al.* 2004; Birtalan *et al.* 2008). Antibody-antigen interfaces are of standard size, ranging from 1200 to 2000Å² (Janin *et al.* 2007).

The second group of protein complexes consists of enzyme-inhibitor assemblies. These complexes can be further divided into two subsets depending on their interface size, standard (1200-2000Å²) or bigger (>2000Å²) (Chakrabarti *et al.* 2002). Usually, standard interfaces show a single recognition patch, whereas the larger interfaces present more than one recognition site (Chakrabarti *et al.* 2002).

Electron-transfer complexes comprise the third group of protein-protein interactions. These complexes have a short half-life and low affinity and it is therefore difficult to obtain them in crystal form. Most of the few electron-

transfer proteins characterised so far have interfaces of 900-1200Å² in size (Mathews *et al.* 2000).

The last group can be described generically as comprising complexes taking part in signal transduction and cell cycle regulation, such as G-proteins and protein-receptor assemblies. These complexes exhibit exquisite sensitivity to changes in the environment, usually forming transient interactions and presenting low-to-medium affinity range (low mM to high nM) (Hyvönen *et al.* 2000; Janin *et al.* 2007).

Although this categorisation is often useful, especially as the great majority of complexes are enzyme-inhibitor and antibody-antigen, this is generally unsatisfactory as it is mostly a reflection of what has been feasible to study. For example structural proteins are not well represented and not included in the categories.

Specific vs. crystallographic complexes

X-ray crystallography provides the majority of the experimental structures of protein complexes. However, the distinction between functional complexes and crystallographic artefacts must be drawn in order to extract information pertaining to specificity, evolution and function. Artificial crystal contacts can occur simply as a result of the protein packing in the crystals. The task of identifying these unnatural contacts is far from trivial and automatic classification is still an open challenge. Nevertheless, the size and composition of interfaces is a useful guide to identifying the correct interface. These predictions can be improved if the sequence conservation of related proteins and estimates of the stability of the predicted assembly are utilised. The PISA resource (Protein Interactions, Surfaces and Assemblies, http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html, (Krissinel *et al.* 2007)) is an example of automated software that predicts quaternary structure from estimation of its thermodynamic stability. Janin and co-workers compared a

set of specific interactions (without including short-lived assemblies or electron-transfer complexes) to a set of non-specific interactions (Bahadur *et al.* 2004). In this comparison, crystal contacts were found not only to be smaller than specific ones, with average interface area of 570\AA^2 , but also contained fewer hydrogen bonds per unit surface area. In addition, interfaces between monomers in crystals were less closely packed than interfaces between protomers in multimeric complexes.

1.2.2.2 Constituents and lifetime of protein complexes

With regard to the constituents and the lifetime of the protein-protein complexes, Nooren and Thornton (Nooren *et al.* 2003) suggested three ways to classify protein-protein interactions. The first divides complexes into those that are homo-oligomeric (composed by identical chains) or hetero-oligomeric (non-identical chains). Homo-oligomers can be further sub-divided into those that are isologous, where interfaces are composed of the same region from each protomer and those that are heterologous, where protomers interact through different regions. Heterologous homo-oligomers can either form a cyclic structure or aggregate into an endless repeated structure. The second distinction that Nooren and Thornton considered is whether the protomers forming the complexes can exist independently *in vivo*. An obligate complex has to be denatured in order to dissociate, whereas a non-obligate complex is formed by stable self-standing monomers. Examples of non-obligate assemblies include antibody-antigen, enzyme-inhibitor and signal transduction complexes. The third division is by complex lifetime; one can distinguish between permanent and transient interactions *in vivo*. Usually obligate interactions are permanent, like most homodimers, whereas transient interactions present a whole range of affinities and kinetics. The authors emphasised these classifications aren't discreet absolute values, but a continuum in the scales of lifetime and stability, see Figure 1.5. Nevertheless, these definitions could be important tools in the quest to understand protein-protein interactions. For example, it is apparent that interfaces of permanent

complexes are more similar to those in the protein interior than on the surface. In addition, permanent interfaces tend to be dryer, more hydrophobic and larger than the interfaces of transient complexes (De *et al.* 2005; Janin *et al.* 2007). However, as mentioned before, the vast diversity of function, flexibility, affinity and specificity of protein assemblies is difficult to capture in a set of general rules.

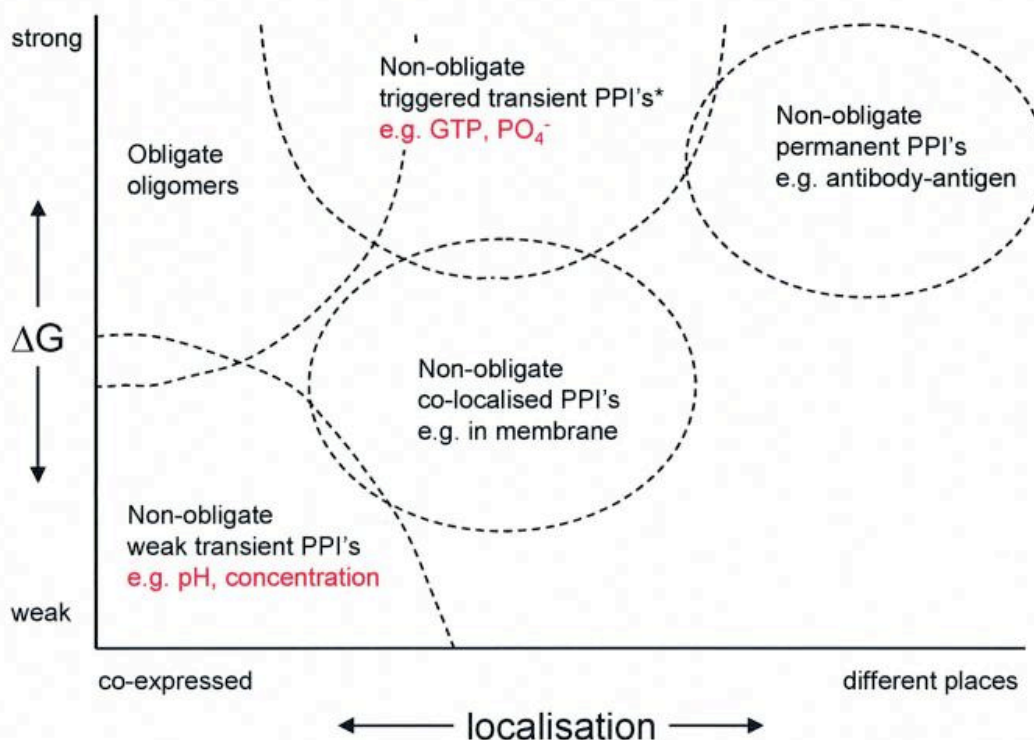


Figure 1.5. Definition of types of protein-protein interactions by function of their binding affinity (Y axis) and the localisation of the protomers (X axis). In red the factors that affect transient interactions. * denotes large conformational changes that usually occur with the association. Reprinted from (Nooren *et al.* 2003).

1.2.2.3 Descriptors and topology of protein-protein interfaces

Classical computational characterisation of interfaces includes size, shape, packing, electrostatic interactions, amino acid composition and amino acid pairing preferences.

Size

The size of an interface is commonly expressed as the change in the ASA (solvent-accessible surface area) between the monomers/protomers and the complex. For example, for a hetero-dimer, the interface size B , is $B = ASA1 + ASA2 - ASA12$ (Janin 2000). Some authors prefer to report $B/2$ in spite of the fact the ASA is not exactly the same for both surfaces unless they are completely flat. The average size of protein complex interfaces is between $1200-2000\text{\AA}^2$ with an average of 23 residues in each protomer (Janin *et al.* 2007). Richard Bickerton's analysis of the PICCOLO database (Bickerton 2009) found that the average size of the protein-protein interfaces is greater than previously reported. The improvement of the structural characterisation techniques allows larger complexes to be resolved. Bickerton found that the average of interface size is $2400 \pm 1900\text{\AA}^2$. Looking at the type of interface, obligate complexes interact on average through larger interfaces than the transient ones.

Shape

Interacting protein surfaces are usually flat overall, but examples of concave-convex interfaced have been found. In these cases, generally, the smaller partner shows convexity, binding to the concave cavity on the bigger component. An exception to this trend is the antibody-antigen complexes where the antigenic site is generally convex independent of antigen size (Janin *et al.* 2007). For large interfaces ($> 2000\text{\AA}^2$) it has been found that the binding site is closer to the centre of mass of the protein than the average location of the surface (Nicola *et al.* 2007).

Packing

Packing density is another measured structural feature of the interacting protein surfaces. This measure is used to estimate the degree of steric complementarity of monomers. The most reported packing indices are

Shape Complementarity score (Sc) (Lawrence *et al.* 1993) and Gap Volume index (GV) between proteins (Laskowski 1995). It has been found that homodimers, enzyme-inhibitor and permanent hetero-complexes are more closely packed than antibody-antigen and transient hetero-complexes (Jones *et al.* 2000).

Electrostatic interactions

It is known that electrostatic complementarity between partners in the complexes confers specificity (Jones *et al.* 2000). On average, there is one hydrogen bond per 200\AA^2 of interface area (B) (Janin 2000). Typically, obligate protein complexes have fewer intermolecular hydrogen bonds per buried ASA than non-obligate complexes with 0.9 HB per 100\AA^2 in homodimers, compared to enzyme-inhibitor complexes, which have 1.4 HB/ 100\AA^2 and antibody-antigen with 1.1 HB/ 100\AA^2 (Jones *et al.* 2000). Additionally, protein-protein interfaces have water-mediated hydrogen bonds, which present the same average distribution as the direct protein-protein hydrogen bonds, that is 10 water molecules per 1000\AA^2 (B/2). However, these waters are not always evenly distributed across the interface; in fact interacting surfaces present a whole topology range of dry/wet patches (Rodier *et al.* 2005). Salt bridges or hydrogen bonds involving at least one charged residue do occur; Lo Conte *et al.* found that 30% of the hydrogen bonds in their data set that occur at the interfaces were salt bridges (Lo Conte *et al.* 1999). However, almost half of the homodimeric structures analysed do not have this type of interaction (Jones *et al.* 2000). Disulphide bonds can be also found between interacting proteins but they are rare (Jones *et al.* 2000).

Amino acid composition

Analysis of the amino acid composition of protein-protein interfaces and pairing preferences between chains have demonstrated different

frequencies that are probably due to the different datasets used and how the interface is defined (Ofran *et al.* 2003; Headd *et al.* 2007). Ofran and Rost (Ofran *et al.* 2003) divided their data set into six different types of protein-protein interfaces. While they found each interface had its own residue propensities, there were some generalities. For example, lysine was found to be underrepresented in all types of interfaces, whereas arginine was overrepresented. However, arginine is common on all protein surfaces, not only protein-protein interfaces (Janin *et al.* 2007). Large hydrophobic amino acids were found to be favoured in all interfaces (His, Met, Try), whereas Ser, Ala and Gly were underrepresented. They corroborated previous findings that hydrophobic residues were more frequent at homo-multimers than hetero complexes. However, when they further divided their dataset into transient and obligate interaction, this distinction no longer held (Ofran *et al.* 2003). Nevertheless, Bickerton (Bickerton 2009) found in his analysis that the obligate interfaces are more hydrophobic than the transient ones. Bickerton's findings also highlight the parallelism between the interface core and the protein core, and the interface periphery and the exposed protein surface. The core of the interface is more hydrophobic than the interface periphery; it is enriched with hydrophobic residues (Ile, Val, Leu, Phe, Met and Ala) and depleted of polar and charged residues (Asp, Gln, Asn, Glu, Lys and Arg).

Pairing preferences

With respect to the residue interactions at protein-protein interfaces, Ofran and Rost (Ofran *et al.* 2003) found hydrophobic-hydrophilic contacts were prevalent at intra-domain, inter-domain and transient hetero-complex interfaces; disulfide bridges occurred more often than expected; salt bridges were less frequent at homo-complexes interfaces and identical amino acid interaction was favoured by obligate homo-complexes. Additionally, Headd *et al.* (Headd *et al.* 2007) studied 135 transient hetero complexes and found that 32% of contacts at the interfaces are formed by interactions involving backbone atoms. After separating backbone from side chain atoms and

calculating relative frequencies (both per residue count and area-weighted per residue at the interface), they found Glu, Ser, Asp, Lys and Arg were the most frequent interacting side chains at the interface, each forming more than 7% of contacts. Whereas Met, Cys, Trp and His were the least frequent with less than 3.5%. In this data set, the most frequent occurring amino acid pairs are salt bridges (Glu-Arg, Asp-Arg, Glu-Lys and Asp-Lys, when only side chains are taken into account and they are weighted by the area they occupy). This evidence highlights the importance of electrostatic complementarity between interacting surfaces, at least for the dataset analysed. After the charge-charge interactions, the next most frequent interactions are Tyr with Arg, Asn, Lys and Glu, followed by Arg with Trp and Asn. Similar results were found by Bickerton (Bickerton 2009), namely hydrophobic interactions, salt bridges and disulphide bonds are important in macromolecular recognition. Pairing preferences are normalised by residue abundance in the data set and also by solvent accessible area per residue. These show that hydrophobic residues favour other hydrophobic ones and avoid polar and charged residues. Aromatic residues prefer other aromatic or hydrophobic residues, although they also engage pi-cation and NH-aromatic interactions. Prolines interact significantly more with aromatic than other residue types. Positive charged residues (Arg, Lys and His) favour negative charged ones (Glu and Asp) but Arg-Arg, His-His and Arg-His are also common due to the versatile capability of these side chains: aromatic interactions, pi-cation and hydrogen bond (with the main chain atoms).

1.2.3 Structural characteristics of protein-small molecule complexes

In a similar way to protein-protein complexes, the immense diversity both in terms of chemical composition and function of the ligands bound to proteins makes it virtually impossible to find general rules for the characteristics of protein-small molecule complexes. Indeed, disparate results are found depending on the type of molecules studied, the accepted level of

redundancy and the size of the sample analysed. This section starts with a broad classification of the types of small molecules found in biology, and follows with a brief summary of the relevant studies published to date regarding molecular recognition for protein-small molecules from experimental structures. For convenience, the section is further divided by the type of molecules considered in these studies.

1.2.3.1 Classification of small molecules bound to proteins

In the context of therapeutic applications, perhaps the broadest classification that one can make is to distinguish between natural small molecules that are products of evolutionary selection and synthetic small molecules produced in a lab.

1.2.3.1.1 Natural molecules

In general terms, natural molecules are produced by living organisms and they are the result of evolutionary selection. Therefore, they sit in the biologically relevant section of the chemical space (Koch *et al.* 2005). However, their production “in situ” often does not confer them with the appropriate properties to cross membranes and distribute elsewhere in the organism. Nevertheless, many natural molecules produced in one organism can be active in another. For example, hormones, therapeutically used plant extracts or penicillin, just to mention a few. The KEGG resource provides a classification of “Compounds with biological roles” from the KEGG BRITE hierarchies (Kanehisa *et al.* 2008). These include carbohydrates (including lipids), nucleic acids, peptides, cofactors, steroids, hormones and transmitters, phytochemical compounds (biological active molecules from plants), marine natural products and antibiotics. In addition, KEGG RPAIR (reactant pairs) labels compounds as substrate or product when they are involved in enzymatic transformations.

1.2.3.1.2 Synthetic small molecules

Synthetic molecules do not have the many millennia to evolve but they are the products of a vast range of chemical transformations and starting materials. In the context of drug discovery, the interest is centred on synthetic molecules with drug-like properties. However, the definition of drug likeness is far from trivial, and typically involves a range of molecular properties derived from known drugs. The pioneer Lipinski's "rule of five" (Lipinski *et al.* 1997) set simple ranges of molecular weight, partition coefficient and hydrogen bond features count, for molecules with increasing likelihood of being absorbed (a crucial characteristic oral drugs must have). Recently, and especially for new targets, it has been argued that more adventurous exploration of the chemical space may be needed, in particular regions sampled by natural products (see for example (Dobson 2004; Bauer *et al.* 2010)). Several studies compare properties of natural products with drugs and synthetic drug-like molecules, (see for example (Feher *et al.* 2002; Singh *et al.* 2009)) and all conclude that natural molecules occupy a different region of the chemical space to that occupied by synthetic molecules. In particular, drug-like molecules are more hydrophobic, have less 3D and stereochemical complexity and have more aromatic rings than natural molecules. These properties may reflect the characteristics a synthetic molecule must possess to overcome all the hurdles before reaching its target in the body. However, these properties are also influenced by the drug discovery settings where these molecules are generated, as we have seen in previous sections.

1.2.3.2 Peptide binding sites

Peptides bound to proteins have been studied mainly in the context of short linear motifs (SLM, LM or SLiMs). These motifs are defined as short regulatory modules, around ten contiguous residues, which are recognised by globular domains in transient manner with typically low affinities (Dinkel *et al.*

2011). These modules are often part of a larger usually disordered structure, but can bind to their globular targets often adopting an organised structure upon binding, often described as concerted folding and binding. The importance of these motifs has been recognised in recent years, not only for the crucial role they play in cell function but also for being attractive drug targets (Blundell *et al.* 2006; Neduva *et al.* 2006). Stein and Aloy (Stein *et al.* 2008) analysed SLiMs found in the PDB using computational alanine scanning. The authors differentiate between the residues in the motifs as defined by the ELM (eukaryotic linear motif database) and the residues (they called them context) that interact with the globular domain but are not defined in the motif pattern. Computing the contribution to binding for each residue, Stein and Aloy found that the amino acids in the motifs are optimised for maximal affinity while the residues in “the context” form suboptimal interactions, however they are most likely to be crucial for specificity. Their argument was that motifs are large enough to secure binding but too small to justify the exquisite specificity *in vivo*.

1.2.3.3 Nucleotide and natural molecule binding sites

Kahraman *et al.* challenged the common assumption that different proteins binding similar ligands would have similar binding sites in terms of physical and chemical properties (Kahraman *et al.* 2007; Kahraman *et al.* 2010). In order to address this question, a special data set was manually collected. Biological relevant protein structures were selected from the PQS resource (protein quaternary structure) binding to its cognate ligand. Proteins were defined as belonging to a distinct CATH homologous superfamily and the resolution of the crystal structure was used to select the representative structure from each family. For each ligand found, only those bound to at least five distinct proteins were kept. On this basis, one hundred protein-ligand complexes were selected comprising ten different chemical ligands (Figure 1.6). These ligands are common natural substrates, products and cofactors. Therefore, in order to avoid generic misleading statements, the

results derived from this analysis have to be kept within the functional context where the molecules operate.

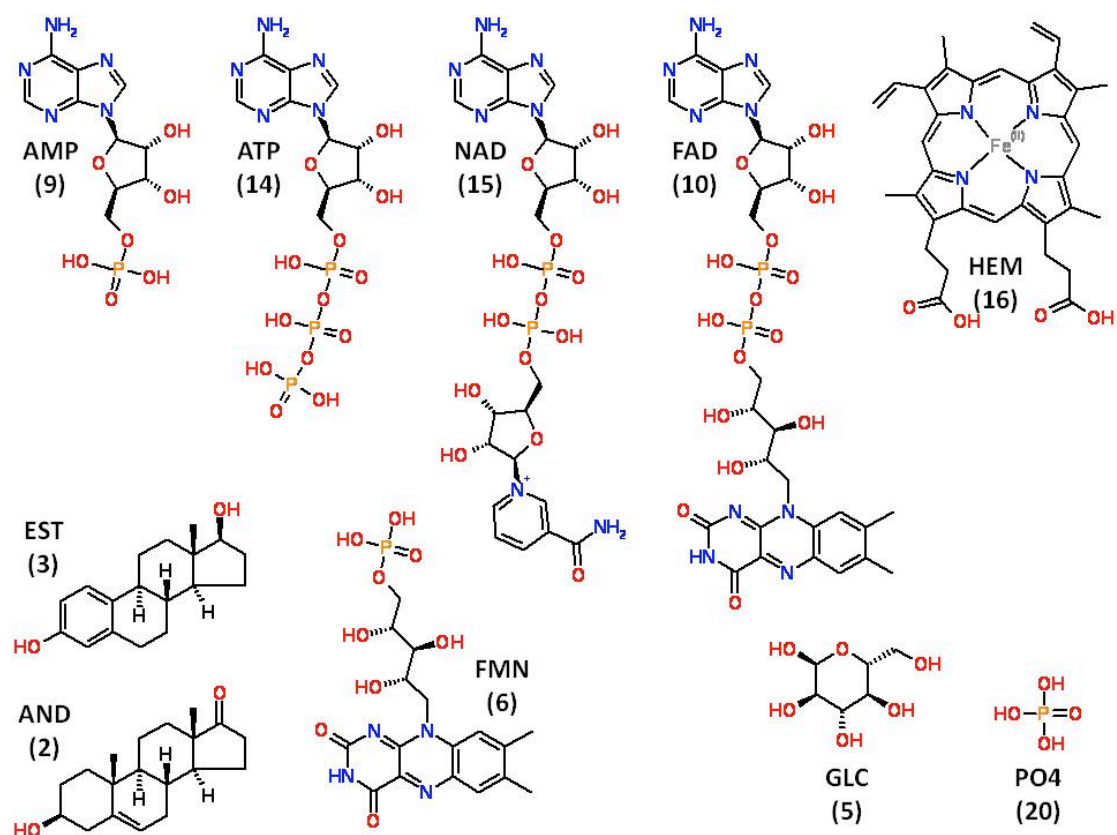


Figure 1.6. Chemical structures of the set of molecules studied by Kahraman *et al.* (Kahraman *et al.* 2007; Kahraman *et al.* 2010). The labels show the three-letter code of the molecule in the HET entry and the number in parenthesis denotes the number of instances used in the Kahraman study.

For these natural molecules, the authors found that the binding pockets were on average three times larger than the volume of the ligand bound. This led to the definition of a “buffer zone” as the free space between the protein and the ligand partially occupied by water (Kahraman *et al.* 2007), and arguably by the partners of these cofactors as discussed below. The authors concluded that the assumption of similar geometrical characteristics for diverse pockets binding the same ligand is only partially true. Looking at the structures and their frequencies used in this study (Figure 1.6), it is worth highlighting that these molecules would need extra room in their binding pockets to bind cofactors or to transfer groups to carry out their function. It

seems unsurprising then, that authors could not find a single perfect fit in the dataset after careful inspection of the crystal structures. In addition, residual flexibility upon binding is entropically favourable (Böhm *et al.* 1996) and arguably it is fundamental for function.

In a more recent analysis of the same data set, Kahraman and co-workers (Kahraman *et al.* 2010) studied the variation of physicochemical environments these ligands experienced in the non-homologous binding sites. For each ligand-protein complex, electrostatic potentials (ESP), hydrophobicity scores, hydrogen bonds and van der Waals potential energies were calculated and used graphically to visualise the physicochemical fields that the ligand experiences in each binding site. The analysis showed that there was no correlation between the average physicochemical properties of the binding site and the ligand bound to them, in other words, the same ligand can be recognised by one protein by electrostatic interactions and by entropic effects by another. In terms of the electrostatic potential experienced by the same ligand in different binding sites, the authors found significant variation, often assisted by the diversity of the neighbouring chemical compounds (NCC such as metals, cofactors and coenzymes within 9Å of any ligand atom). In comparison, the hydrophobicity to which the ligands were exposed in the different cavities varied much less. The authors warned about the use of methods that predict function from structure based on similarity of known functional sites, as their results shown no complementarity between sites binding the same ligand.

This divergence can be explained in terms of binding-site modularity (Gherardini *et al.* 2010) derived by the intrinsic modularity of the nucleotides (base, sugar, phosphates and cofactors). Protein sites binding nucleotides are normally composed by 3D motifs repeated in different protein folds, which recognise the same nucleotide moiety. For example, the acceptor-donor-acceptor motif interacting with the nucleotide base, or the glycine rich loop that often recognises the phosphate groups. Gherardini and co-workers

discuss the evolutionary implications of this modularity. If functional sites are not single functional entities but can be decomposed into small modules instead, the convergence of the whole binding site is a rare event.

1.2.3.4 Synthetic molecule binding sites

After the seminal paper by Hopkins and Groom, "The druggable genome" (Hopkins *et al.* 2002), many studies have tried to capture the characteristics of "druggable pockets" by analysing the known structures of protein-drug or drug-like molecules. Often, these analyses are driven by the development of pocket detection algorithms and scoring schemes in order to differentiate "druggable" cavities (i.e. will bind preferentially a small molecule drug) from those that are "non-druggable". Few analyse the interactions between the drug-like molecules and the proteins, which is the focus of this section.

In 2005 the Abagyan group developed the program PocketFinder (An *et al.* 2005), an algorithm based on estimating the potential for van der Waals interactions with the protein in order to identify binding envelopes. The authors validated the method, predicting 96.8% of known protein-ligand binding sites. These binding sites were extracted from the PDB, removing entries with common cofactors and substrates (like heme, ATP and other high frequent ligands) or with ligands with less than seven atoms. Further filters were applied to remove entries with proteins outside the length range of 50-2000 residues, structures with resolution poorer than 2.5Å were also removed. The final set contained 5,616 protein-ligand binding sites. The results from this analysis confirmed previous results from smaller data sets. The number of envelopes is roughly proportional to the overall volume of the protein, approximately one pocket per 10,000Å³. For the majority (81%) of cases, the ligand-binding pocket coincided with the largest of the predicted envelopes and for 12% the second largest. Regarding the volume of the pockets with respect to the volume of the bound ligands, this study found that on average

the pockets were 1.4 times larger than the ligands they encapsulate. In addition, the surface area buried by the predicted envelope had an average surface ratio with the whole protein surface of 4.7%. In further investigation focussed on binding pockets for human proteins (943 sites from 160 human proteins), An and co-workers (An *et al.* 2005) clustered the binding sites based on shape, hydrophobicity and electrostatic descriptors and compared the resulting tree with the clustering based on the chemical similarity of the ligands. In most cases, similar ligands bound to similar pockets and proteins, however there were also instances of the same ligand binding to different pockets, as well as one pocket binding to chemically diverse ligands. These differences, the authors concluded, highlight the complex relationship between chemical properties of the ligands and the protein sites where they bind. In other words, there is no an absolute prevalent matching of site and ligand properties.

In 2009, Chen and Kurgan (Chen *et al.* 2009) published an analysis of the atomic interactions between proteins and small molecules. Their data set was extracted from the PDB removing entries with proteins bound to peptides or nucleotides. The level of redundancy for the proteins studied was low, as only proteins with less than 25% sequence identity were accepted. In contrast, the level of redundancy of the small molecules is very high, as all ligands in all pockets were kept, yielding 7,759 ligand-protein pockets from 2,320 protein chains. Indeed, 59% of the protein-ligand complexes analysed involved ligands that were bound in more than 100 pockets. The authors classified these ligands into four categories based on the ligands with high occurrence, namely organic compounds, metals, inorganic anions and inorganic cluster. The analysis focussed in the organic compounds subset composed by 3,685 pockets of 560 distinct small molecules. However, the high level of redundancy of small molecules biased the results presented by this analysis. Although the explicit content of the organic subset was not disclosed, the examples of the organic compounds present more than 100 times were disappointing: acetate, glycerol and 1,2-ethandiol. In addition,

calculation of hydrogen bonds was performed with programs developed to work only with proteins and therefore hydrogen bond estimation might be approximate at best. For example, the authors found that the most commonly observed hydrogen bond involved the backbone NH and an oxygen atom in the ligand. Whether or not this is a genuine result, the oxygen content in the ligands analysed and the redundancy of high oxygen content ligands was not taken into consideration.

In a recent report, Schmidtke and Barril (Schmidtke *et al.* 2010) argued that druggability predictions can also underline the keys of molecular recognition between drugs and their targets. Previous models scored druggable pockets by only taking into account shape and hydrophobicity, but the authors reminded us that polar interactions give selectivity by assisting 3D specific orientations. Moreover, druggable cavities have on average fewer polar atoms than non-druggable ones (20-40% versus 40-60%); therefore electrostatic interactions are stronger due to the hydrophobic environment. In addition, analysis of the change in the ratio of polar and apolar surface areas with the radii of the probe to calculate the surface show that in druggable pockets polar atoms stick out from the cavity surface as anchor points for molecular recognition.

This is not an exclusive observation for drug-binding sites. In protein folding the importance of polar interactions compared to the hydrophobic effect is being reassessed. For instance, in the energetics of protein folding (Baldwin 2007), the hydrophobic effect has been long considered the driving force and most relevant factor. However, the importance of the peptide hydrogen bond is increasingly gaining relevance, and compelling evidence is accumulating to justify its place as one of the two major factors for protein folding.

In thermodynamic terms, it is worth highlighting here that polar interactions contribute enthalpically to binding but not in a linear fashion.

Polar groups lose the enthalpy of their hydrogen bonds with water to gain only a little more enthalpy by forming successful contacts with the protein, as well as eventually forcing hydrophobic groups to be exposed to the solvent (Freire 2008). Besides, correlation of structural interactions with entropic and enthalpic changes upon binding is still a major challenge (Klebe 2006; Ladbury 2010). Analysis of ITC data from protein complexes with biological and synthetic ligands (Olsson *et al.* 2008) found no correlation between burial of polar and apolar surface with enthalpy and entropy respectively, however it is accepted that that successful polar interactions will increase the enthalpic component of the free Gibbs energy and apolar interactions will reflect in the entropic part.

1.2.3.5 Comparisons between different types of small molecules binding sites

In 2007, Ji and colleagues (Ji *et al.* 2007) analysed the content of the sc-PDB (annotated database of druggable binding sites from the PDB, (Kellenberger *et al.* 2006)). This set was composed of 2,186 small molecules (MW 70-800Da) bound to 5,740 different SCOP domains belonging to 591 different folds. Water molecules, metals, solvents, detergents and covalently bound ligands were removed, as well as ligands that had more than 50% solvent exposed surface. Ji and co-workers (Ji *et al.* 2007) found that the number of ligands versus the number of domains they bound to follows a power law. Almost one third of the ligands interact with two or more domains, and few ligands are bound to more than 100 distinct domains. These most promiscuous ligands are the hubs for metabolic networks, like for example ATP, the most common ligand in this set bound to 35 different folds.

Furthermore, the authors compared these two thousand small molecules bound to proteins (ligands) with a similar number of random screening molecules extracted from the ACD-SC (available chemical directory for screening compounds, from MDL). The comparison was based on factor

analysis using principal component analysis (PCA) of 70 molecular descriptors. The loadings of the two components explaining most of the variance revealed that polar surface area (PSA), hydrogen bond donor count (HBD), hydrogen bond acceptor count (HBA) and partition coefficient (logP) could discriminate between random screening compounds and ligands. Further analysis of the distributions of these properties for the two sets of molecules highlighted that on average ligands had higher PSA, HBD and HBA and lower logP than screening molecules. However, this study was centred on the chronology of evolution of protein-ligand binding and no further insight or distinction into the types of ligands considered was given.

Adrian Schreyer found results more relevant to drug discovery in his analysis of the CREDO database (Schreyer 2010). Schreyer compared the atomic interactions of proteins with drug-like molecules (several filters yielding a group of molecules with molecular weight range of 100-600Da) and with endogenous molecules (identified with the KEGG database (Kanehisa *et al.* 2008)). Drug-like molecules engaged on average more hydrophobic and aromatic interactions and less hydrogen bonds than the endogenous molecules. Analysis of the polar and apolar accessible surface area (PASA and AASA) versus molecular weight (MW) of these molecules revealed that drug-like molecules only increase AASA with MW whereas PASA remains constant. Conversely, endogenous molecules increased PASA with MW while AASA remained constant in comparison with drug-like molecules. This result is in agreement with findings by Olsson and co-workers (Olsson *et al.* 2008), which revealed that on average synthetic ligands have greater entropic contributions than native ones, in consonance with the higher lipophilic character of drug-like molecules.

1.3 Aims of this thesis

Drug discovery is at an inflexion point: old practises are being scrutinized to try to optimise outcomes and new areas, such as tackling protein-protein complexes with small molecule therapeutics, are being explored.

In chapter 2 of this thesis, a resource to aid the latter is described and analysed. Published reports of small molecule inhibitors are collected into a relational database called TIMBAL. Analysis of these successful small molecules in comparison with other compounds relevant to medicinal chemistry is discussed.

The Blundell group has established structural databases with atomic interaction and annotated data for all protein complexes in the PDB. These are: CREDO, holding protein-small molecules complexes (Schreyer *et al.* 2009); BIPA, protein-nucleic acids complexes (Lee *et al.* 2009) and PICCOLO, protein-protein complexes (Bickerton 2009). These databases are powerful tools that enable the study of molecular recognition at atomic level from the different types of molecules. The interest in small molecules disrupting protein-protein interfaces leads to the question of how these small molecules mimic the interactions of the protein partner. Thus, much insight can be gained extracting interactions profiles for protein-small molecules (CREDO) and protein-protein complexes (PICCOLO).

Chapter 3 of this thesis examines CREDO and PICCOLO with the objective of assessing whether comparisons across both databases are feasible. For example, calculation of atomic interactions, specifically hydrogen bonds, is not a trivial task, and the possibilities of making accurate calculations varies with the knowledge of tautomeric forms, atomic hybridisation and formal charges that have to be assigned in order to ensure compliance with the geometrical constraints that this directional interaction

requires. In chapter 3, differences between databases arising from such challenges are highlighted and the development of simpler atomic contacts that allow straightforward comparisons between them is described.

Chapter 4 defines non-redundant subsets of molecules - drugs, drug-like, small peptides, natural molecules and proteins - interacting with proteins and explores the atomic interaction patterns presented by these subsets in the context of medicinal chemistry. The objective of these analyses is to learn from natural patterns and migrate this knowledge into the design of new therapeutics.

Finally, chapter 5 studies the structural features of the binding sites and interaction surfaces of the same subsets of molecules.

2.1 Introduction

2.1.1 Protein-Protein interactions (PPI) as drug targets

The central role played by protein-protein interactions in living organisms make them attractive targets for therapeutic intervention. Successes in antibody therapies targeting extracellular protein complexes (Adair *et al.* 2005) encourage drug researchers to seek small molecules that modulate these pivotal interactions. Small molecules bring a number of advantages over antibody therapies, not least in cost of goods and ease of delivery. However, the quest for an ideal small molecule, which can compete with one of the partners in a multi-protein complex, will be challenging. Just a decade ago, this quest was thought to be insurmountable; nevertheless two experimental findings have made protein-protein interactions more attractive for drug discovery: the existence of hot spots and the adaptability of the interfaces targeted (Whitty *et al.* 2006). In fact, in recent years there have been an increasing number of studies reporting small molecules disrupting protein-protein interactions. Figure 2.1 shows the increase of citations regarding protein-protein interactions, including inhibition by small molecules.

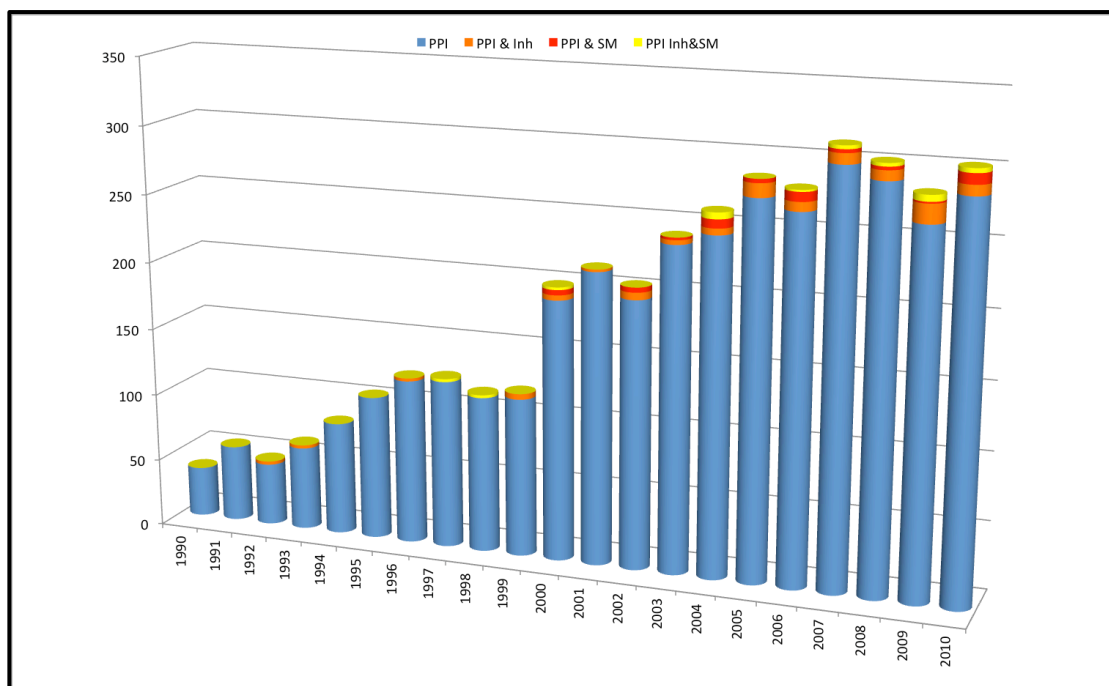


Figure 2.1. Number of publications per year (normalised by the total number of publications per year) containing in the title "protein-protein interaction". The colour code is as follows: blue, only PPI in the title; red, ppi and small molecule (SM) in the title; orange, ppi and inhibitor (inh) in the title; yellow, all the above in the title. Searches have been done in PubMed.

2.1.2 Survey of literature reviews of small molecules inhibitors of PPI

One of the first reviews of protein-protein interactions, which included non-peptidic small molecules, was published in 2000. Cochran described (Cochran 2000) several approaches to disrupt protein complexes, one of which was small molecules that modulate (agonize or antagonize) cytokine signalling. This review collates four molecules and highlights the rigidity of the scaffolds and their richness in aromatic rings and indoles.

In 2002, Toogood (Toogood 2002) wrote the first dedicated perspective/review of the use of small molecules to inhibit protein-protein interfaces with therapeutic purpose. This review describes in detail the methods and frameworks of the early projects delivering small molecules inhibiting protein complexes. However, not all these have validated binding to

one of the protein partners. The author did not try to derive common features of these molecules, but rather advocated doing this when the field had become more consolidated.

It was not until 2004 that the first analysis of small molecules inhibiting PPI was published. Pagliaro and co-workers (Pagliaro *et al.* 2004) collected 19 molecules from 12 different multi-protein complexes, half of which were not covered by the standard screening libraries and only eight of which fulfilled the Lipinski drug-like criteria (Lipinski *et al.* 1997).

Since then, several studies have focused on subsets of small molecules that disrupt protein-protein interactions. In 2005, Fischer reviewed protein-protein interactions in drug discovery (Fischer 2005) and collected 38 small molecules from 11 complexes. Again, only a small proportion of these molecules passed the Lipinski (Lipinski *et al.* 1997) and Veber (Veber *et al.* 2002) drug-like criteria. In the same year, Fry and Vassilev (Fry *et al.* 2005) reviewed targeting protein-protein interactions for cancer therapy. From the few cases where small molecules were found to inhibit protein complexes, most of these small molecules had properties that were not drug-like, such as too many rotatable bonds, insolubility issues, charged moieties or reactive groups. In 2007, Neugebauer and co-workers (Neugebauer *et al.* 2007) extracted known inhibitors from the literature excluding peptides and small proteins. These authors collected 25 structurally diverse small molecules (all of them with a molecular weight higher than 400 Da) from seven targets and compared them with 1057 FDA (The Food and Drug Administration) approved drugs. More than 600 molecular descriptors were calculated and decision trees discriminated between PPI inhibitors and FDA drugs. The most relevant descriptor to distinguish PPI inhibitors was based in molecular shape and size, however no clear guidelines of the value range was given. Later in the same year, Wells and McClendon (Wells *et al.* 2007) analysed six multi-protein complexes where structural and binding data showed small molecules competing directly with one of the partners. In most of the cases, site

adaptability occurs when the small molecule binds to the interface. The authors compared these small molecules with other sets of drug-like small molecules and found little similarity. For 13 diverse molecules, each optimised against one of these six complexes, they demonstrated a linear correlation between number of atoms of the small molecule and its energy upon binding with the protein. This ratio, known as the ligand efficiency, LE (Hopkins *et al.* 2004), had a value of about ~ 0.24 for these molecules. Assuming that this is a general threshold for these interfaces, the authors concluded that a 10nM binder would require a molecular weight of 645Da, which is well above the classical 500Da Lipinski limit (Lipinski *et al.* 1997). Similar findings were reported by Fry (Fry 2008); small molecules binding to protein interfaces tend to be large rigid structures with complex 3D shape.

At that time we were developing and analysing the resource described in this chapter, results of which were published in 2009 (Higueruelo *et al.* 2009) and which I report here in greater detail. Since then, other studies have also been published.

In 2010, Sperandio *et al.* (Sperandio *et al.* 2010) studied the chemical space occupied by small molecule inhibitors of protein-protein interfaces. These authors suggested that as the interfaces are richer in tyrosine, phenylalanine, tryptophan and methionine the primary chemistry to tackle these interfaces would be aromatic and hydrophobic. Analysing 66 PPI inhibitors versus 557 small molecule drugs from the DrugBank (Knox *et al.* 2011), both sets passing filters of structural diversity and loose drug-like properties, Sperandio and co-workers found that PPI inhibitors were bigger and more lipophilic than the drug set (mean of molecular weight 421 vs 341 and mean of $\log P$ 3.58 vs 2.61, P values $5E10^{-9}$ and $6E10^{-6}$ respectively). In the same study, more than 1600 molecular descriptors were used to derive decision trees, which found that shape descriptors and accounting for unsaturated bonds had the most discriminative power for distinguishing PPI inhibitors from standard drugs. Being complex shapes and high number of

unsaturated (including aromatic) bonds favoured by PPI inhibitors. This result, matches the previous findings of Neugebauer (Neugebauer *et al.* 2007) and Fry (Fry 2008), as well as ours as we will see in further sections. The same research group published a second report (Reynes *et al.* 2010) about the applicability of these findings to the design of focused libraries to target protein-protein interactions.

In the same year, Bourgeas et al. (Bourgeas *et al.* 2010) released the 2P2I database, a hand curated database of the structures of protein-protein complexes with known inhibitors. Only targets with structural information for both the protein-protein complex and the protein-inhibitor complex were included in the database, which described 17 protein-protein complexes and 56 protein-small molecule inhibitors. These authors focused the published study on the characteristics of the protein interfaces and this is discussed in chapter 5 of this thesis. Interestingly a more recent update of the 2P2I database has less complexes and small molecules as the authors have removed some of the previous entries (Morelli *et al.* 2011). The 2P2I database has now 12 protein-protein complexes with a non-redundant set of 39 small molecules bound to their protein-protein targets. In this latest study, the authors analysed the molecular properties of the small molecule inhibitors, along side their binding and surface efficiency indexes (BEI and SEI) as defined by Abad-Zapatero et al. (Abad-Zapatero *et al.* 2005; Abad-Zapatero *et al.* 2010). The Rule of 4 is proposed as general profile for possible protein-protein small molecules inhibitors, based in the average of MW 547 ± 154 thus $MW > 400$, $\log P$ 3.99 ± 2.37 thus $\log P > 4$, number of rings 4.44 ± 1.02 thus $NoR > 4$ and number of hydrogen bond acceptor 6.62 ± 2.60 thus $HBA > 4$. When the small molecules inhibiting protein-protein complexes were mapped to the BEI and SEI space of the marketed oral drugs, they appeared in the zone of "sub-optimal series that could not get optimised", as they are too large and lipophilic to fit in the classical oral drug space. However, one of these large molecules, Navitoclax (ABT-263) is progressing in phase II clinical trials for cancer. The authors concluded this report advocating shifting the

paradigm of what it takes to be a drug for this class of targets as well as developing alternative and parallel technologies, like the nanoparticle drug delivery system (NPDDS) (Morelli *et al.* 2011).

2.1.3 Need for collecting existing data and derive knowledge from it

It is clear then, that drug discovery for these challenging targets is in an uncharted area, where we may need to reassess the concept of drug-likeness for this type of target. Classical drug-like properties are largely derived from competitive inhibitors of endogenous small molecules. Current screening decks may not be well suited to identify protein-protein small molecule modulators. We must also increase our understanding of the mutual recognition between small molecules and interfaces in order to develop better methods for growing initial hits and to efficiently maximise their affinity and selectivity, whilst trying to confer on them the appropriate profile of a therapeutic agent.

In order to find hits in this context, tools for the accurate prediction of hot spots and protein flexibility are needed as well as knowledge of which types of molecular interaction are best exploited by small molecules on protein surfaces. I addressed this last point with the creation of a relational database that holds the current small molecules disrupting protein-protein interactions. This database, called TIMBAL, is compatible with other structural databases of protein-protein (Bickerton *et al.* 2011) and protein-ligand interactions (Schreyer *et al.* 2009), providing a useful framework to study small molecule interactions at protein-protein interfaces. This compatibility allows the exploration and comparison of the structural features and interactions of small molecule modulators of protein-protein interactions and the multi-protein complexes they inhibit. This resource also allows profiling and analysing the molecular characteristics of the small molecules that successfully inhibit protein-protein interactions.

2.2 Methods

2.2.1 Creation of a database of small molecule inhibitors of PPI: TIMBAL

TIMBAL is a relational database containing small molecules that inhibit protein-protein interactions. These molecules and the information regarding the systems affected by them have been retrieved from relevant scientific publications. The literature up to 2008 was searched and analysed in order to identify all the known small molecules modulators of protein complexes. Short peptides or peptidomimetic molecules were not included at this stage. Manual updates until 2011 have been carried out only for molecules deposited in the PDB (Berman *et al.* 2000). The growth of data (see Figure 2.1) in the past years makes hand-curated databases a phenomenally time-consuming task. The feasibility of maintaining high quality data alongside other research duties is low. The maintenance of TIMBAL has been envisaged through automated searches on the ChEMBL database (Gaulton *et al.* 2011).

Literature searches were carried out using Ovid (UCB access, <http://ovidsp.tx.ovid.com/>) and Pubmed (public access, <http://www.ncbi.nlm.nih.gov/sites/entrez>). Automated queries were set up in Ovid to keep the data source up to date until 2008.

Data extraction was achieved by critically reading the papers and manually sketching molecules into an Excel spreadsheet with Accord functionality, (<http://accelrys.com/products/informatics/desktop-software.html>) to handle chemical structures.

Not only is it of great importance to collate all known small molecule modulators of protein-protein interactions, but also to relate them to the structure-based database projects within the department, in order to

maximise the information that can be derived for this type of molecule. For this reason it became apparent that TIMBAL should be compatible with other databases within the Biocomputing Group in the Department of Biochemistry (CREDO, (Schreyer *et al.* 2009); PICCOLO, (Bickerton *et al.* 2011) and BIPA, (Lee *et al.* 2009)). To achieve this compatibility the spreadsheet mentioned above is post-processed with a Python script, which generates TIMBAL as a relational database in MySQL (open source database engine, <http://www.mysql.org>). The database is normalised to remove redundancy and is constituted by different tables. The TIMBAL schema is shown in Figure 2.2.

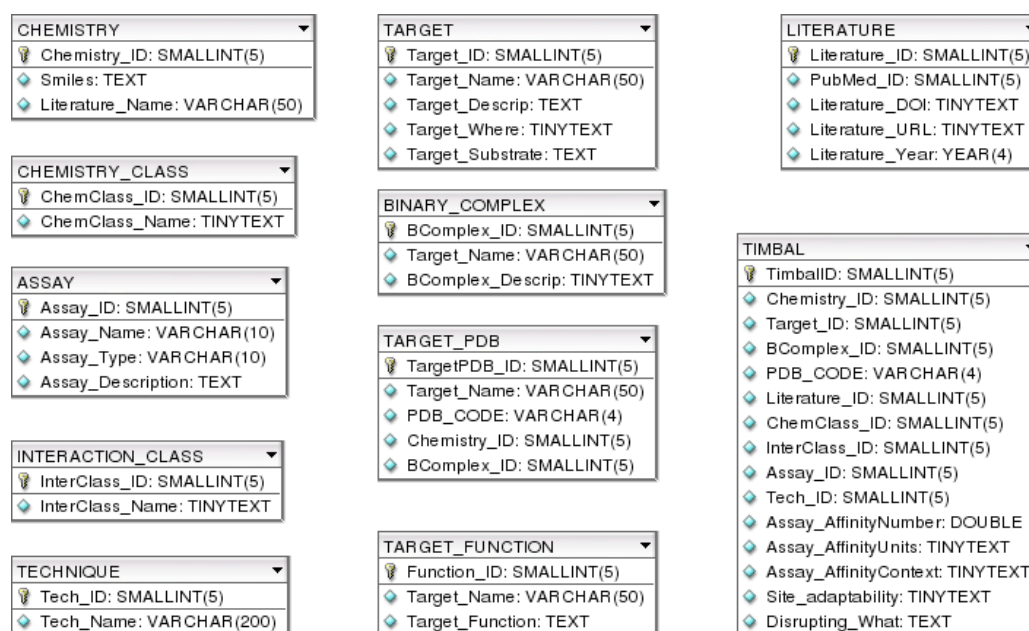


Figure 2.2. Complete schema of TIMBAL database. Chemical structures are held as SMILES (Simplified Molecular Input Line Entry System), generated with the Accord functionality within Excel. These sets of tables have been defined to normalise TIMBAL and avoid redundancy.

The subset of TIMBAL molecules present in the PDB are also a subset of the CREDO database, consequently the CREDO database has a TIMBAL table which allows profiling of the protein-protein modulators and comparison with other ligands in the PDB.

2.2.2 Profile and analysis of TIMBAL

2.2.2.1 TIMBAL profile

The TIMBAL database has been profiled in terms of the molecular properties of the small molecules in it. Typical molecular properties like molecular weight, *alogP*, polar surface area (PSA) and rotatable bonds have been calculated with Scitegic Pipeline Pilot software (<http://accelrys.com/products/pipeline-pilot/>). In order to put TIMBAL molecules in a medicinal chemistry and structural context four other sets of molecules have also been profiled. The aim is to analyse possible trends and differences between sets. These are:

- **Drugs:** Preclinical, phase I to IV, and launched drugs from the MDDR database (MDL® drug data report) with molecular weights below 900Da; this cut-off has been set up to have only small molecules. The biggest molecule in TIMBAL now has a molecular weight of 813Da. An amino acid SMARTS (Smiles ARbitrary Target Specification, query extension of SMILES) filter (Daylight) is applied to these molecules in order to remove peptide-like molecules from the set. This set contains 11,843 molecules.
- **Screening compounds:** A random selection of small molecule screening compounds from the catalogues of three different suppliers: Enamine, Asinex and Maybridge. The same cut-offs have been applied, and molecular weights below 900Da and peptide-like molecules filtered out. This set contains 12,022 molecules.
- **Ligands from the PDB:** Molecules in CREDO that are not in TIMBAL. The same cut-offs have been applied: molecular weight below 900Da and removal of peptide-like molecules. Molecules with 10 atoms or less have been also filtered out to remove most of the small molecules solvent and salts. This set contains 7,841 molecules.
- **DL-ligands from the PDB:** The drug-like subset from the above set. Ligands in the PDB (as extracted from the HETATM entries) are very diverse. These small molecules are not appropriate for a drug-

like comparison. For example, they can be heavy metal complexes, detergents and nucleotide analogues. However, the distinction between drug-like molecule and non drug-like is not trivial. The pragmatic approach used to select this drug-like subset is as follows: molecules from the above CREDO set with at least one carbon atom; with composition of carbon, nitrogen, oxygen, sulphur and halogen only; with at least one ring; and with no chains longer than six carbons sp^3 -CH₂. The resulting molecules have been clustered based on chemical structure using MDL public keys as descriptors and maximum dissimilarity method to find the centre of the clusters as implemented in Scitegic Pipeline Pilot software. Clusters with nucleotide analogues or detergents have been removed. This subset contains 3,048 molecules.

2.2.2.2 Pharmacophoric analysis of the interface

TIMBAL has also been profiled and compared with the three sets of small molecules described in 2.2.2.1 in terms of the chemical functionality of the compounds that are present. Thirty-nine medicinal chemistry functional groups have been used in a single substructure search against all the molecules in the different sets. Scitegic Pipeline Pilot software has been used to perform these searches. This analysis should highlight any particular functional group favoured within the small molecules modulators of protein-protein interactions.

The TIMBAL subset that it is contained in CREDO (i.e. structural data is available for the complex small molecule protein) has been analysed and compared with the rest of CREDO molecules in terms of types of interactions and contacts between the ligand and the protein. This has been done using all the pre-calculated contacts derived from the structural data developed by Adrian Schreyer in CREDO. These contacts are defined by distance and atom type between atoms in the ligand and the protein. I have used: covalent, van der Waals clash, van der Waals, hydrogen bond, ionic, piCation, aromatic and

hydrophobic. These contacts are also pre-calculated in the PICCOLO database containing protein-protein interactions in the PDB. Therefore a cross comparison between PICCOLO, CREDO and TIMBAL interactions has also been analysed, in order to plot trends regarding the type of interactions and contacts favoured by TIMBAL molecules.

CREDO also pre-calculates the protein surface buried by ligand binding. This measurement has been compared between the TIMBAL subset and the rest of CREDO molecules.

Affinity data for TIMBAL molecules have been also collected from literature when available. Ligand Efficiency (LE (Hopkins *et al.* 2004), free energy of ligand binding $\Delta G = -RT \ln K_d$ divided by number of non-hydrogen atoms) has been calculated on the assumption that K_i and IC50 are good approximations of K_d and the temperature is set to 300K. This LE has been compared with the threshold described in literature by Wells and McClendon (Wells *et al.* 2007) for the most optimised small molecule inhibitors of protein-protein interactions (~0.24).

2.3 Results and discussion

2.3.1 TIMBAL database

TIMBAL can be publicly accessed through the Department web site (<http://www-cryst.bioc.cam.ac.uk/timbal>). It now contains 117 small molecules of which 39 are in the PDB co-crystallised with their PPI targets and therefore are also included in CREDO database. The analysis described here however was performed in 2008 when the database was created; at that time TIMBAL had 104 small molecules, 27 of which were in the PDB. TIMBAL also holds 247 small fragments; these are tether (Erlanson *et al.* 2004) hits from Cys mutations in the IL-2 cytokine (Arkin *et al.* 2003). This set of fragment molecules has not been included in the analysis, as they would bias towards the IL-2 interface. Overall, the TIMBAL database contains small molecules disrupting 17 protein-protein complexes, a summary of the contents of TIMBAL in terms of protein-protein systems, techniques used to identify the small molecules and number of compounds per system at the time of the analysis is shown in Table 2.1.

	Complex	Complex Type	Therapeutic Area	Techniques	N of SM (series)
Extra-cellular	IL-2/IL-2R α	Heterodimer	Immuno-suppressor	Peptidomimetics Tethering ²⁵	6 (2) + 247 tethers
	CD80/CD28	Heterodimer	Immuno-suppressor	Cell based screening HTS ELISA	4 (2)
	TNF α trimer	Homotrimer	Inflammation	CTGFA with ELISA	2 (1)
	ZipA/FtsZ	Heterodimer (small peptide)	Antibacterial	Screening SPR and FPA SBDD	21 (7)
Intra-cellular	Bcl2 or BclXL /Bax or Bak or Bid	Heterodimer (small peptide)	Oncology	SBDD. VS-Docking. FPA HTS FPA SAR by NMR	26 (9)
	β -Catenin /Tcf4 or Tcf3	Heterodimer (flexible peptide)	Oncology	SBDD. VS-Docking. NMR and ITC ELISA screening	4 (2)
	c-Myc/Max	Heterodimer binding to DNA	Oncology	Screening FPA	1
	ESX/Sur-2	Heterodimer (small peptide)	Oncology	Cell based screening + binding assay	1
	p53/MDM2	Heterodimer (small peptide)	Oncology	SBDD. VS-Docking. FPA LBDD. VS-Pharmacophore. FPA Peptidomimetics. Natural products HTS ThermoFluor® ²⁷ , ELISA, SPR, FPA	16 (7)
	p53/S100B	Heterodimer (small peptide)	Oncology	SBDD. VS-Docking. Trp Fluorescence assay	7 (4)
	XIAP/Casp-9 or SMAC	Heterodimer	Oncology	Peptidomimetics. Natural products SBDD. VS-Docking. FPA	5 (2)
	UL30(Po) /UL42	Subunits HSV	Antiviral	HTS FPA	3 (3)
	E1-E2 /DNA(HPV)	Heterodimer binding to DNA	Antiviral	SAR by NMR	4 (2)
	ToxT dimer	Homodimer	Antimicrobial	HTS phenotypic screen	1
	iNOS dimer	Homodimer	Inflammation Immunology	CombiChem	1
	RGS4/G α o	Heterodimer	Modulation of GPCRs	Screen FCPI assay	1
	CMR1/NES	Heterodimer	Antiviral	Cell based screen	2 (1)

Table 2.1. Protein-protein complexes found in literature modulated by small molecules. Green background for systems with structural information. Red background for systems without structural information. Blue background highlights the systems with more small molecules.

Table 2.1 key: Tethering (Erlanson *et al.* 2004). HTS (High Throughput Screening). ELISA (Enzyme-Linked Immunosorbent Assay). CTGFA (Combinatorial target-guided fragment assembly, Sunesis). SPR (Surface Plasmon Resonance). FPA (Fluorescence Polarization Assay). SBDD (Structure-Based Drug Design). VS (Virtual Screening). SAR (Structure Activity Relationship). NMR (Nuclear Magnetic Resonance). SAR by NMR (Hajduk 2006). ITC (Isothermal titration calorimetry). LBDD (Ligand-Based Drug Design). ThermoFluor® (Cummings *et al.* 2006). CombiChem (Combinatorial Chemistry). FCPI (Flow Cytometry Protein Interaction assay).

The data shown in Table 2.1 highlight the importance of structural information for these challenging targets. Virtually all PPIs that have been successfully disrupted by small molecules have crystallographic or NMR structural data for the protein-protein complex (IL2/IL2Ra, Bcl-XL/Bad, MDM2/p53, S100B/p53, TNF trimer, XIAP/Smac, B-catenin/Tcf4, ZipA/FtsZ, cMyc/Max, E1/E2, iNOS dimer, and UL42/HSV-Pol); or for one of the complex components (CD80, CMR1). Also apparent from the table is the connection (highlighted in blue in Table 2.1) between complexes where one of the partners is a small peptide and the success in finding small molecules binding to the interface. These examples lead to the hypothesis that these types of interfaces are more druggable than the interfaces from globular constituents, as the existence of one partner that becomes ordered on binding allows a larger interaction surface between ligand and protein and often better formed pockets (Blundell *et al.* 2006). In addition, these complexes may be more amenable to the development of scalable competitive binding assays to identify small molecule inhibitors. A good example of this is the Fluorescence Polarisation Assay (FPA), a homogeneous assay that gives robust results if the size ratio between components of the complex is high (Berg 2003). The small peptide is fluorescent labelled and put in solution with its bigger partner. The whole complex is excited with polarized radiation, which emits highly polarized fluorescence as the bigger protein maintains the orientation of the fluorescent peptide. If a competitive inhibitor is added to the system, the peptide is released in solution free to rotate and translate which will cause decrease of the polarization of the emitted fluorescence.

Another point to highlight from the structural data of these complexes is that almost all targets with a large number of reported successful small molecules modulators have preformed small and deep pockets in the interface (Bcl-XL, MDM2, S100B) with ZipA as a remarkable exception. Small molecules have successfully modulated the interaction between ZipA and FtsZ, binding to a shallow hydrophobic interface. However, these molecules did not progress in the path to be therapeutic agents because the required cell penetration, solubility and specificity were not married to their ability to bind to ZipA. This example suggests that assessment of druggability for a PPI target cannot be limited to finding small molecules bound to the interface. However extraordinary, PPI binders need to also have the appropriate molecular profile to achieve the approved drug status.

2.3.2 Profile and analysis of TIMBAL

2.3.2.1 TIMBAL profile

The molecular properties profile of the TIMBAL database is shown in Figure 2.3 and Figure 2.4, as well as the profile of the other four sets described in the Methods section. In order to have all profiles on the same scale, the frequencies for the binned properties have been normalised by the total number of molecules per set. TIMBAL profile (in pink) is more spiky as a consequence of the smaller number of molecules in this set. Table 2.2 summarises the average value and standard deviation for each molecular property and set.

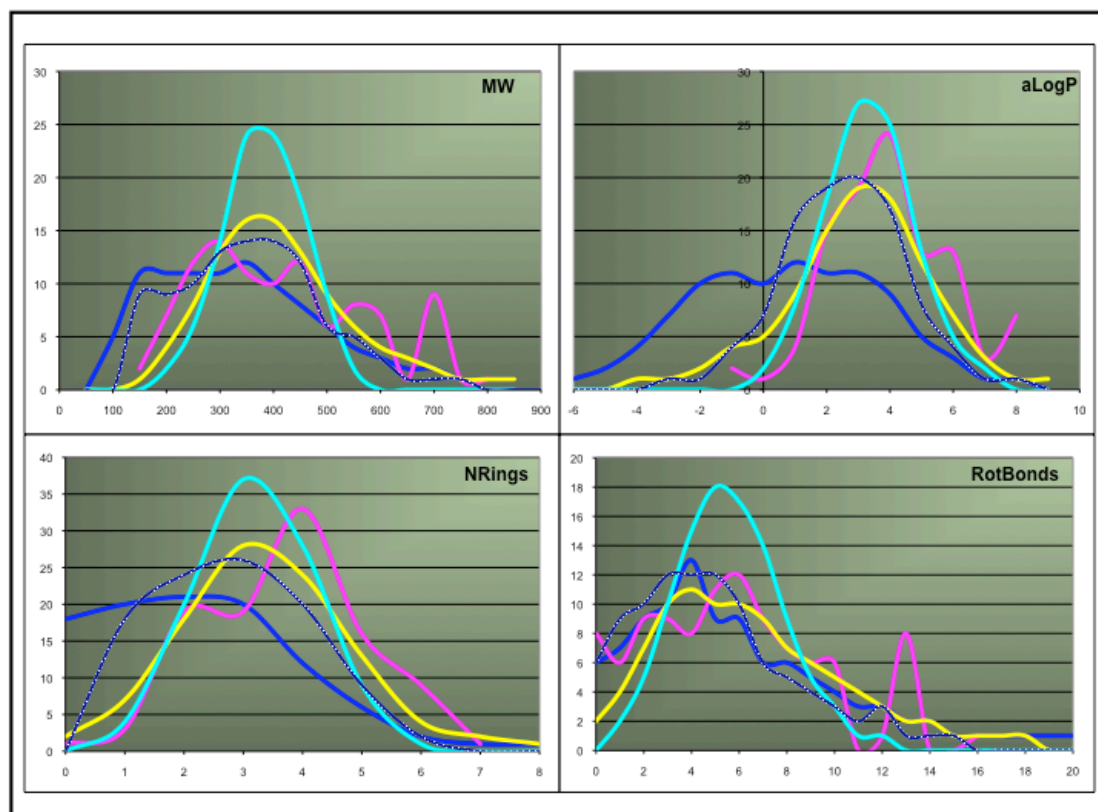


Figure 2.3. Distribution of molecular properties for the different sets of molecules described in the main text. See section 2.2.1. Colour coded: dark blue (PDB ligands), grid dark blue (PDB ligands drug-like subset), yellow (Drugs from MDDR), cyan (Screening compounds), pink (TIMBAL, small molecule inhibitors of protein-protein interactions). MW: Molecular weight; alogP: Calculated logarithm of the partition coefficient; NRings: Number of rings; RotBonds: Rotatable bonds.

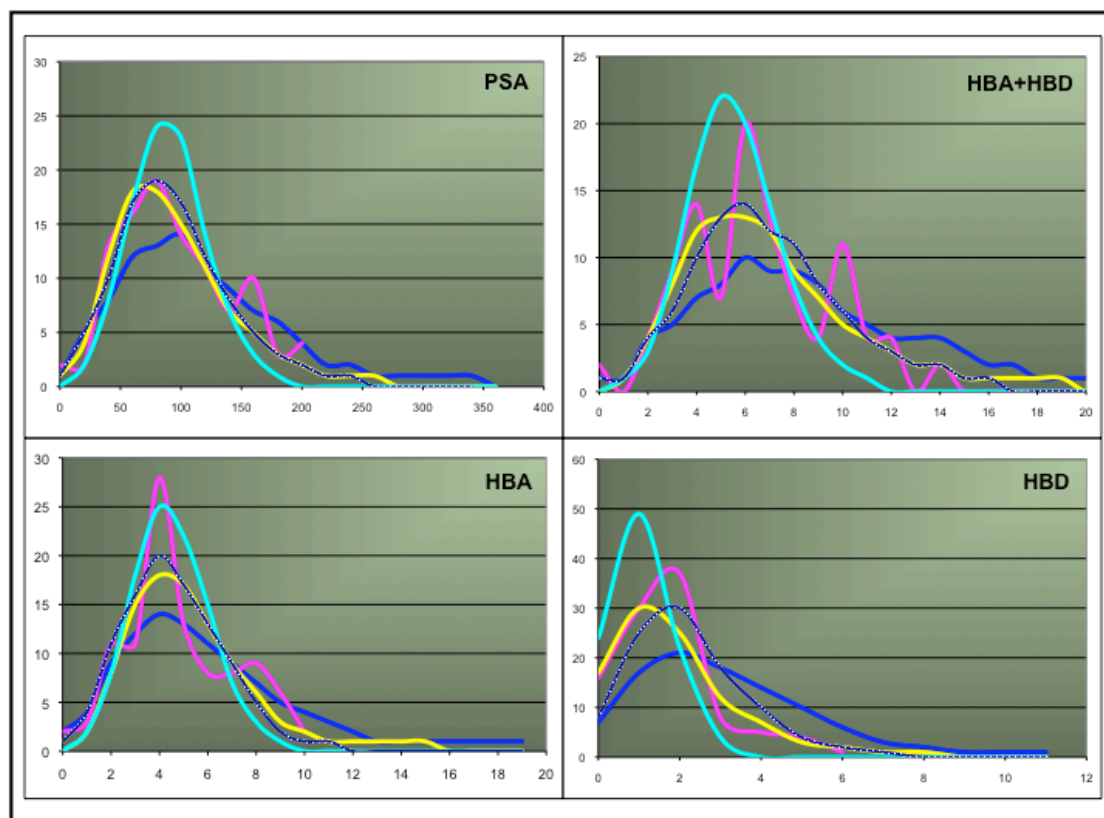


Figure 2.4. Distribution of molecular properties for the different sets of molecules described in the main text. See section 2.2.1. Colour coded: dark blue (PDB ligands), grid dark blue (PDB ligands drug-like subset), yellow (Drugs from MDDR), cyan (Screening compounds), pink (TIMBAL, small molecule inhibitors of protein-protein interactions). PSA: Polar surface area; HBA: Number of hydrogen bond acceptors; HBD: Number of hydrogen bond donors.

	PDB Ligands	PDB Ligands-DL	Drugs	Screening compounds	PPI Inhibitors
MW	352 ± 172 330 340	360 ± 139 350 380	417 ± 140 400 380	384 ± 79 380 370	420 ± 156 400 690
alogP	0.9 ± 3.4 0.9 1.0	2.6 ± 2.3 2.6 2.2	2.9 ± 2.6 3.1 3.8	3.3 ± 1.5 3.3 3.6	4.0 ± 2.0 3.9 3.1
NRings	2.3 ± 1.8 2 2	2.9 ± 1.4 3 3	3.3 ± 1.5 3 3	3.2 ± 1.1 3 3	3.7 ± 1.4 4 4
NAromRings	1.3 ± 1.3 1 0	2.0 ± 1.3 2 2	2.0 ± 1.3 2 2	2.5 ± 1.0 2 2	2.8 ± 1.3 3 3
RotBonds	6.5 ± 5.8 5 4	5.4 ± 4.1 5 4	6.9 ± 4.7 6 4	5.6 ± 2.2 6 5	5.7 ± 3.7 5 6
PSA	122 ± 77 105 65	95 ± 46 90 65	98 ± 57 85 60	90 ± 32 90 85	95 ± 46 85 65
HBA+HBD	9.4 ± 6.2 8 6	6.9 ± 3.3 6 6	7.4 ± 4.1 7 6	5.6 ± 1.9 5 5	6.5 ± 3.0 6 6
HBA	6.2 ± 4.2 5 4	4.7 ± 2.2 4 4	5.4 ± 2.9 5 4	4.5 ± 1.6 4 4	4.8 ± 2.3 4 4
HBD	3.2 ± 2.3 3 2	2.3 ± 1.5 2 2	2.0 ± 1.8 2 1	1.1 ± 0.8 1 1	1.7 ± 1.3 2 2

Table 2.2. Mean ± standard deviation, median and mode of the molecular properties for each set. To calculate the median and the mode, Molecular Weight was binned in 10Da and rounded to integer. Topological polar surface area was binned in 5Å² and rounded to integer. AlogP was rounded to one decimal place.

Although TIMBAL molecules present a spread of molecular properties, for example molecular weight goes from 148Da to 813Da, their overall profile shows a tendency for being big lipophilic molecules. In addition, they have more rings and less rotatable bonds than the molecules from the drugs and DL-ligands from the PDB sets. In spite of the fact of being on the average bigger, TIMBAL molecules show in proportion less features than the molecules from the other sets, as captured by the hydrogen bond donor and acceptor counts. The ratio of hydrogen bond donor and acceptor count by molecular weight for TIMBAL molecules is significantly ($P < 0.05$) smaller than the same ratio for the drugs and DL-ligands from the PDB. In drug discovery, molecules with a similar profile will be regarded as promiscuous and little attractive. It will be interesting to profile these molecules against a

panel of PPI targets to evaluate their selectivity, however there aren't many PPI targets with validated assays, and this type of data has not been found in literature. The same applies to classical later stage properties in the drug discovery time frame like DMPK (Distribution, Metabolism and Pharmacokinetics). These inhibitors are in the very early stage of discovery (with Bcl as a remarkable exception) and the publicly available data are limited.

It is interesting to note that PPI small molecules modulators show a closer profile to the drug set than the screening compounds group. However, only one molecule, an analogue of ABT-737 (TIMBAL molecule of 813Da), has reached the phase I/II of clinical trials (Morelli *et al.* 2011) thus far showing acceptable oral bioavailability despite its huge molecular weight. This might suggest broadening the type of molecules that get screened against a PPI target, in terms of property profile as well as more diverse sources, like natural products for example. On the other hand, products from DOS (Diverse Oriented Synthesis) (Di Micco *et al.* 2009) may be attractive screening candidates as complex shape is common amongst PPI inhibitors (Neugebauer *et al.* 2007; Fry 2008; Sperandio *et al.* 2010).

In order to assess if this profile of TIMBAL molecules is general and not biased by target, Figure 2.5 and Figure 2.6 show the distribution of molecular weight and $\log P$ for TIMBAL molecules colour coded by target. Only targets with more than one molecule have been included in the graph. Seven out of nine targets have molecules with molecular weight greater than 500Da, and all of them have molecules with $\log P$ greater of 4. These properties do not depend on whether the researchers who synthesised the molecules are in an industrial or academic environment (48% molecules generated in industry, 39% in academia and 13% in collaborative efforts). Therefore the general trend for small molecule modulators of protein-protein interactions is being bigger, more rigid, more lipophilic and less hydrogen bonding than molecules in the drug and screening sets.

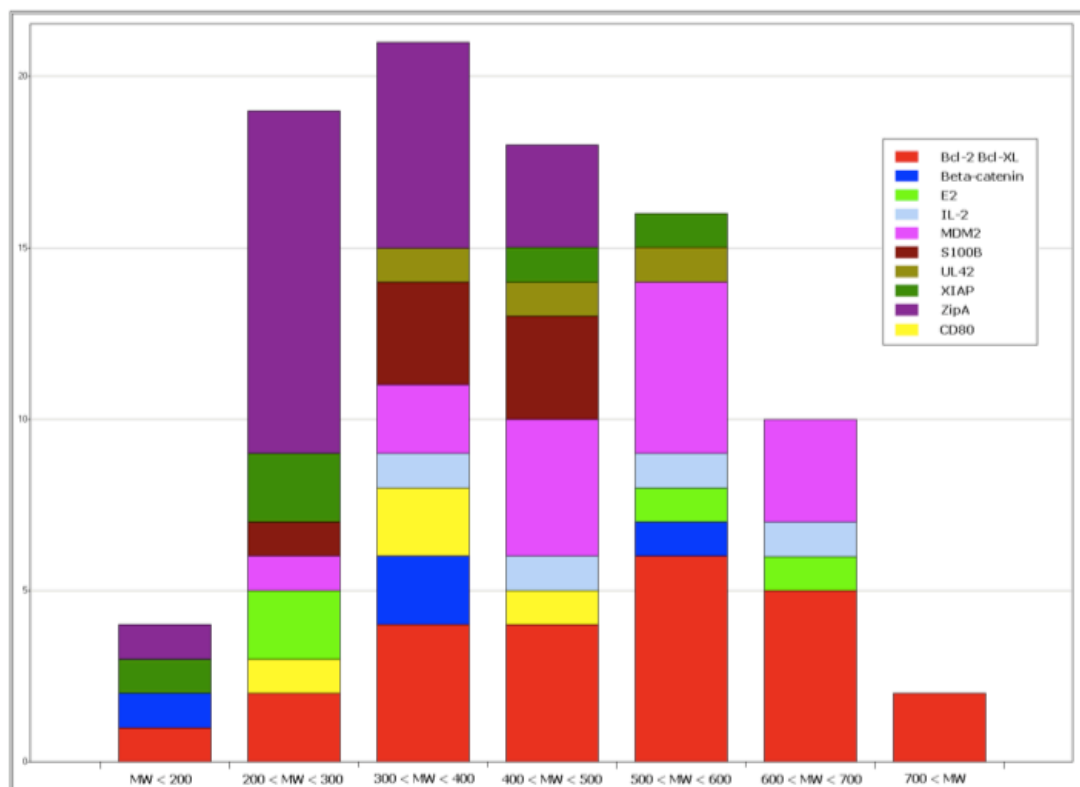


Figure 2.5. Distribution of the Molecular Weight (MW) of the TIMBAL molecules colour coded by target. Only targets with more than one molecule are plotted.

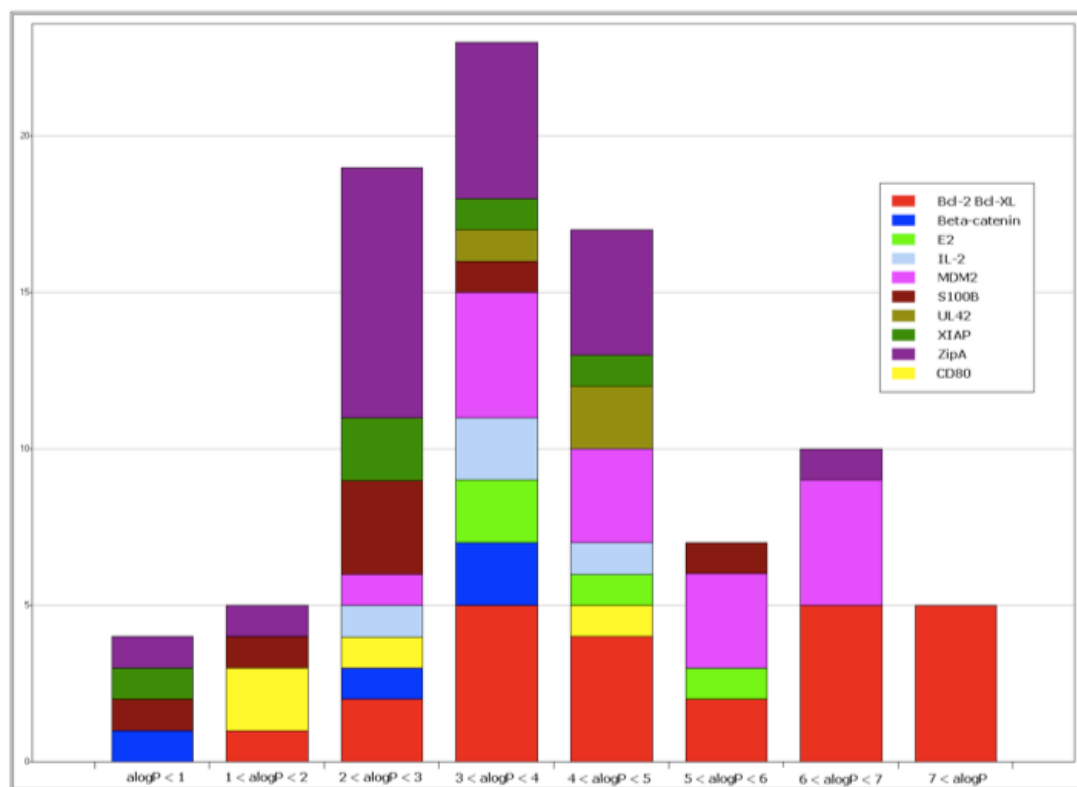


Figure 2.6. Distribution of the calculated logarithm of the partition coefficient ($alogP$) of the TIMBAL molecules colour coded by target. Only targets with more than one molecules are plotted.

In order to evaluate the similarity or dissimilarity, in terms of molecular properties, between the TIMBAL molecules and the other sets of molecules, a principal component analysis has been applied. For this analysis only molecules from the PDB ligand drug-like subset have been used to represent ligands in PDB, as many of the ligands in the PDB are not relevant for drug-like profiling as described in the methods section (2.2.1). The loadings of this PCA are shown in Table 2.3.

Property	PC1	PC2	PC3
StandardDeviation	2.051	1.369	0.953
VarianceExplained	0.526	0.234	0.114
TotalVarianceExplained	0.526	0.760	0.874
MW	0.367	0.438	-0.028
RotBonds	0.319	0.219	-0.666
HBA	0.442	-0.019	0.145
HBD	0.350	-0.271	-0.067
HBA+HBD	0.470	-0.135	0.072
NRings	0.119	0.510	0.664
PSA	0.445	-0.120	0.083
ALogP	-0.114	0.627	-0.280

Table 2.3. Principal Component Analysis (PCA) Loadings for molecular properties used as descriptors of the molecules in the different sets.

Figure 2.7 represents the molecules with their PCA scores, i.e. the projections of the molecular properties onto the first three principal components. As can be seen in the top right quadrant, the drug molecules are spread broadly in this space and it captures most of the TIMBAL molecules. In the bottom left and right quadrants, TIMBAL molecules are less covered by the PDB ligands-DL set and Screening compounds respectively.

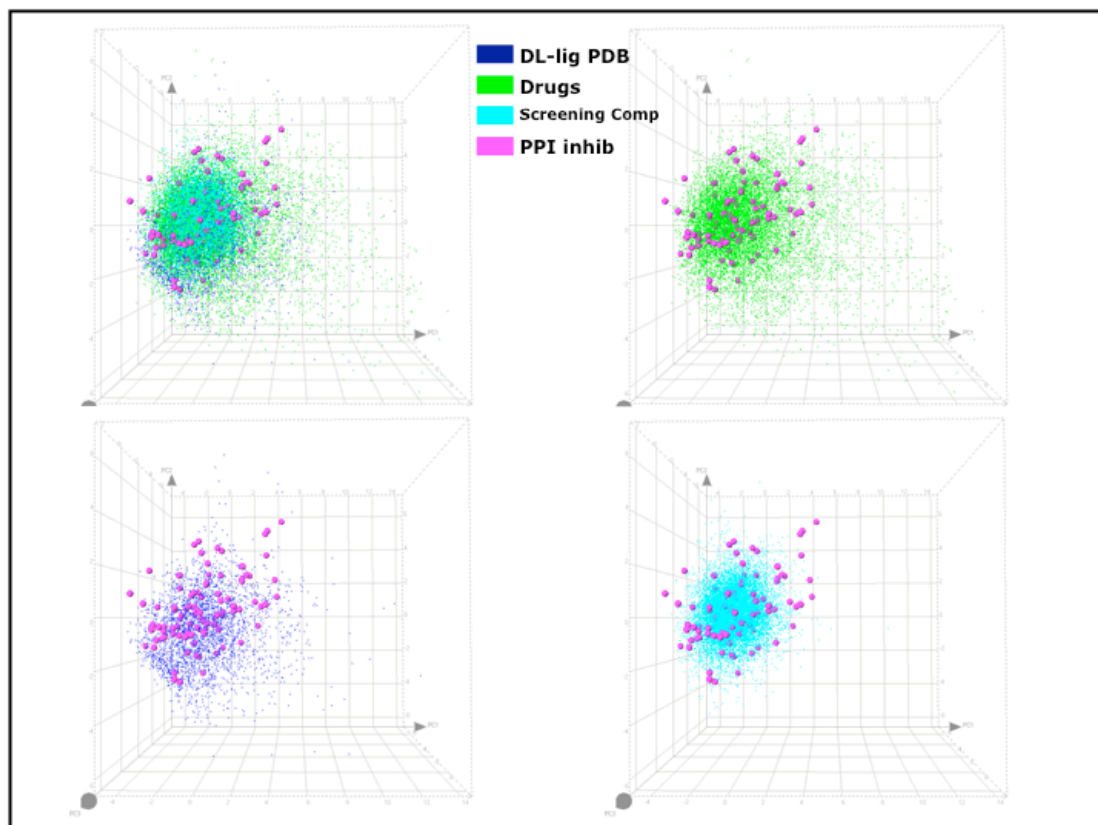


Figure 2.7. Three-dimensional projection of the principal components of the molecular properties for the different sets of molecules.

In order to have a quantitative measure of the distribution of the molecules from different sets within this PCA space (3-dimensional: PC1, PC2, PC3), the distance from the arithmetic centre of this space (distFC) has been calculated for all molecules. The average of this distance is 2.18 with a standard deviation of 1.49. Figure 2.8 shows the distribution of this distance for each set of molecules. As we have seen previously, in terms of molecular properties TIMBAL molecules are closer to developed drug molecules rather than starting point molecules (screening compounds).

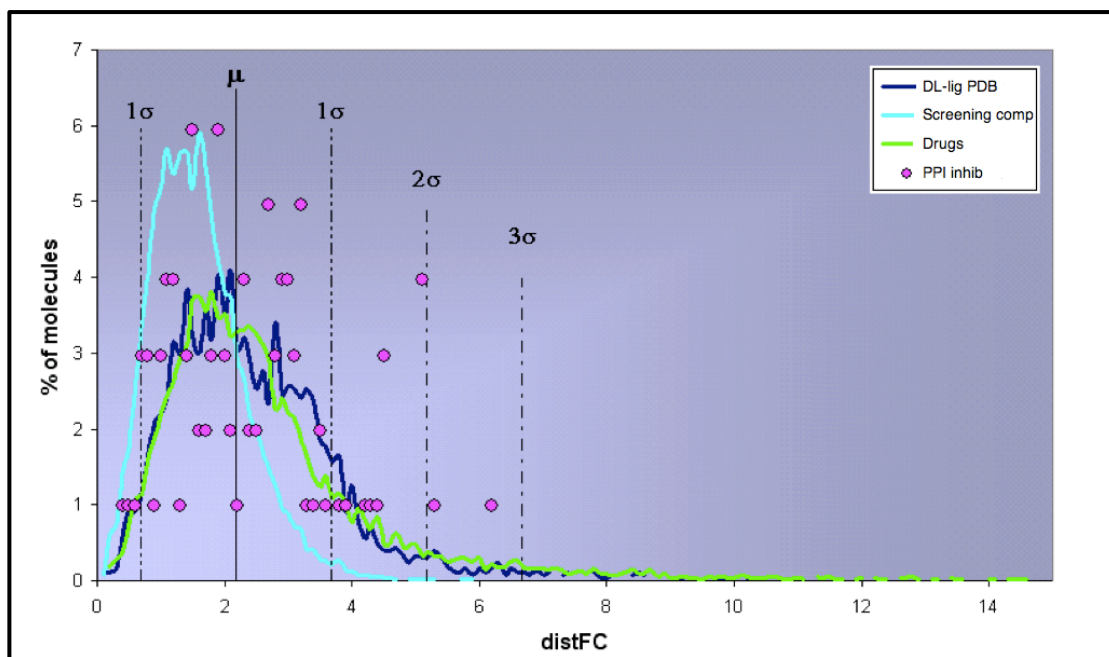


Figure 2.8. Distribution of the distances to the arithmetic centre of the PCA space for each set of molecules. TIMBAL molecules represented by dots for clarity. The mean of this distance is 2.18 with a standard deviation of 1.49. Table 2.4 shows the percentage of molecules in each bin.

$\mu = 2.18,$ $\sigma = 1.49$	% in Screening comp	% in DL-lig PDB	% in Drugs	% in PPI inhib
$\mu \pm 1\sigma$	92	84	80	83
$\mu \pm 2\sigma$	8	12	12	15
$\mu \pm 3\sigma$	0	3	4	2
$\mu \pm \text{more } 3\sigma$	0	1	4	0

Table 2.4. Percentage of each molecule set for each standard deviation bin in the distribution of distances to the centre in the PCA space. Distributions are shown in Figure 2.8.

In conclusion, molecules disrupting protein-protein interactions tend to be big lipophilic molecules with fewer hydrogen bonds than the average drug-like molecules. In order to assess whether these characteristics are due to surface complementarity and molecular recognition, the next section analyses the interface of these complexes.

2.3.2.2 Pharmacophoric analysis of the interface

2.3.2.2.1 Chemical functionality

In terms of chemical functionality, TIMBAL molecules contain more carboxylic acids and sulfonamides and less ether groups than drugs but similar proportions to the DL-ligands from the PDB set. Carboxylic acids are present in drugs, in DL-ligands from the PDB and TIMBAL sets but underrepresented in the Screening compounds. As reported previously (Whitty *et al.* 2006), PPI inhibitors tend to contain aromatic rings. Table 2.2 shows they have the highest content of aromatic rings and phenyls of the different sets. Perhaps the most surprising result is the high nitro-group content of TIMBAL molecules. This holds true across series, originator environment and target. Six out of the 17 TIMBAL targets have molecules with a nitro group. In general, aromatic nitro groups are avoided in drug development due to toxicity problems when the nitro group is reduced in the body (Boelsterli *et al.* 2006). This could explain the lower proportion of nitro groups in the drug like sets. However, nitro aromatic rings are poor in electrons due to the strong electron withdrawing effect of the nitro group. This result could lead to further investigation of the properties of these rings in the context of the molecular recognition at the protein interfaces. Table 2.5 shows the details of the functional group analysis for the different sets of molecules.

	PDB Ligands	PDB Ligands-DL	Drugs	Screening compounds	PPI Inhibitors
OH	71	52	46	13	41
NH	68	68	64	68	67
Aldehyde	3	2	0	0	1
Ketone	9	12	14	10	17
Amide	37	42	44	70	54
Alkyne	1	1	2	0	3
AkylHalide	1	1	1	1	0
Carbamate	3	3	5	1	2
Imide	5	6	5	2	6
COO	28	23	16	2	22
Epoxide	1	1	1	0	0
Ester	10	11	18	17	8
Ether	24	29	41	42	29
Pyridine	7	11	14	13	7
4ari_Nitrogen	1	0	1	0	0
1ari_aniline	2	2	2	0	2
2ari_aniline	8	15	13	31	33
3ari_aniline	4	7	12	14	17
Acetal	5	4	5	4	3
Butyl	6	2	6	2	3
CF3	2	4	5	5	3
Cyano	1	3	4	6	2
IsoPropyl	7	8	8	5	12
Nitro	2	3	3	7	15
tBu	2	3	4	3	5
sulfonamide	6	12	7	16	17
thioether	4	3	5	19	12
Phenol	9	15	9	2	11
urea	8	6	8	7	5
AminoPyr	2	5	2	1	2
sulfoxide	0	0	1	0	0
N	81	85	91	98	88
O	93	91	93	96	89
S	21	26	26	54	25
Halogen	18	30	35	43	42

Table 2.5. Percentage of molecules per set that have a least one of the functional groups present in the chemical structure.

2.3.2.2.2 Atomic contacts

With respect to the types of interactions and contacts favoured by TIMBAL molecules, Table 2.6 shows the average and standard deviation of the contact types (in percentage) extracted from the CREDO (PDB ligand subsets analysed so far) and PICCOLO databases. TIMBAL molecules considered in this analysis are the subset present in CREDO (in total 27 molecules from 26 PDB entries, see Table 2.7 for details). The last column of Table 2.6, PICCOLO(T), shows the average numbers of contacts for the protein interfaces of the multi-protein complexes that are disrupted by TIMBAL molecules (in total 28 interfaces from 16 PDB entries, note only relevant interfaces are considered, see Table 2.7 for details).

Table 2.6 shows that on the average the 27 molecules co-crystallised with protein interfaces from multi-protein complexes (see column headed TIMBAL) present more hydrophobic and aromatic and less hydrogen bond contacts than the average CREDO-DL and PICCOLO interfaces. This result correlates with the molecular property profile described in section 3.2.1 for these molecules and with slightly more hydrophobic character than the TIMBAL interfaces in PICCOLO. An interesting result that emerges from this analysis is that the PICCOLO interfaces (slightly higher even in the TIMBAL subset) are more ionic in character than the CREDO (including TIMBAL) molecules. This result will be further investigated in chapter 3, as the contacts in PICCOLO and CREDO are calculated with different algorithms and estimation of the ionisation state for small molecules is not a trivial task.

<i>avg ± STD</i>	CREDO	CREDO-DL	TIMBAL	PICCOLO	PICCOLO(T)
Covalent	3.1±9.6	0.1±0.6	0±0	0.1±0.5	0.0±0.12
vdW	54.2±22.4	60.4±11.7	56.1±8.0	52.3±17.7	47.0±8.8
vdWclash	34.2±23.5	15.7±9.1	11.7±6.5	11.9±9.0	11.9±5.8
Hbond	7.3±10.1	5.3±4.9	2.3±3.1	4.2±4.9	3.3±1.3
Ionic	7.1±18.2	4.1±8.5	2.5±4.2	13.1±17.4	16.1±8.8
piCation	0.1±0.6	0.1±1.0	0.4±1.0	8.9±13.6	8.1±7.7
Aromatic	2.3±7.0	7.8±12.1	9.3±9.3	6.8±12.5	6.8±8.1
Hydrop.	9.8±20.6	37.3±22.4	53.0±13.9	33.8±19.2	41.2±9.9
Buried_PA	360±193	281±135	386±106	1788±1909	1953±1106
Surface_A	613±285	533±246	754±167		

Table 2.6. Average and standard deviation of the contact types in percentage extracted from the CREDO (lig-protein complexes, and DL 'drug-like' subset) and PICCOLO (protein-protein complexes) databases. For instance, TIMBAL having 57% of hydrophobic contacts means that the TIMBAL molecules on average have 57% of the total contacts as hydrophobic. TIMBAL molecules considered in this analysis are the subset present in CREDO (in total 27 molecules from 26 PDB entries, see Table 2.7 for details). The last column, PICCOLO(T), shows the average numbers of contacts for the protein interfaces of the multi-protein complexes that are disrupted by TIMBAL molecules (in total 28 interfaces from 16 PDB entries, note only relevant interfaces are considered, see Table 2.7 for details). Buried_PA: buried protein area upon binding. Surface_A: surface area of the small molecules in the binding conformation. Values in bold denote significant differences between TIMBAL and CREDO-DL ($P < 0.05$).

Target	Complex	PDB PICCOLO	PDB CREDO
IL-2	IL-2/IL-2Ra	1Z92 (A:B)	1M48
			1PW6
			1PY2
			1YSG
			1YSI
			1YSN
			1YSW
Bcl-2 Bcl-XL	Bcl-2 and Bcl-XL with BAX; BAK and BID	2BZW (A:B) 1G5J (A:B) 1BXL (A:B)	2O1Y
			2O21
			2O22
			2O2F
			2O2M
			2O2N
			2YXJ
MDM2	p53-MDM2	1YCR (A:B) 1T4F (M:P)	1RV1
			2AXI 1T4E
CD80 (B7-1)	CD80-CD28 (or CTLA4)	1I8L (A:C) 1I8L (B:D)	
S100B	S100B-p53	1DT7 (A:X) 1DT7 (B:Y)	
TNFa	TNFa trimer	1TNF (A:B)	2AZ5
XIAP	XIAP/Caspase9 or SMAC (BIR3 domanin)	1G3F (A:B)	1TFQ
			1TFT
Beta-catenin	BetaCatenin/Tcf4 and Tcf3	1JPW (A:D) 1JPW (B:E) 1JPW (C:F)	
ZipA	ZipA-FtsZ	1F47 (A:B)	1S1J
			1S1S
			1Y2F
			1Y2G
c-Myc/Max	c-Myc/Max	1A93 (A:B)	
E2	E1-E2-DNA	1TUE (A:B) 1TUE (D:E) 1TUE (F:G) 1TUE (H:J) 1TUE (K:L) 1TUE (M:Q)	1R6N
iNOS	iNOS dimerization	3NOS (A:B)	1DD7
UL42	UL30(Pol)-UL42 subunits of HSV	1DML (A:B) 1DML (C:D) 1DML (E:F) 1DML (G:H)	

Table 2.7. Summary of the PDB entries for the PPI systems used in this analysis. PDB entry 1YSG has 2 SM.

2.3.2.2.3 Buried surface area

Buried surface area by TIMBAL molecules is higher on average than the CREDO-DL molecules, but this is due to the bigger size of TIMBAL molecules. In fact the ratio between buried surface and molecular surface is the same for both sets.

2.3.2.2.4 Ligand efficiency

Finally, the affinity data for the TIMBAL molecules have been analysed in terms of the Ligand Efficiency (LE (Hopkins *et al.* 2004)). The threshold described by Wells and McClendon (Wells *et al.* 2007) for the most optimised small molecules inhibitors of protein-protein interactions with structural data is 0.24.

TIMBAL holds at the moment 76 affinity data points for all the targets present in the database. Figure 2.9 shows the spread of Ligand Efficiency per target. Most of the data points fall in the range of 0.15 - 0.35 LE with an average LE of 0.27 with a standard deviation of 0.10 for all 76 molecules. Three targets are above this average, XIAP with 5 molecules averaging 0.40 (range 0.29-0.57), beta-catenin with 4 molecules averaging 0.37 (range 0.18-0.6) and CD80 with 4 molecules averaging 0.37 (range 0.36-0.38). None of these three targets were considered in the analysis of Wells and McClendon (Wells *et al.* 2007).

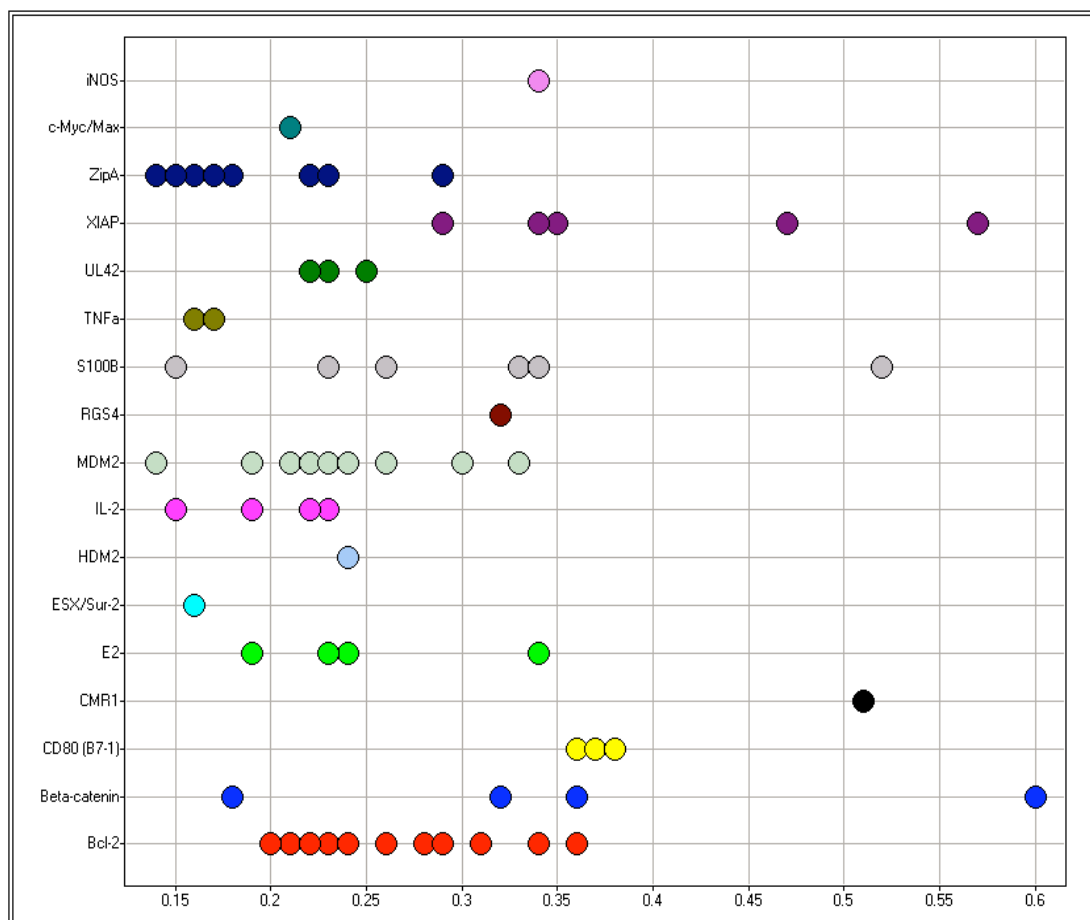


Figure 2.9. Range of Ligand Efficiency, LE (X axis) of the TIMBAL molecules separated by target.

As mentioned previously, only one TIMBAL molecule has reached phase-I/II clinical trials. Therefore, it is interesting to compare these LE values with typical ranges of LE in a hit to lead medicinal chemistry campaign for traditional targets. Table 2.8 shows these ranges. The average LE of 0.27 is reached for these TIMBAL molecules with an average of 30 atoms. Therefore, TIMBAL molecules are slightly less efficient than typical medicinal chemistry leads with the same number of atoms (LE range 0.32)

Kd(μ M)	MW	N atoms(*)	LE
10	200-250	15-19	0.46-0.36
1	250-300	19-23	0.44-0.36
0.1	300-400	23-30	0.43-0.32
0.001	500	38	0.33

Table 2.8. Range of affinities and sizes for a typical medicinal chemistry campaign from hit to lead. Ligand Efficiency and number of atoms are calculated following the original paper.

In addition, Figure 2.10 shows the average molecular property values for TIMBAL molecules binned by LE. By definition of LE the black bars (average of number of atoms) in Figure 2.10 should decrease for higher LE, in fact one can see this trend, as well as alogP which correlates with molecular weight (and therefore with number of atoms). The interesting result of this plot is that more efficient binders as well as the lesser ones have virtually the same average of hydrogen bond features. This recalls the previous result that TIMBAL molecules make fewer hydrogen bond contacts, but what this plot suggests is that the few hydrogen bonds achieved by the efficient binders should be kept. Moreover, the hydrogen bond features remain more or less constant with the increase of molecular size. This observation is in agreement with those made by Olsson et al. (Olsson *et al.* 2008) and it will be explored in detail in chapter 4.

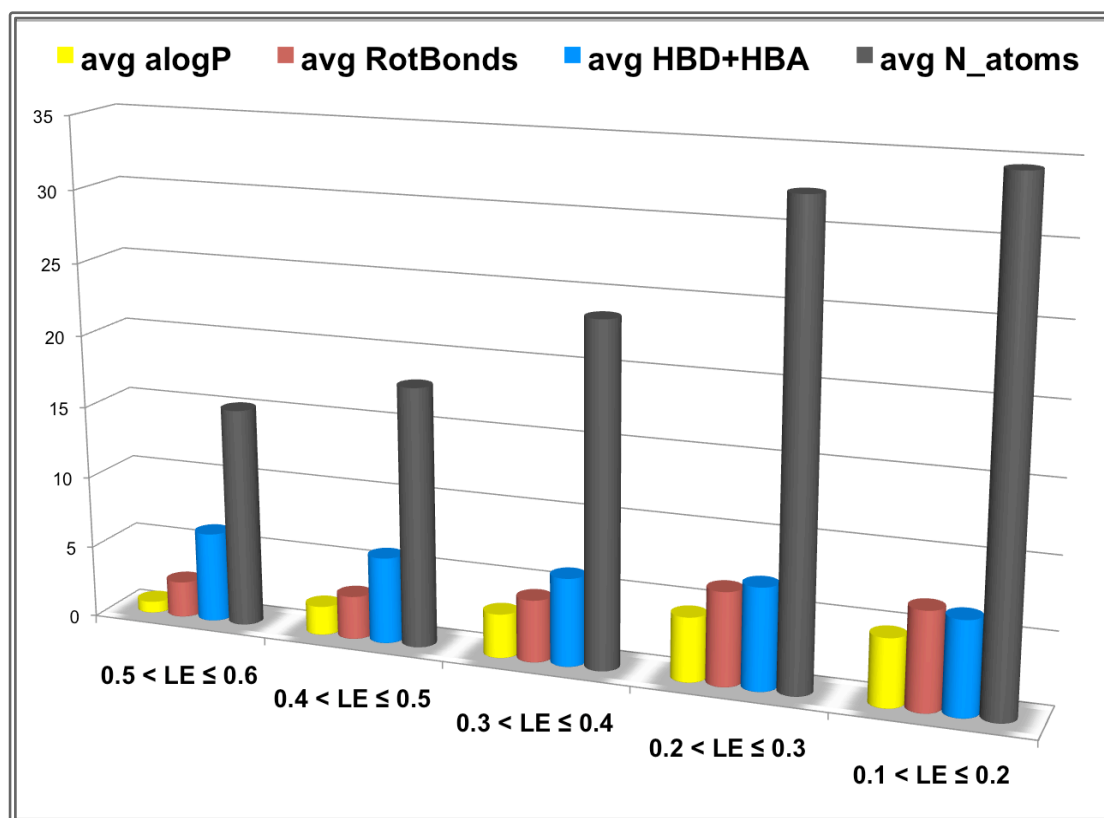


Figure 2.10. Average of the molecular properties for TIMBAL molecules binned by LE. Blue: Average of the sum of hydrogen bond donors and acceptors. Red: Average of rotatable bonds. Yellow: Average of alogP. Black: Average of number of atoms.

The calculated ΔG from the affinity data in the TIMBAL database has been plotted against all molecular properties and contact types (16 TIMBAL molecules with affinity data are also present in CREDO), but no correlation has been found. Further filtering and classification might be needed to extract meaningful relationships, like for example compare contacts per surface area (suggestion from Richard Bickerton, personal communication) rather than absolute numbers.

2.4 Conclusions

I have described the creation of a database containing small molecule modulators of protein-protein interactions. The database has been profiled and compared with other sets of molecules and interactions. TIMBAL molecules tend to be bigger, more rigid, more lipophilic and less hydrogen bonding than molecules in the drug and screening sets. This result is consistent with types of interactions these molecules make; as has been discussed, TIMBAL molecules present more hydrophobic and aromatic and less hydrogen bond contacts than the average CREDO-DL and PICCOLO interfaces. In terms of functional groups, protein-protein modulators seem to favour nitro groups, carboxylic acids and sulfonamides. LE for these molecules has been found to be slightly lower than the typical hits and leads from more traditional targets.

Several analyses have highlighted the ease with which medicinal chemistry programs can deliver high affinity molecules by increasing the lipophilicity, see for example (van de Waterbeemd *et al.* 2001; Leeson *et al.* 2007; Keserü *et al.* 2009). However, increasing the hydrophobicity of small molecules also increases the likelihood of a compound failing in the development phase (Leeson *et al.* 2007). This observation seems to be particularly relevant for protein-protein targets. These molecules are not classical drug-like molecules, and their profile suggests they may not be selective binders either. Lesson and Springthorpe (Leeson *et al.* 2007) have shown in their analysis of small molecule drugs a positive correlation between clogP and promiscuity. However, these are the first small molecule modulators of a class of targets long believed to be undruggable. Drug-like properties are derived from drugs developed many years ago that hit a limited set of historical targets. Perhaps druggability as well as selectivity have to be addressed on a case-by-case basis for this diverse target class, see for example ABT-263/Bcl2 (Morelli *et al.* 2011). My conclusion is that we should continue to collate and connect the available information regarding the

small molecules and systems they modulate in order to extract any trends and thereby be in a better position to develop new therapeutic agents for these emerging targets. This will be the focus of the following chapters.

3.1 Introduction

Structural databases are powerful resources to study molecular interactions. In chapter 2, we have seen an application of these resources, namely CREDO (protein-ligand interactions (Schreyer *et al.* 2009)) and PICCOLO (protein-protein interactions (Bickerton *et al.* 2011)), which I used to compare atomic contacts of different sets of molecules. However, atomic contacts are not defined in the same way in these two databases. Indeed, one of the results reported in chapter 2 – that protein-protein interfaces have a more ionic character than those of protein-small molecules - prompted me to examine in detail the contact definitions and frameworks of the two databases. Using the definitions in the original databases, 13% on average of the atomic contacts made at protein-protein interfaces were ionic compared to 7% of those at the protein-ligand interfaces. If genuine, that was a remarkable result. Thus, efforts were made to make sure the differences in contact patterns were really due to differences between the molecules and not due to database definitions. My objective in this chapter is to examine the differences of definitions between the structural databases used in this thesis and to resolve problems where they arise.

3.1.1 CREDO

CREDO is a comprehensive database of protein-ligand interactions, storing structural data, sequence annotation and chemical information (Schreyer *et al.* 2009). It is the centre of Adrian Schreyer's PhD thesis (Schreyer 2010), and it was developed as a resource to support drug discovery and virtual screening.

CREDO identifies ligands in the Protein Data Bank (PDB) through information stored in the mmCIF dictionaries (macromolecular Crystallographic Information Files). Ligands are either single residues in the non-polymer entities, or short polypeptides up to eight residues long. These

ligands are used to extract ligand-protein complexes from the PDB. CREDO does not consider entries with ligands only, nucleic acid ligand complexes, protein backbone-only structures and entries violating the PDB format.

The interatomic protein-ligand contact data are then derived using OpenEye's OEChem toolkit (<http://www.eyesopen.com>). With this toolkit, all protein atoms within a radial distance of 6.5Å to any ligand atom are found. Atom types are assigned and hydrogens atoms are added to classify each atom pair into the following non-exclusive contact types: covalent, van der Waals, van der Waals clash, hydrogen bond types depending on atom types and geometries, halogen bond, ionic, metal complex, pi-cation, pi-donor, pi-carbon, aromatic types depending on the geometries of the aromatic rings involved, hydrophobic and carbonyl. CREDO was updated weekly in an automated manner until April 2010.

3.1.2 PICCOLO

PICCOLO is a comprehensive database of structurally characterized protein interactions; storing structural data and sequence annotation (Bickerton *et al.* 2011). It is the main focus of Richard Bickerton PhD thesis (Bickerton 2009), and it was developed as a resource for protein modelling including protein-protein docking, prediction of the effect of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability and function, derivation of environment-specific substitution tables and analysis of hot spots at protein interfaces.

PICCOLO derives data from the mmCIF dictionaries at structure, chain and residue levels. Structures are handled with the PDB module in BioPython (Hamelryck *et al.* 2003) and are "sanitized" into clean PDB flat files to ensure every protein residue is uniquely identified and inconsistencies are removed. Only polymer protein standard residues are considered and the three most common non-standard amino acids are modified into their standard analogue.

These are selenomethionine (MSE), methyllysine (MLY) and hydroxyproline (HYP). PICCOLO uses the PISA resource at the EBI (Krissinel *et al.* 2007) to generate quaternary assemblies from these clean files. PICCOLO considers protein-protein interactions as pairwise interactions between chains in the generated files. Therefore, PICCOLO has two flavours: "PDB" that stores interactions between chains of the entry in the PDB as they are in the asymmetric unit, and "Quaternary" that stores interactions between the chains of the quaternary assemblies generated by PISA transformations.

The interatomic protein-protein contact data are derived in a pairwise manner between two distinct chains in the same structure. The PDB module in BioPython is used to find all atoms in one chain that neighbour any atom in the second chain with a cut-off distance of 6.05Å. PICCOLO structures are composed of only 20 standard amino acids, thus atom types and atomic radii are tabulated and manually curated. HBPLUS (McDonald *et al.* 1994) is used to derive hydrogen bonds and water-mediated hydrogen bonds. The other contact types assigned in PICCOLO are: covalent, van der Waals, van der Waals clash, ionic, pi-cation, several aromatic types depending on the geometries of the aromatic rings involved, hydrophobic, disulphide and aromatic-sulfur. PICCOLO did not have an update procedure in place and the exponential growth of the PDB required the group to provide one. I undertook this responsibility by testing and merging more than 30 scripts (written by Richard Bickerton to create PICCOLO) into a single procedure that runs monthly. Information about this process can be found at: http://tetra.bioc.cam.ac.uk/mediawiki/index.php/PICCOLO_AliciaNotes

3.2 Methods

3.2.1 PICCOLO-CREDO intersection

The fact that CREDO considers polypeptides up to eight residues long as ligands allows straightforward comparison between both databases. Structures with these short peptides are the intersection of PICCOLO and CREDO. The number of contacts (for each common contact type) from each resource can be plotted against each other. If the databases are identical these plots will show a straight line of slope one.

Further filtering of these structures is required, as PICCOLO considers only standard amino acids and CREDO considers only the asymmetric unit deposited in the PDB. For these reasons, the contacts analysed here are from the PDB flavour of PICCOLO, and from CREDO only protein-ligand complexes with standard amino acids are considered. PICCOLO stores the interatomic interactions from pairwise protein chains; therefore the queries used in CREDO consider pairwise interactions only. For example, for a ligand interacting with two different protein chains, the query retrieves the contacts to compare them with PICCOLO and then sums them by ligand-one-chain-protein at the time instead of summing all the interactions that the ligand presents, which is the normal philosophy in CREDO. The subset for the PICCOLO-CREDO comparison is composed of 962 pairs from 468 distinct PDB entries, summing more half a million atomic contact pairs.

3.3 Results and discussion

Figure 3.1 shows the differences found between the original PICCOLO and CREDO databases for the subset of structures that were present in both.

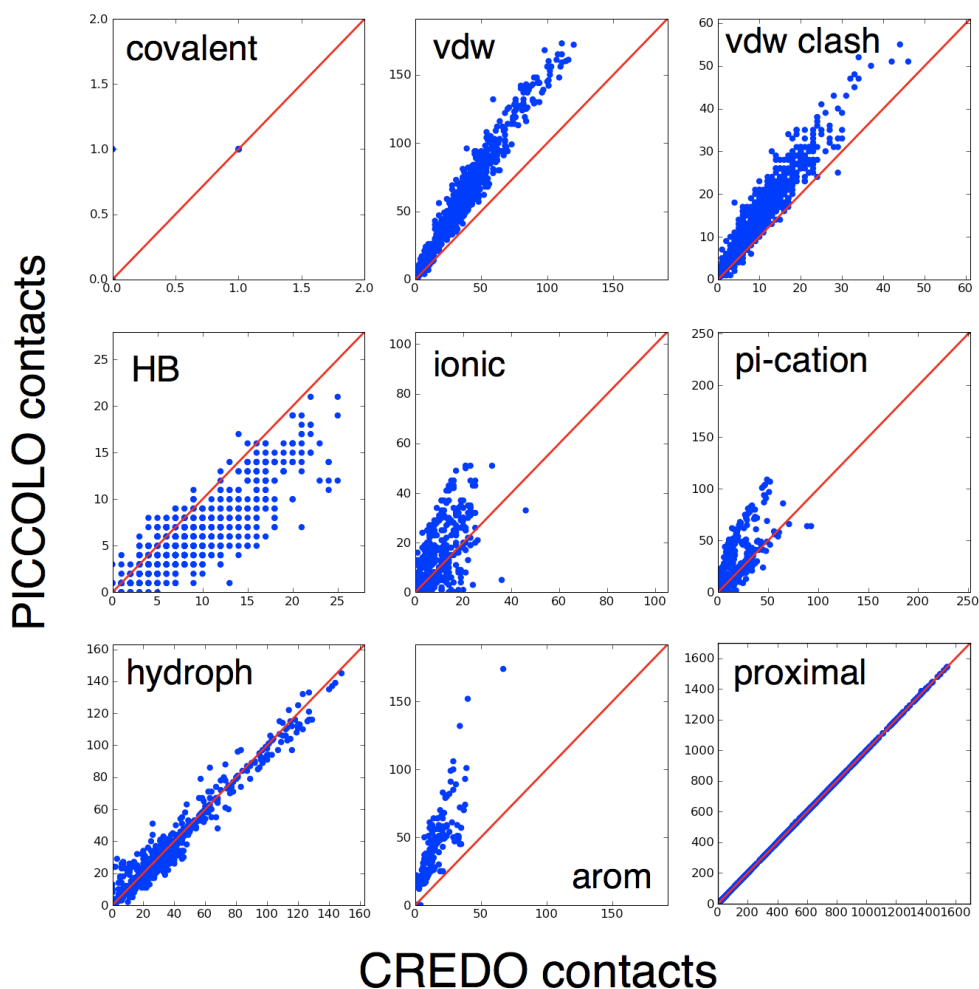


Figure 3.1. Scatter plots of the comparison of PICCOLO and CREDO contacts. In all nine plots, X-axes are for CREDO contacts and Y-axes for PICCOLO contacts. Each scatter plot is for one of the common contact types in both databases, from top left to bottom right: covalent, van der Waals, van der Waals clash, hydrogen bond, ionic, pi-cation, hydrophobic and proximal. Proximal is defined as when the two atoms are less than or equal to 6.05Å apart, the maximal distance of a water-mediated hydrogen bond. The red line in each plot denotes the slope that is given when the two databases give identical results.

3.3.1 Stored distances

The scatter plot of proximal contacts for CREDO and PICCOLO in Figure 3.1 shows that there were minor differences between databases. Two factors were found to explain these differences. First, CREDO stores distances with two decimal places whereas PICCOLO stores distances with three decimal places. Secondly, from the half a million contacts studied in this comparison, 0.06% of them had different stored distances, i.e. the absolute difference between them was greater than 0.005. One of the reasons for this difference is the conformers for certain residues that CREDO retains and PICCOLO cleans up. In addition, the fact that PICCOLO pre-processes PDB files and uses BioPython to extract neighbour atoms and distances, whereas CREDO uses raw PDB files and uses the OEChem toolkit causes minor differences in recorded distances too.

Although there are discrepancies between databases caused by these stored distance issues, the proportion of these differences is small and the differences are comparable to the standard experimental error.

3.3.2 Atomic radii

The differences in covalent, van der Waals and van der Waals clash contacts seen in Figure 3.1 are due to different atomic radii used for each database. Both databases use the same criteria to define these contact types (essentially sum of radii). However, CREDO used the radii from the OEChem implementation, which in turn uses data from the Cambridge Crystallographic Data Centre (CCDC) for covalent atomic radius and from (Bondi 1964) for van der Waals radii. PICCOLO used, for both covalent and van der Waals radii, the set of residue-specific atomic radii from (Tsai *et al.* 1999).

The resolution of these differences was for CREDO to use the same residue-specific atomic radius as PICCOLO for the protein atoms. Therefore,

extra tables with hybridisation labels and PICCOLO radii for each atom type in the 20 standard amino acids were generated to allow CREDO to use the same van der Waals radii for protein atoms.

3.3.3 Ionic, pi-cation, hydrophobic and aromatic contacts

For ionic, pi-cation, hydrophobic and aromatic contact types, both databases initially used the same SMARTs queries to label atoms as positive ionisable, negative ionisable, aromatic and hydrophobe. Continuous and independent development of the resources led to divergence in the initial queries and certain atoms were labelled differently. A detailed assessment of the atom labels for the 20 standard residues in both databases was performed, and atom types were modified to be identical in both CREDO and PICCOLO. In addition, the distance cut-off criterion for each contact type was different in each database. Therefore, a consensus for these distances was also reached. Table 3.1 describes details of these contact criteria.

	PICCOLO	CREDO	Consensus	References
Ionic	d(pi-ni) <= 6 Å	d(pi-ni) <= 4 Å	d(pi-ni) <= 4 Å	(Barlow <i>et al.</i> 1983) (Marcou <i>et al.</i> 2007)
Pi-cation	d(pi-ar) <= 6Å	d(pi-ce) <= 4Å	d(pi-ce) <= 5Å	(Gallivan <i>et al.</i> 1999)
		arcsin >= 30		
Hydrophobic	d(hyd-hyd) <= 5Å	d(hyd-hyd) <= 4.5Å	d(hyd-hyd) <= 4.5Å	(Tina <i>et al.</i> 2007) (Marcou <i>et al.</i> 2007)
Aromatic	d(ar-ar) <= 6Å	d(ar-ar) <= 4Å	d(ar-ar) <= 5Å	(Chakrabarti <i>et al.</i> 2007) (Marcou <i>et al.</i> 2007)

Table 3.1. Details of the criteria for the different contact types. Distances are between atom types: pi (positive ionisable), ni (negative ionisable), ar (aromatic), ce (centroid of aromatic ring), hyd (hydrophobe).

The differences in distances indicate that contact definitions are not canonically established. These definitions depend of the context in which they are applied. For instance, the original CREDO criteria were tighter than those used in PICCOLO. This can be understood in terms of the accuracy and disorder level of the side chains of a small molecule-binding site in comparison with the side chains at protein-protein interfaces. In addition, resolutions of the crystal structures deposited in the PDB differ. Therefore, the exact numerical distance used is less important than the consistency across comparisons.

3.3.4 Hydrogen bonds

The differences in the hydrogen bond contacts were due to different algorithms used to calculate these contact types. CREDO uses OEChem to add hydrogens to the structures and SMARTs (SMiles ARbitrary Target Specification, www.daylight.com/dayhtml/doc/theory/theory.smarts.html) queries, to label heavy atoms as donors or acceptors. A contact is then labelled as a hydrogen bond if it meets the distance and angle criteria described in Table 3.2. PICCOLO uses an external program HBPLUS (McDonald *et al.* 1994) to assign hydrogen bonds. This program adds hydrogen to the structures, assigns donor, acceptor, donor antecedent and acceptor antecedent labels to heavy atoms, and calculates hydrogen bond contacts with the geometry criteria described in Figure 3.2 and Table 3.2. In addition, for structures with resolution greater than 1.0Å, hybridisation and atom type of the atoms in the side chains of asparagine, glutamine and histidine are also assigned to optimise hydrogen bonds for these residues.

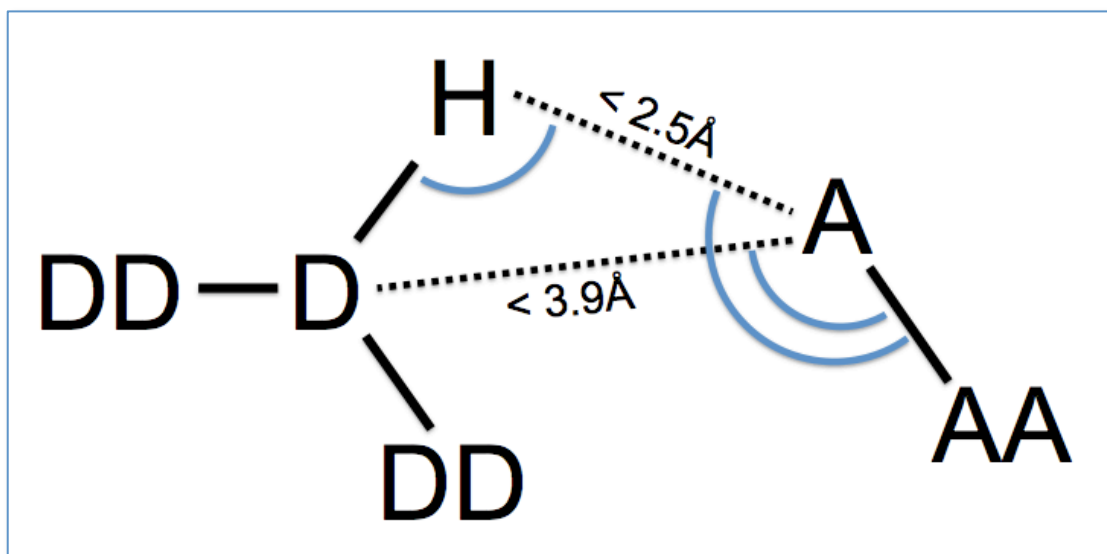


Figure 3.2. Geometric criteria for hydrogen bonds used in HBPLUS, adapted from figure 1 in (McDonald *et al.* 1994). D is the donor heavy atom. H is hydrogen, A is the acceptor heavy atom. DD is donor antecedent (an atom two covalent bonds away from the hydrogen). AA is acceptor antecedent. All three angles highlighted in the figure are required to be greater than or equal to 90 degrees to meet the hydrogen bond criterion.

	PICCOLO	CREDO	Consensus
Atom types	D, DD, A and AA from HBPLUS, see Fig 3.2	Donor and acceptor from SMARTS queries	Not possible
Distance	$d(D-A) \leq 3.9\text{\AA}$ $d(H-A) \leq 2.5\text{\AA}$	$d(D-A) \leq 3.6\text{\AA}$	$d(D-A) \leq 3.9\text{\AA}$
Angle	$a(D-H-A) \geq 90$ $a(H-A-AA) \geq 90$ $a(D-A-AA) \geq 90$	$a(DHA) \geq 120$	$a(DHA) \geq 90$

Table 3.2. Details of the original hydrogen bond calculation in PICCOLO and CREDO databases and the consensus achieved.

The initial discrepancy in geometrical criteria used in the two databases was resolved. Both databases currently use the same distance and angle criteria for hydrogen bonds. However, after these modifications, the differences remained and the CREDO hydrogen bond count was greater than the PICCOLO hydrogen bond count for the subset studied, as described below.

Unfortunately deriving hydrogen bond contacts from PDB files is not a trivial task. On the one hand, most structures deposited in the PDB do not have hydrogens, therefore algorithms to add them need to be in place taking into account the likely uncertainties in the structural models. For example, the tautomeric form of histidine and the true positions of the O and N atoms in the amide groups of asparagine and glutamine will all depend on their environments. In the case of proteins where there is a finite set of building blocks, this challenge can be addressed with a high percentage of success using algorithms like HBPLUS. However, this is not easily solved for small molecules where the diversity and lack of connectivity information for them in the PDB files means that a pragmatic approach must be taken.

In PICCOLO, computation of hydrogen bonds between proteins is achieved using the aforementioned HBPLUS program, which gives high specificity (low rate of false positives). In CREDO, the addition of hydrogens and donor-acceptor labelling uses the OEChem toolkit. Due to the difficulty of estimating pKa and tautomerism for small molecules in protein environments, the calculation of hydrogen bonds is somewhat more generous and less specific than that for the protein complexes. In practical terms, comparison across databases is not possible for these types of contacts.

3.3.5 Creation of simple contact definitions

Although consensus has been achieved between CREDO and PICCOLO, the issues identified with respect to computation of hydrogen bonds prevented full compatibility between the two databases. At this point, three options seemed feasible. First, use the OEChem toolkit in PICCOLO to derive hydrogen bond contacts. However, this option would be somewhat detrimental to PICCOLO and would inevitably lead to its regeneration and the maintenance of two parallel versions. In addition, the performance of the hydrogen-bond calculations for protein-ligand complexes compared to protein-protein complexes is unknown and should be investigated before

interpreting any results. The second option was to develop a hydrogen bond calculator that was not biased by molecule type. The third was to define simple contact definitions that were software and molecule-type independent. The pragmatic option chosen was the last due to the time that remained for me to complete the project. The first two options could easily be stand-alone projects.

Therefore, new tables with simple contacts were generated from the existing CREDO and PICCOLO tables. These contacts are simple distance cut-offs between atom pairs, labelled as polar or apolar depending on which atom types constitute the pair. The distance criterion used for all pairs is 4.5Å. The selection of this distance is somewhat arbitrary. As mentioned earlier, resolutions of the experimental structures analysed are not homogeneous, therefore it is best to keep this cut-off consistent and simple across comparisons (both for sets of molecules analysed and atomic types). 4.5Å was favoured as a compromise between the distances used to define hydrophobic (4.5Å) and ionic (4.0Å) contacts in the different databases. Using this distance, then, the polar and apolar contacts were defined as follows:

Protein-protein complexes

Apolar contacts: C...C, C...S, S...S (not in Cys-Cys bridges)

Polar contacts: N...O, O...O, N...N, O...S, N...S (S from Cys)

Protein-small molecules complexes

Apolar contacts: C...C, C...S, C...X, S...X (X = Cl, Br, I)

Polar contacts: N...O, O...O, N...N, O...S, N...S, N...F, O...F, S...F (S from Cys)

Figure 3.3 shows that, as expected, these simple contacts are less specific and introduce false positives. In fact, the majority of the points are above the green line of slope = 1, i.e. there is a greater number of simple contacts (y axis) than specific contacts (x axis). Nevertheless, there is a strong correlation (r value > 0.9, (Townend 2002)) between specific and simple apolar contacts, and that dominates the global correlation for the sum

of contacts. However, there is a poor correlation between specific and simple polar contacts. This can be explained by two opposing effects. Firstly, simple polar contacts do not have the geometric and atom type constraints that hydrogen bonds must meet, nor the charge complementarity that is required of ionic contacts. Secondly, polar specific contacts such as pi-cation will not be considered in the simple polar definition.

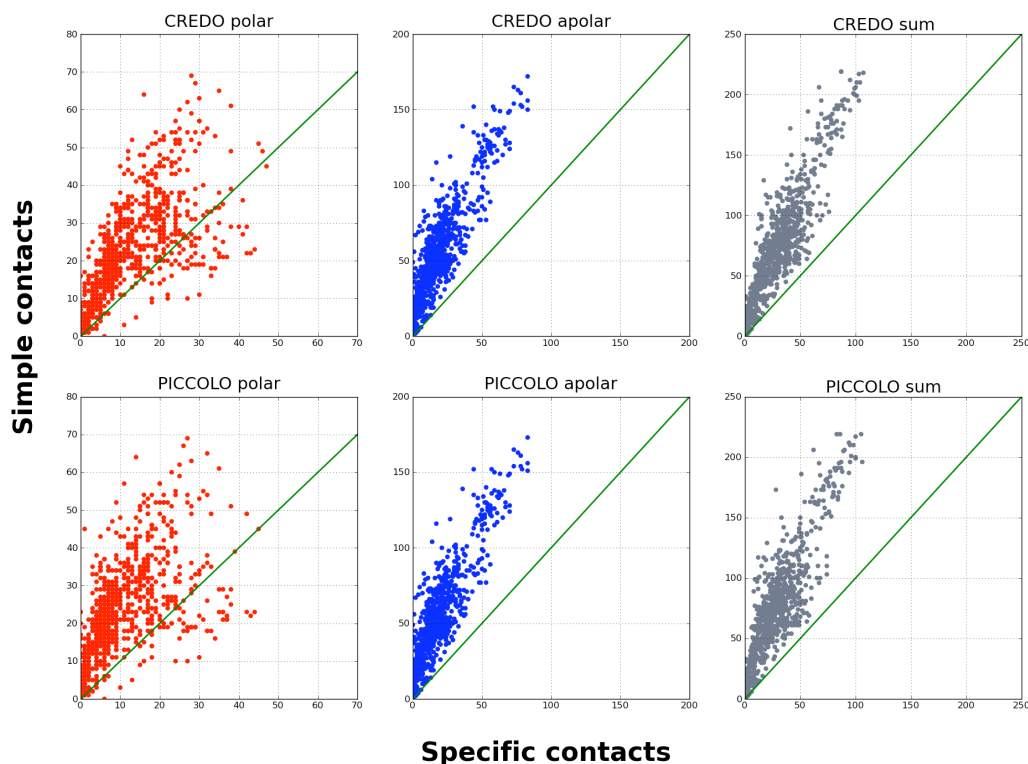


Figure 3.3. Scatter plots of specific contacts versus simple contacts for each database and type for the subset common to both databases. Simple polar and apolar contacts are distance cut-offs between polar-polar and apolar-apolar atom type as described in the text. Specific contacts refer to the contacts defined in CREDO and PICCOLO. Hydrogen bond, pi-cation and ionic are considered as polar contacts and hydrophobic is considered as apolar. The green line has a slope = 1 to aid visualisation. See Table 3.3 for details of the linear correlation.

Database	Contact type	r value	P value
PICCOLO	Sum of contacts	0.90	0.00
	Polar	0.65	0.00
	Apolar	0.90	0.00
CREDO	Sum of contacts	0.91	0.00
	Polar	0.72	0.00
	apolar	0.90	0.00

Table 3.3. r and P values from linear correlation calculations between specific and simple contacts. The P value has been rounded to zero when $P < 1E-100$.

Although much less specific, these simple contact types can unravel patterns in molecular recognition. The number of contacts at the binding interface is analogous to the burial of surface area upon binding. Indeed, Figure 3.4 shows strong correlation between the number of contacts and the buried surface area. This correlation is maintained through polar, apolar and sum of contacts with r values of 0.95, 0.90 and 0.94 respectively. All three correlations are significant with P values $< 1E-100$.

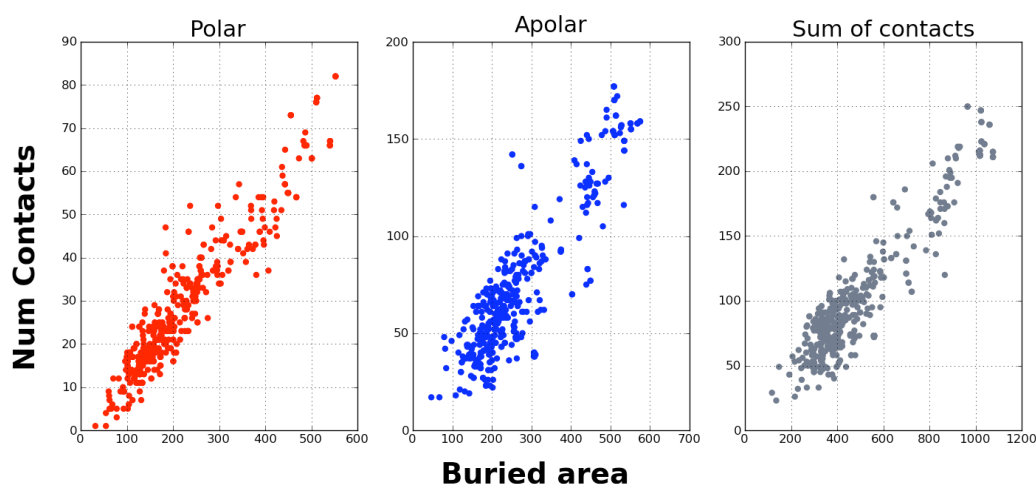


Figure 3.4. Scatter plots of buried surface area upon binding and the number of atomic contacts (polar, apolar and sum of contacts) for the subset of complexes common to both databases. The sum of contacts has been calculated over all interacting chains for comparison with buried area.

3.4 Conclusions

PICCOLO and CREDO were created with different research questions in mind. They were designed to deal with different types of molecules and therefore used different software to parse them. In addition, they are the product of PhD projects addressing specific needs to resolve these different questions. As a result, a database consolidation step was needed before performing any analysis that involved cross comparison of data from the different resources.

Detailed analyses of the database generation process and the contact definitions led us to reach a consensus in order to unify PICCOLO and CREDO. However, there were issues that could not be resolved. PICCOLO does not consider non-standard amino acids, and contacts involving them are simply not recorded. This in turn allows PICCOLO to use more accurate calculations of hydrogen bonds because it deals only with a finite number of atom types. In contrast, CREDO covers all residue types occurring in small molecule ligands and therefore cannot calculate hydrogen bonds with the same level of accuracy. On the other hand, PICCOLO stores inter-chain interactions, including assemblies predicted to be biologically relevant, whereas CREDO only considers interactions between proteins and ligands from the deposited asymmetric unit in the PDB. Inter-ligand interactions are not contemplated in CREDO. Furthermore, neither database registers interactions with nucleic acids or carbohydrates.

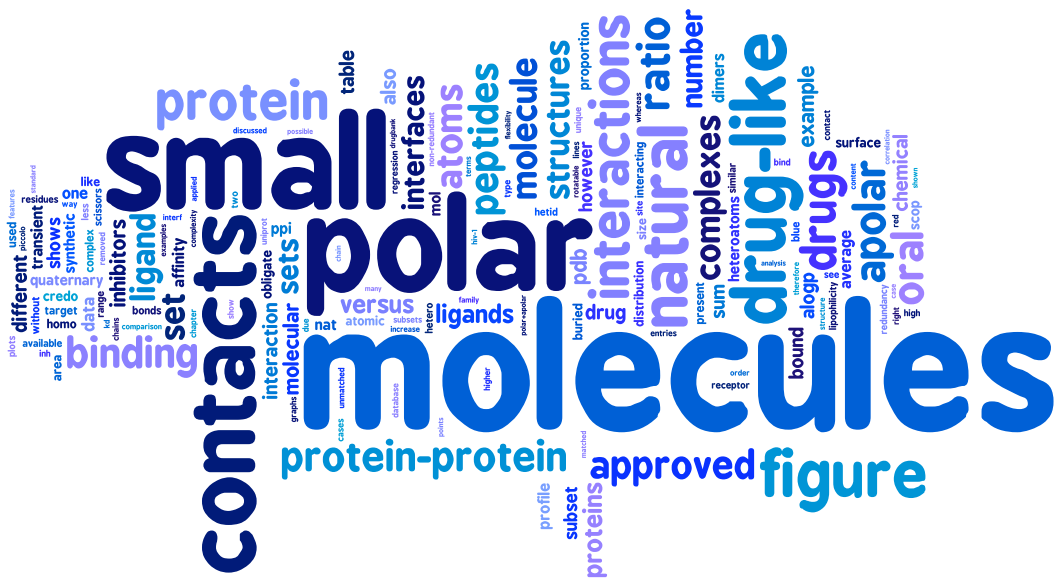
The results of this analysis and the feedback generated using both databases has helped Adrian Schreyer, now a post-doc in the group, to define a "new CREDO" database that negotiates these issues. The new database includes quaternary assemblies and considers interactions between all different entities in the PDB: proteins, polypeptides, nucleic acids, carbohydrates and other chemical entities. The database also stores interactions between the same types of entity; in this way it encloses under

one umbrella, protein-protein, protein-nucleic acids, ligand-ligand and so forth. It also stores intra-chain interactions for all atoms within a 5Å threshold distance, excluding atoms in close contact where the length of the shortest covalent bond path between them is less than three. The contact definitions are identical regardless of the type of molecules analysed. Hydrogen bonds are computed with the OEChem toolkit. This new database will be released in the first quarter of 2012. In addition, PICCOLO is maintained to keep the higher level of specificity only possible for the subset of protein-protein interactions.

Before this new resource is available, simple contact definitions have been generated. Although, these contacts are less specific, they allow cross comparisons between databases and resemble the measurement of buried surface area used in other studies, providing a coarse description of the interfaces. These contacts are the ones used in the remainder of this thesis.

Chapter 4

Structural interaction profiles of protein-protein and protein-small molecules



4.1 Introduction

In chapter 2 we have seen how the first small molecule inhibitors of protein-protein interactions are large, lipophilic and with few polar features. As I have discussed in the introduction of this thesis, lipophilic molecules are bad news for drug discovery, as they have to overcome more hurdles to become safe drugs. This in turn not only increases the cost of development but also the probability of failure as drug candidates. It seems natural to ask if this size and lipophilicity is a requirement that small molecules need to fill in order to bind to protein interfaces. The aim of this chapter is to understand how nature effects interactions in order to migrate this knowledge to the design of small molecule modulators of biological targets. However, molecular recognition laws are far from simple and unravelling their complexity is not achievable from representative frozen structures only (van Regenmortel 1999). Reality is closer to dynamic molecular ensembles living in crowded cellular environments, where solvent and local concentrations have a role that is difficult to model. In addition, multi-protein complexes present a huge diversity of protein-protein interfaces in terms of function, lifetime, size, shape, affinity, plasticity and specificity, making it almost impossible to establish common rules for all protein-protein complexes in order to translate them into the design of small molecules. However, one can elucidate general trends of molecular recognition in terms of atomic interactions from the experimentally determined structures of natural protein complexes (not only multi-protein complexes but also endogenous small-molecule protein complexes) and compare them with trends from drug-like small molecule protein complexes. In this way, we can guide the design of synthetic molecules to resemble better their natural counterparts.

I generated modified versions of our in-house databases derived from the PDB (Berman *et al.* 2000), PICCOLO (Bickerton *et al.* 2011) and CREDO (Schreyer *et al.* 2009) in order to analyse, in atomic detail, the patterns of interactions between the different classes of molecules. With the caveat that

data are from static structures instead of the dynamic ensembles, I looked into the interaction profiles that characterise different complexes namely: protein-protein, protein-natural molecules, protein-small peptides and protein-synthetic small molecules. Keeping in mind that current drug candidates and hits for protein-protein interactions are somewhat too lipophilic to succeed, it is appropriate to define these interaction profiles in terms of polar and apolar contacts, with the aim of migrating natural patterns into the design of new therapeutics.

4.2 Methods

4.2.1 General considerations and filtering

All subsets of protein coordinates were extracted from two of the in-house structural databases derived from the PDB: CREDO (protein-ligand interactions) and PICCOLO (protein-protein interactions). The PDB holds almost 75,000 (August 2011) experimentally determined structures of proteins, nucleic acids and complex assemblies. This wealth of data allows researchers to investigate the various aspects of molecular folding and recognition. However, it also brings the challenges of removing redundancy to avoid bias and data curation to minimise the noise. The size of the data held requires automated treatment for clustering and data-cleaning procedures. The approach applied here has been to minimise the amount of noise in order to have cleaner sets of molecules, even when this implies reducing the number of structures analysed.

4.2.1.1 Interactions outside of each database scope

Small molecules and small peptides were identified using the CREDO database, protein-protein interfaces were extracted from the PICCOLO database. These resources are powerful tools, but they also have limitations that have to be taken into account when comparing structures across and within databases.

For instance, neither CREDO nor PICCOLO considers interactions with nucleic acids. Therefore atomic contacts that ligands or proteins engage with nucleic acids are not recorded. A good example of these cases is 1HNX, 30S ribosomal subunit in complex with Pactamycin (Figure 4.1) (Brodersen *et al.* 2000). This structure has 22 chains, two of which are polyribonucleotide. The ligand Pactamycin (PCY, 40 heavy atoms) is interacting mainly with the ribosomal RNA (chain A) and a fragment of messenger RNA (chain X),

however it is also proximal to protein S7 (chain G) engaging in a single hydrogen bond interaction with it. In the same fashion, protein S7 is interacting with RNA but it is also proximal to protein S11 (chain K). In this case, both CREDO and PICCOLO underestimate the atomic contacts for these entities.

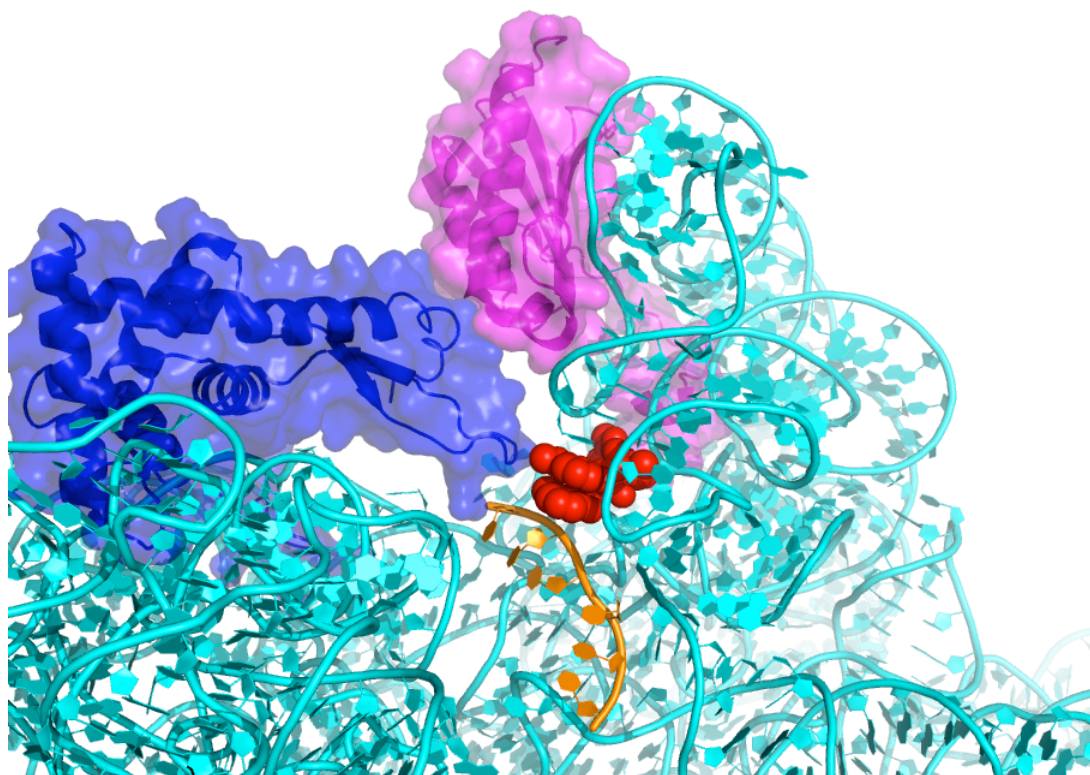


Figure 4.1. Structure 1HNX (30S ribosomal subunit in complex with Pactamycin). Small molecule ligand (PCY) represented by red spheres. Ribosomal RNA in cyan cartoon, fragment of messenger RNA in orange cartoon. Protein S7 in blue cartoon with surface and Protein S11 in magenta cartoon with surface.

The filter applied to avoid these cases was to remove structures that contain nucleic acids interacting with proteins, using BIPA database (containing 2380 structures from PDB, June 2010).

The same situation can be observed when there is a saturation of ligands in the protein crystal or solution. Proximal ligands can interact between themselves making atomic interactions that are not recorded in the database. For example, 1F6A, Fc fragment of human IgE bound to its

receptor (Figure 4.2) (Garman *et al.* 2000), where five ligands sit between two protein chains.

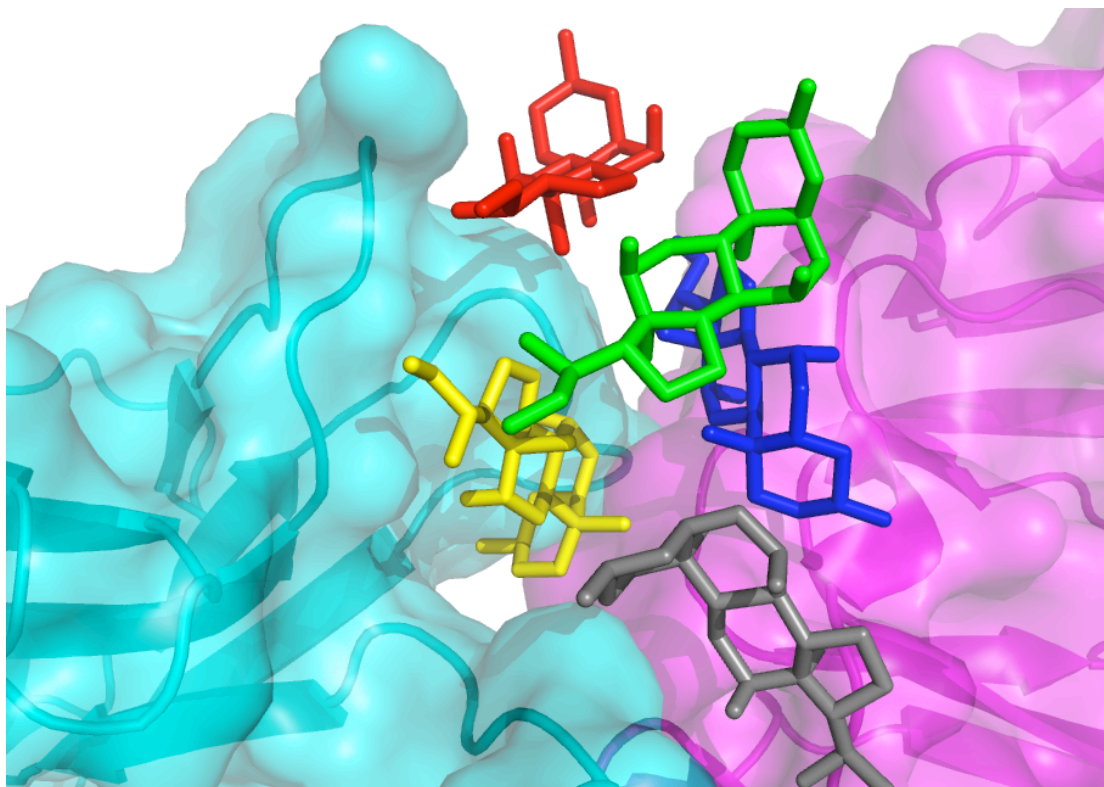


Figure 4.2. Binding interface between human immunoglobulin epsilon chain C (IgE-FC in cyan) and its high affinity immunoglobulin epsilon receptor alpha subunit (magenta) from PDB entry 1F6A. At this interface, electron density is also observed for five molecules of the CHAPS detergent (only steroid heads resolved, in stick representation with different colour for each CHAPS molecule).

The filter applied was to remove ligands that share one or more residues in the binding site. In CREDO, residues in the binding site are those that are within 6.5\AA of the ligand. To avoid removal of structures with ligands in remote sites not interacting with each other, only residues that are at 4.5\AA around the ligand are considered as binding site residues.

This filter also removed ligands that interact with metal in catalytic sites. The CREDO database considers these metals as independent ligands, therefore metal interactions with organic ligands are not recorded. An example of these cases is depicted in Figure 4.3, where the ligand brinzolamide binds to the human carbonic anhydrase II through the catalytic Zinc coordinated with 3 histidines (Stams *et al.* 1998).

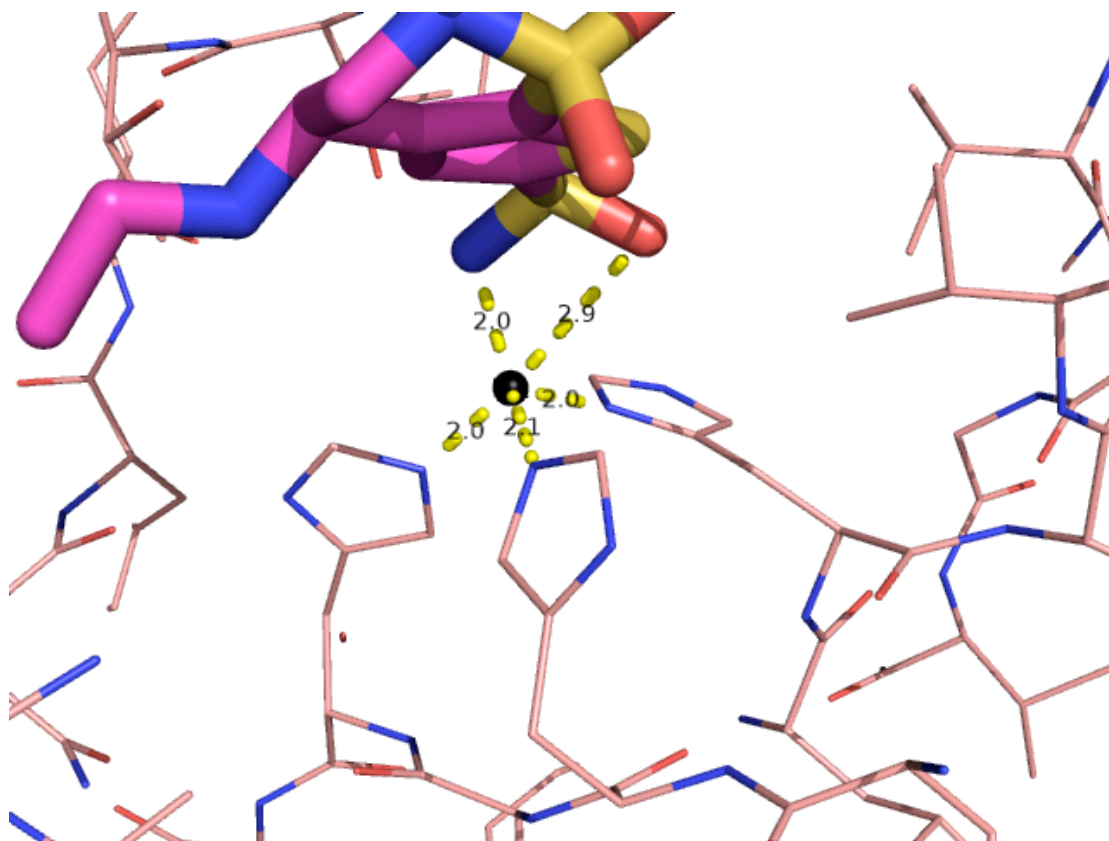


Figure 4.3. Structure 1A42, human carbonic anhydrase II complexed with brinzolamide. Zinc atom is represented by a black sphere, protein atoms by pale pink lines and brinzolamide ligand by magenta sticks.

4.2.1.2 Crystallographic interactions

By definition, PDB shows interactions only within the asymmetric unit, defined by the crystallographer, and not those between them in the crystal lattice, and therefore these possible interactions are not stored in CREDO. Unless one simulates the crystal lattice and recomputes the interactions, there is no trivial filter that can be applied to flag these cases. However, it is relatively easy to avoid structures where the ligands or proteins seem to be

floating in the solvent for this or other reasons. The filter applied to remove these situations involving small molecules (example 1T6J, phenylalanine ammonia-lyase (Calabrese *et al.* 2004), Figure 4.4 left) is to keep only those structures that have at least twice as many contacts as the number of ligand atoms. For structures that have more than one ligand bound to independent sites, the ligand with more contacts is kept (example 1T9U, Acriflavine resistance protein B (Yu *et al.* 2005), Figure 4.4 right).

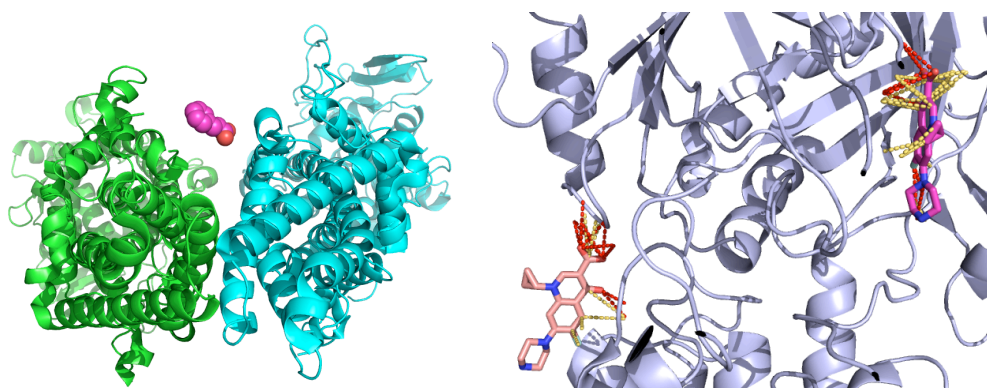


Figure 4.4. Left: Structure 1T6J, phenylalanine ammonia-lyase with carboxycinnamic acid (magenta spheres). Right: Acriflavine resistance protein B with Ciprofloxacin. This molecule (stick representation) binds into two independent sites, the interaction with more atomic contacts is kept for the analysis.

In the case of protein complexes, the asymmetric unit may or may not be the same as the biological assembly (quaternary structure). The protein-protein interfaces studied here are from the predicted quaternary assemblies using the PISA resource from the EBI (Krissinel *et al.* 2007). Moreover, the database stores interactions between pairs of chains of these assemblies. For example, in a trimer with chains A, B and C, PICCOLO stores the contacts between AB, AC and BC. For this reason, in the set of categorised complexes (Obligate and Transient, see 4.2.3.6) only structures that are true dimers are kept. An example of a transient complex not considered in the analysis is structure 1IS8, composed by 15 protein chains that are the complex of GTP-cyclohydrolase I (GTPCHI) with its feedback regulatory protein (GFRP) (Maita *et al.* 2002). The transient interaction is between the GTPCHI decamer and

the two GFRP pentamers; thus the transient interface is between different chains in the PDB.

4.2.1.3 Ligands to remove

Before selecting complexes for the small molecule sets, entries with certain type of ligands were omitted. These included complexes with ligands that:

- Have covalent or metal bonds with proteins
- Are recognised solvents (initial set from (Hartshorn *et al.* 2007) and manually extended by Adrian Schreyer in CREDO) or have less than 10 atoms
- Belong to structures containing nucleic acids
- Are small molecule inhibitors of protein-protein interactions (from TIMBAL)
- Have alternate locations for the ligands or residues in the interface
- Resolution of the crystal structures is lower than 3.5Å

4.2.1.4 Identifiers

For proteins I have used the UniProt identifier (The UniProt 2011), for SCOP domains the SCOP family identifier (Murzin *et al.* 1995) and for small molecules the HET identifier (hetID, from the PDB [<http://www.wwpdb.org/documentation/format23/sect4.html>]). HetID is a three letter code used in the HETATM entries to group heteroatoms in residue-like level. These entries are also known as Het Groups or Chemical Components. Ligands can also be composed of more than one hetID, as in the case of small polymeric peptides. For these ligands, the list of concatenated hetIDs is used as an identifier.

4.2.1.5 Redundancy removal

For protein-protein interfaces, the non-redundant set in PICCOLO (Bickerton 2009; Bickerton *et al.* 2011) was used. Pairwise interfaces were

clustered based on unique UniProt pair identifiers with more than 75% of identical residue interacting pairs. This clustering sampled complexes with the same constituent proteins but different binding modes.

For small molecules I recorded the number of interacting chains and kept entries with distinct ligand names (as hetID or list of hetIDs), UniProt identifiers and numbers of interacting chains. When more than one entry had the same three identifiers, the one with highest quality score was kept. This Qscore was implemented in PICCOLO, for the whole PDB, by Richard Bickerton.

$$Qscore = \left(\frac{1}{resolution} + (0.1 - R_{factor}) \right) \times (1 - PMR)$$

where PMR means proportion of missing residues. Note that this score prioritises X-ray structures. Assessment of the redundancy of small molecules has been done by unique hetID or list of hetIDs, redundancy at protein level by unique UniProt identifier and redundancy of SCOP domains by SCOP family identifier.

4.2.2 Contact definitions

As discussed in chapter 3, software to calculate hydrogen bonds for all types of molecules (proteins, nucleic acids and small molecules), with the same level of specificity, is not available at the moment. For this reason, simple polar and apolar contacts were defined. See 3.3.5 for details. In brief, distance criterion used for all pairs is 4.5Å, depending on the atom type of the pair, they are labelled as follows:

Protein-protein complexes

Apolar contacts: C...C, C...S, S...S (not in Cys-Cys bridges)

Polar contacts: N...O, O...O, N...N, O...S, N...S (S from Cys)

Protein-small molecules complexes

Apolar contacts: C...C, C...S, C...X, S...X (X = Cl, Br, I)

Polar contacts: N...O, O...O, N...N, O...S, N...S, N...F, O...F, S...F (S from Cys)

4.2.3 Subset definitions

4.2.3.1 *Small molecule protein-protein interactions inhibitors*

Small molecules inhibiting protein complexes were identified using TIMBAL. The subset analysed here is composed of the TIMBAL molecules present in CREDO. As this subset is small, it was possible to curate manually the entries in order to have a clean set.

4.2.3.2 *Natural molecules*

Natural small molecules were identified with KEGG (Kanehisa *et al.* 2010), HMDB (Wishart *et al.* 2009), ChEMBL (Gaulton *et al.* 2011), MGEx (pure natural products from AnalytiCon Discovery, <http://www.ac-discovery.com>) databases implemented in CREDO. This set contains molecules that are flagged as substrate, product or cofactor from KEGG and ligands that are labelled as endogenous from the HMDB. Also, natural products from MGEx and molecules classified as such in ChEMBL. For the ChEMBL natural products, the Openeye OEChem toolkit (<http://www.eyesopen.com/oechem-tk>) was used to find ligands in CREDO that were at least 90% similar to them. There was no overlap with the small molecules from the previous set. Filters and redundancy removal (described in section 4.2.1) were used to produce a non-redundant set of small natural molecules interacting with proteins. Further manual classification was performed with these molecules based in their chemical structure and annotated function, so they were labelled as antibiotics, lipids, natural-product-like, nucleotides, peptide-like, steroids and sugars.

4.2.3.3 *Small peptides*

CREDO includes small peptide ligands up to eight residues long, containing both standard and non-standard amino acids. The criterion to belong to this set was that at least half of the chemical components (i.e. a

HET group or residue) are standard amino acids. Small molecules from the previous set were removed to avoid overlap between sets. The same filters and redundancy removal were applied here.

4.2.3.4 Drug-like molecules

Small molecules in the PDB have been extracted from CREDO, applying the same procedure as in chapter 2 (section 2.2.2.1) to select drug-like ligands and to filter out small molecules belonging to the previous sets to avoid overlap. The same filters and redundancy removal was applied here as for previous sets.

4.2.3.5 Approved and oral drugs

Approved drugs characterised in the PDB were retrieved from the implementation of DrugBank (Knox *et al.* 2011) in CREDO. ChEMBL (Gaulton *et al.* 2011) resource was queried to retrieve oral drugs and Scitegic Pipeline Pilot (<http://accelrys.com/products/pipeline-pilot/>) software was used to find the subset of these that are present in CREDO. The same filters and redundancy treatment were applied as before. This set is the only one allowed to overlap with the other ones. In this way, it was possible to identify approved drugs, for instance those that come from natural sources or are peptide like. Further manual classification was performed with these molecules based on their chemical structure and annotated function, so they were labelled as antibiotics, lipids, natural-product-like, nucleotides, peptide-like, steroids, sugars or nota (none of the above, which captures more classical drug-like synthetic molecules).

4.2.3.6 Obligate and transient dimers

These sets were extracted from PICCOLO. Data were taken from two published sets (Zhu *et al.* 2006) and (Mintseris *et al.* 2005). Protein redundancy was removed using UniProt identifiers for the protein pairs,

keeping the structures with the highest Qscore. As discussed previously, only true dimeric entries were kept for these sets. Crystal structures with resolution higher than 3.5Å have also been removed.

4.2.3.7 Quaternary interfaces

As discussed in 4.2.1.5 the non-redundant set of protein interfaces has been extracted from PICCOLO as described in G.R. Bickerton PhD Thesis (Bickerton 2009). In summary, pairwise interfaces with proteins constituted by less than 15 amino acid residues are not considered. Also, interfaces in which the product of the number of interacting residues in each chain is less than 25 are also removed. The remaining pairwise interfaces are clustered together where they have the same pair UniProt identifier and more than 75% of the interface residues are identical. From each cluster the pair with the highest Qscore (see 4.2.1.5) is chosen as representative for that cluster. Pairwise interacting interfaces have been divided into homo and hetero according to whether the proteins in the pair are the same or different respectively. Crystal structures with resolution higher than 3.5Å have also been removed.

4.2.4 Data representation

4.2.4.1 Scissors plots

It has been shown by Olsson et al (Olsson *et al.* 2008) that molecular recognition as a binding event can be studied in terms of polar and apolar interactions due to the aqueous environment where biological interactions occur. The authors show correlation between binding ΔG and burial of apolar surface in the complex formed, due to the more constant contribution of the polar interactions. The authors display this observation in a scatter plot presenting the polar and apolar buried area versus the total buried area upon binding. We call this representation "Scissors plot" (Figure 4.5). I found these graphs useful in detecting different interaction patterns between different

types of molecules. Domination of apolar contacts present a “scissors open” pattern, whereas an increase of polar interactions “closes” the scissors. In addition, the way they are constructed imposes interesting trigonometric proprieties that are used to compare different graphs.

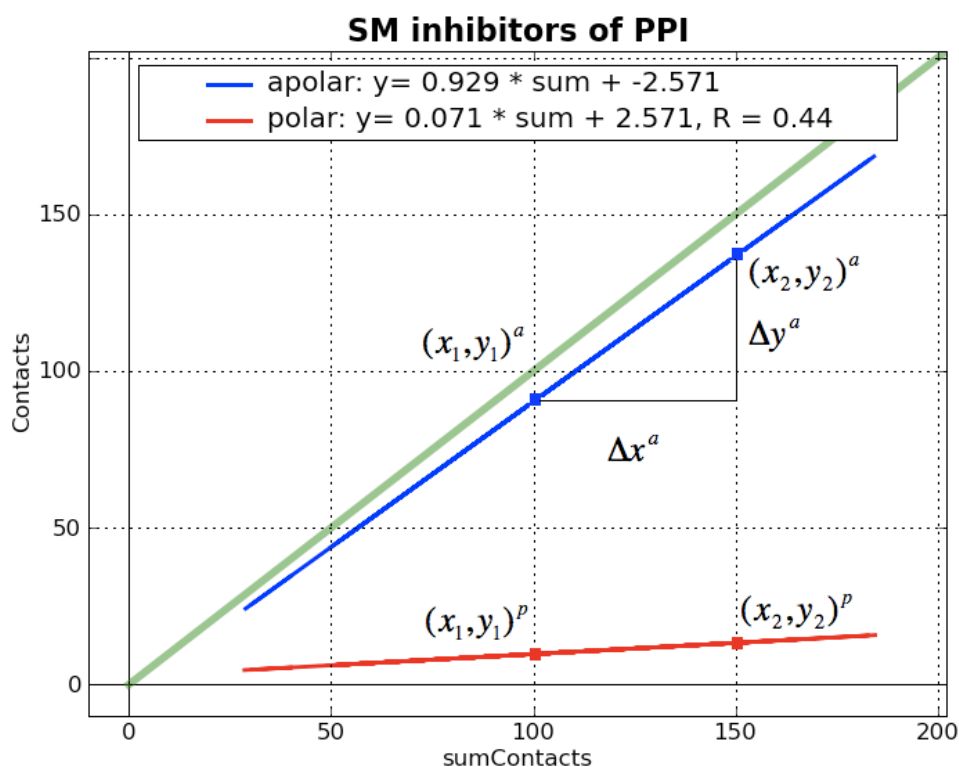


Figure 4.5. Example of a scissor plot. X axis represents sum of contacts (as polar + apolar). Y axis represents the contacts, apolar in blue and polar in red. See text for discussion about these graphs.

Each ligand (or interface) represented in these plots will have two points $(x,y)^a$ for apolar contacts (blue in Figure 4.5) and $(x,y)^p$ for polar contacts (red in Figure 4.5). As the sum of contacts is defined as apolar plus polar, for each ligand the pair of points will have the same $X = X_a = X_p (= Y_a + Y_p)$. This confers certain properties to these graphs. First, all points will be under the line $y = x$ (green diagonal in Figure 4.5). Secondly, if we applied linear regression to each contact type the sum of the slopes will be equal to 1 and the intercepts at the origin will have the same absolute value with opposite signs. This is an important characteristic, as we can compare only one of these regression lines across different sets of molecules. In other

words, in these plots one regression line determines the other. Demonstration of these properties is as follows:

Slopes :

$$m_a = \frac{\Delta y^a}{\Delta x^a} \quad m_p = \frac{\Delta y^p}{\Delta x^p}$$

$$\begin{cases} x_1 = x_1^a = x_1^p \\ x_2 = x_2^a = x_2^p \end{cases} \quad \text{and} \quad \begin{cases} x_1 = y_1^a + y_1^p \\ x_2 = y_2^a + y_2^p \end{cases}$$

Therefore :

$$x_2 - x_1 = y_2^a - y_1^a + y_2^p - y_1^p$$

Sum the slopes :

$$m_a + m_p = \frac{\Delta y^a}{\Delta x^a} + \frac{\Delta y^p}{\Delta x^p} = \frac{y_2^a - y_1^a}{x_2 - x_1} + \frac{y_2^p - y_1^p}{x_2 - x_1} = \frac{y_2^a - y_1^a + y_2^p - y_1^p}{x_2 - x_1} = 1$$

Linear regression lines :

$$\begin{cases} y_a = m_a x + b_a \\ y_p = m_p x + b_p \end{cases} \quad \text{where} \quad y_a = x - y_p \quad \text{and} \quad m_a = 1 - m_p$$

$$\begin{cases} x - y_p = m_a x + b_a \\ y_p = m_p x + b_p \end{cases} \quad \Rightarrow \quad x - m_p x - b_p = m_a x + b_a$$

So,

$$x(1 - m_p) - b_p = (1 - m_p)x + b_a$$

$$-b_p = b_a$$

4.2.4.2 Multiple linear regression

Scissors plots for different sets of molecules can be compared in terms of comparison for only one of the regression lines. I choose to compare the apolar contacts versus the sum of contacts across sets due to the consistent superior r value in all the sets analysed. I follow the method described by Townend (Townend 2002) and the OLS (Ordinary Least Squares) module in Python. However the residuals of these regression lines present heteroscedasticity, i.e. the residuals versus the independent variable are not homogeneously distributed (Figure 4.6). In the case of the scissors plots, the residuals are fan shaped and so errors increase with the independent variable. When heteroscedasticity is pronounced, the chances of Type I error (rejecting true null hypothesis) increases (Osborne *et al.* 2002). For this reason I have

backed up these comparisons with histograms distributing the ratio of polar versus sum of contacts (polar + apolar).

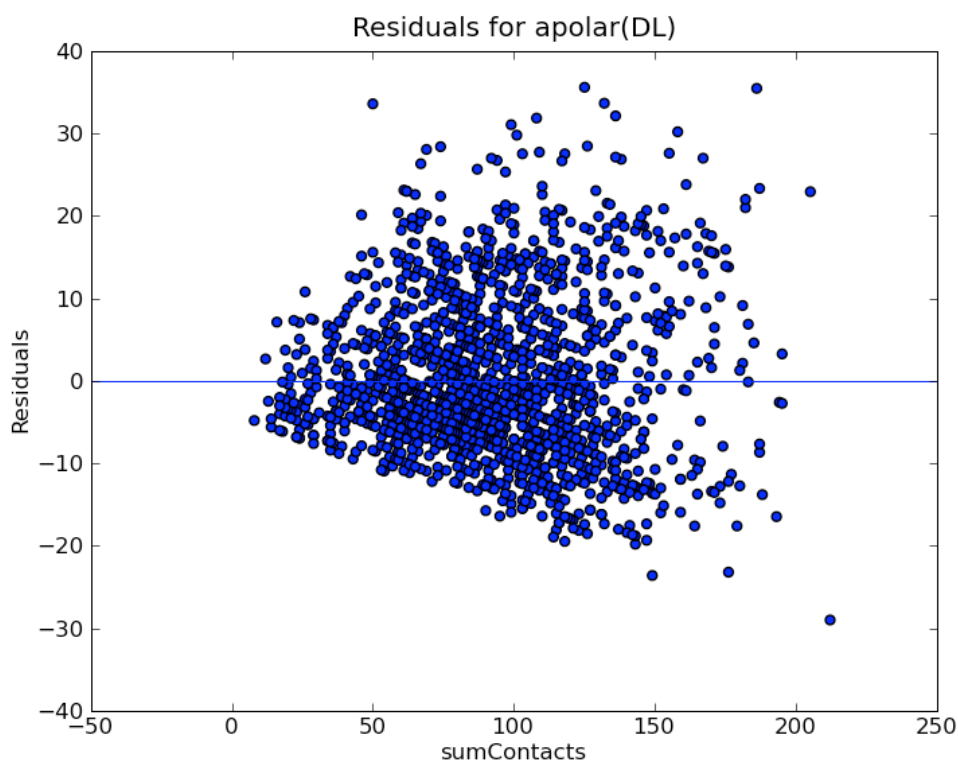


Figure 4.6. Heteroscedasticity. Fan shape of the residuals for the apolar regression line of the drug-like (DL) set.

4.2.4.3 Distribution polar versus sum of contacts

Another way to compare the interaction characteristics in the different sets of molecules is to compare the distribution (with normalised smoothed histogram charts) of the ratio of polar contacts with the apolar contacts. This ratio is described by the number of polar contacts divided by the total number of contacts, as suggested by Dr Will Pitt. In this way, the ratio gives normalised proportion of the polar and apolar contacts. For example a ratio of 0.4 means 40% of the contacts are polar and 60% apolar. As some of these histograms are not normal distributions, non-parametric tests are used for comparison. The Kolmogorov-Smirnov test was used for this purpose and the Kruskal-Wallis test for comparison of medians. I used the stats Python module (Jones *et al.* 2001 -).

4.2.4.4 Bar charts of the polar/sumContacts ratio binned by sum of contacts

These graphs show the mean and standard errors of the polar versus sum of contacts ratio binned by sum of contacts. For clarity, only contacts up to 300 are shown, as that is the maximum number of contacts for small molecules. Sum of contacts is polar + apolar contacts.

4.2.4.5 Contour plots

Dr Will Pitt has developed these charts. They show a scatter plot of ratio of polar contacts by sum of contacts versus the sum of contacts in the X axis; superimposed with the population of each grid square (10 x 0.1) in the scatter plots. This population is represented by a heat map (red to blue), with contour lines in a similar fashion to the contour lines of a topographic map showing elevation.

4.2.4.6 Molecular properties

The number of rotatable bonds and the number of heteroatoms for the small molecule subsets were calculated with Scitegic Pipeline Pilot (<http://accelrys.com/products/pipeline-pilot/>).

4.2.4.7 Bar charts of matched and unmatched atoms

For each subset, the polar and apolar atoms within 4.5Å of the interacting parts, i.e. small molecule-protein or protein-protein, are considered. Then, I record how many of these atoms are engaged in polar and apolar interactions, respectively. These graphs represent first, the mean and standard error of the percentage of matched atoms (within 4.5Å radius, polar atoms doing polar interactions in red and apolar atoms doing apolar interactions in blue) for both binding partners. Secondly, they represent the mean and standard error of the percentage of unmatched atoms (within 4.5Å

radius, polar atoms not making polar contacts in red and apolar atoms not making apolar contacts in blue) for both binding partners.

4.2.4.8 Buried surface area calculation

For the small molecule sets, PyMOL (<http://www.pymol.org/>) has been used to calculate the surface area of the unbound protein, unbound ligand and the complex protein-ligand. In this way, the buried surface area was calculated as follows (sa stands for surface area, psa polar surface area and asa apolar surface area):

$$\text{Buried_sa} = (\text{protein_sa} + \text{ligand_sa} - \text{complex_sa}) / 2$$

$$\text{Buried_psa} = (\text{protein_psa} + \text{ligand_psa} - \text{complex_psa}) / 2$$

$$\text{Buried_asa} = (\text{protein_asa} + \text{ligand_asa} - \text{complex_asa}) / 2$$

Polar surface area for the ligand is calculated from atoms N,O and F. In the protein side, polar surface area is calculated considering N and O atoms and S from cysteine not involved in disulphide bonds. Apolar surface area for the ligand is from atoms C, S, Cl, Br and I. In the protein side, apolar surface area is from C and S from methionine.

4.2.5 Hann's complexity model

This refers to a theoretical model than Mike Hann and co-workers developed at GlaxoSmithKline in 2001 (Hann *et al.* 2001). I summarise it here for its relevance to the results of this chapter.

Although High Throughput Screening (HTS) was widely popular in the 90's, the in-house collections of pharmaceutical companies have often proved insufficient and too costly for the identification of initial hits for lead-development programs. In a seminal paper Hann and colleagues (Hann *et al.* 2001) showed how molecular complexity works against the chances of finding a hit in a biological assay. This analysis is widely considered to provide the

theoretical background that identifies the limitations of HTS and supports a fragment-based approach.

The authors elaborated a simple model to describe the binding event as a match of all ligand features with the features of the surface of the receptor. Ligand and receptor features are conceptually reduced to +/- localized recognition points where a + ligand needs to match a - in the receptor. In this way one can calculate the probability of a binding event for a randomly chosen ligand of a particular size (accounted as number of features) within a given active site described by a finite number of points or features. This probability is computed by enumeration of all possible configurations of ligand and active site and considering the binding event as a complete match of all the ligand recognition points with those of the receptor. For a given active site, this probability can be plotted against the number of ligand features (as a measure of complexity). Although this model is simple and takes into account neither the molecular flexibility that may lead to structural reorganisation upon binding, nor the uneven distribution of binding energy at the receptor interface, it clearly shows how the chances of finding a matching molecule decrease as the complexity of the molecule increases.

Having very simple molecules reduces the likelihood of actually achieving measurable binding events. In addition, low complexity ligands can have multiple binding modes. Although a small number of features are easier to match in the active site, they might not give sufficient affinity for binding to be experimentally detected in a biological assay. Therefore, there is an "optimal" complexity that balances the chances of having a perfect fit between ligand and receptor, and enough interactions to reach a detectable binding.

However, complexity is a relative concept rather than a calculable property. Although one can estimate complexity in many different ways (e.g. molecular weight, fingerprints), the precise values that would optimise the

chances of having a binding ligand and the ability of measuring it, would depend on the system being studied and the assays employed.

4.3 Results

4.3.1 Data sets

I extracted from CREDO database non-redundant sets of protein-ligand complexes classified by the type of small molecule involved: drug-like, approved non-oral and oral drugs, protein-protein interactions inhibitors, natural small molecules and small peptides. For each group, bias was assessed in terms of distinct proteins (by unique UniProt), distinct fold (by unique SCOP family) and distinct small molecules (by unique hetID). These more restricted sets (unique by UniProt, SCOP families or small molecules) presented the same trends as the non-redundant groups. Therefore the statistical analysis has been carried out with the bigger non-redundant-by-complex set of interactions. Affinity data from the PDBind (Wang *et al.* 2004) implementation in CREDO was included in these sets when available. From PICCOLO database I extracted non-redundant sets of protein complexes as obligate dimers, transient dimers, homo and hetero pairwise interfaces from quaternary assemblies, as shown in Table 4.1, which summarises the number of entries of each set. PDB codes for each subset are available to download at:

<http://www-cryst.bioc.cam.ac.uk/members/alicia>

Set	Unique by Complex	Unique UniProt	Unique SCOP families
Drug-like	1,525 (1,206)	518 (385)	165 (143)
Approved drugs	201 (95)	155 (76)	67 (46)
Oral drugs	134 (68)	93 (49)	24 (19)
Protein-protein interaction inhibitors	30 (25)	9 (9)	7 (7)
Natural molecules	1505 (283)	1159 (216)	346 (134)
Small peptides	557 (467)	288 (238)	98 (83)
Obligate dimers	161	161	293
Transient dimers	154	154	183
Homo quaternary interfaces	12,034	7,177	2,711
Hetero quaternary interfaces	2,271	1,709	897
Protein-protein complexes SM inhibited	15	15	13

Table 4.1. Number of entries in each set of molecules. The non-redundant sets are considering non-redundant set of interactions for the complexes (protein-ligand or protein-protein interaction). From these sets I removed protein redundancy by selecting unique UniProt identifiers and removed structural domains redundancy by selecting unique SCOP families. Numbers in parenthesis are the number of unique small molecules in each set. Numbers for unique UniProt and SCOP families for protein complexes refer to distinct pairs of UniProt identifiers or SCOP family respectively. See Methods for details.

The generation of these molecular subsets, mainly for the small molecules, has been an iterative trial-and-error exercise. As discussed in the general considerations and filtering (4.2.1 section of this chapter), the vast amounts of data available compel researchers to use automated filters and selection protocols, which are not perfect. For example, ligands covalently bound to proteins are removed. However, covalent contact is defined in CREDO when the distance between two atoms is less than or equal to the sum of their covalent radius (defined by the Cambridge Crystallographic Data Centre (CCDC), this is a really accurate measure that sometimes outperforms the data in the PDB). Example: PDB entry 1FCN (Patera *et al.* 2000), the ligand Loracarbef is covalently acylated to serine 61 in chain A, the distance reported between the carbon of the ligand and the serine oxygen is 1.46Å whereas the sum of covalent radii (CCDC) for O-C is 1.36Å. On the other hand, data in some cases show a certain degree of ambiguity, for example the definition of “natural product” is somewhat variable within the community. By the same argument, “drug-like molecule” classification is not unequivocal; it is more a continuous “likeness” property without rigorous thresholds. Furthermore, the emerging new targets have forced debate about what it takes to be a drug (Macarron *et al.* 2011). In other cases, the annotation seems to be accurate and straightforward but misinterpretation occurs nevertheless. For example, the case of the “citrate anion”, a common buffer to maintain neutral pH in experimental conditions and therefore a common ligand in the PDB. Due to its size (13 atoms), this ligand can be easily labelled as an oral drug, as lithium citrate (or carbonate) is commonly used to treat depression. However, the active ingredient is the lithium, not the counter anion.

4.3.1.1 Small molecule protein-protein interaction inhibitors

Small molecule inhibitors of protein-protein complexes were identified using TIMBAL. Visual inspection of the 39 PDB entries stored in TIMBAL yielded 28 non-redundant protein-small molecule complexes. Entries with

non-biological contacts were removed, example 1PW6 (Thanos *et al.* 2003), Figure 4.7. Figure 4.8 shows examples of chemical structures from this set.

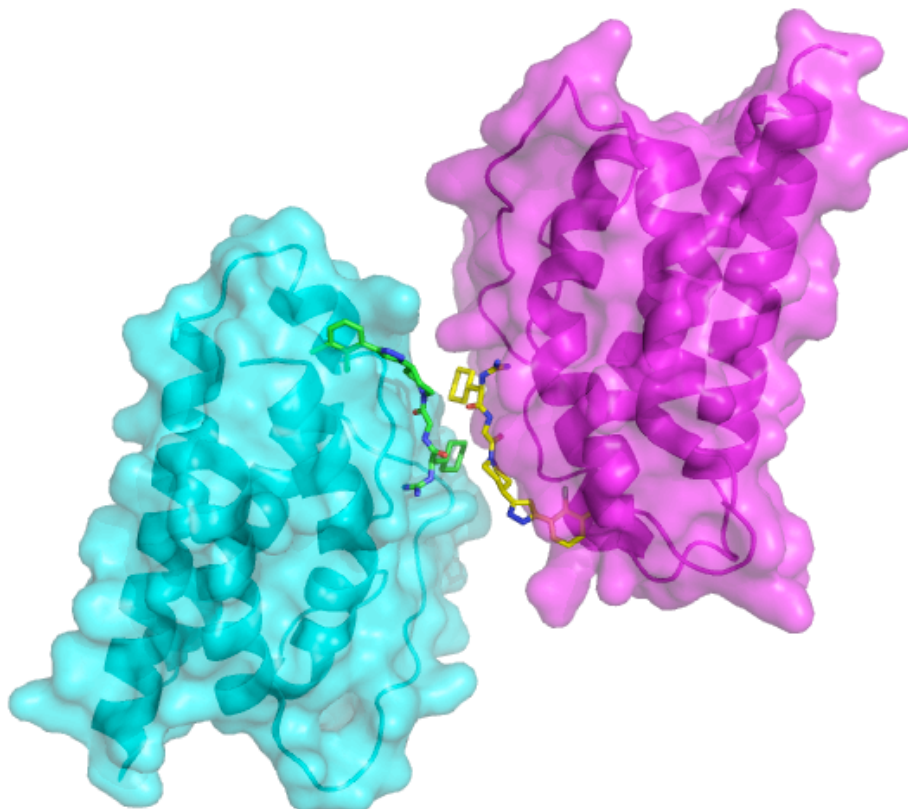


Figure 4.7. PDB 1PW6, crystal structure of IL-2 bound to inhibitor SP2456. This entry was not considered for the non-redundant subset of inhibitors of protein-protein interactions because the small molecule (in stick representation, green and yellow) interacts with itself in the crystal packing. Note these are identical molecules packed in the asymmetric unit.

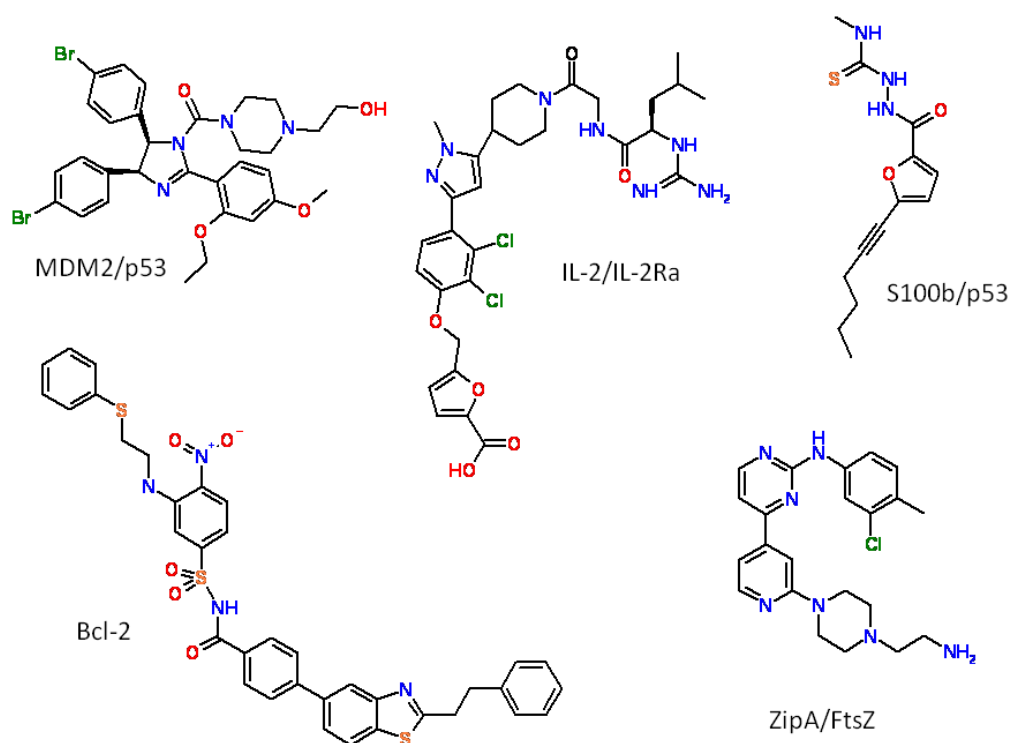


Figure 4.8. Examples of chemical structures of the small molecules inhibiting protein-protein complexes. Each structure is labelled with the protein complex it inhibits.

4.3.1.2 Natural molecules

Natural small molecules in this set are:

- Ligands flagged as substrate, product or cofactor from KEGG
- Ligands labelled as endogenous from the HMDB
- Natural products from MGEx
- Ligands that are a least 90% similar to small molecules classified in ChEMBL as natural products

Filters and redundancy removal yielded 1,505 non-redundant complexes between natural small molecules and proteins, from which there were only 283 distinct small molecules. Figure 4.9 shows that half of this non-redundant subset of interactions was composed of eight nucleotide small molecules: ADP, NAD, NAP, ATP, AMP, FAD, SAH and COA. This redundancy and the chemical composition are taken into account in the discussion. For

example, all of these eight molecules have sugar rings and all but SAH (S-adenosyl-l-homocysteine) have phosphates, therefore these molecules have a high content in heteroatoms.

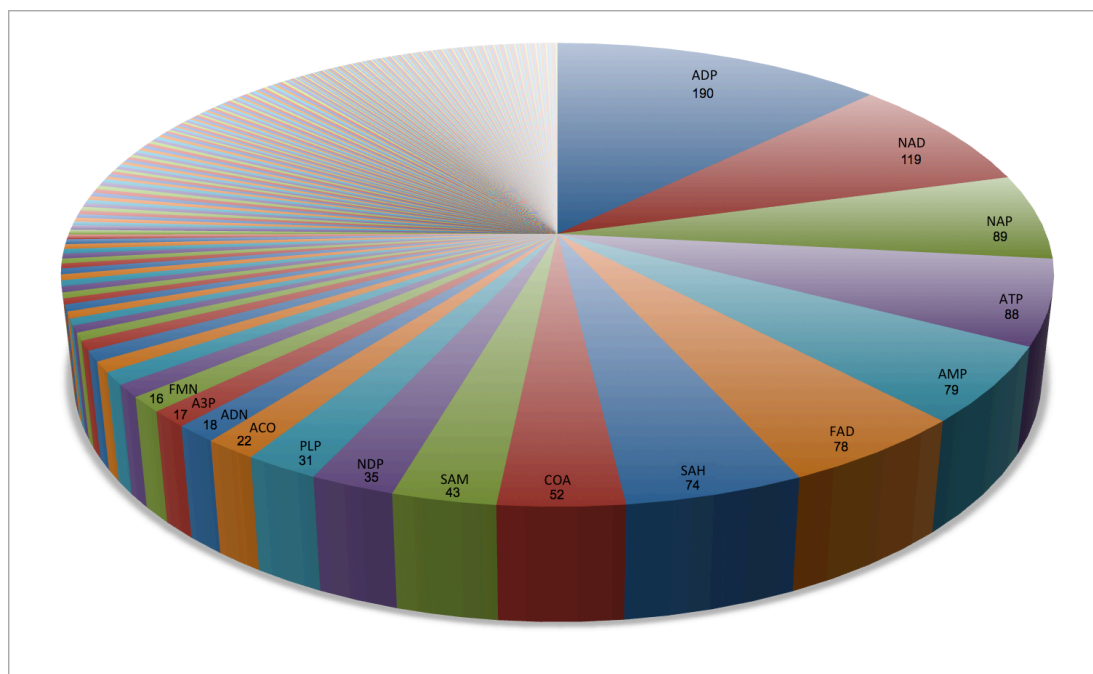


Figure 4.9. Distribution of the natural small molecule subset in terms of entries per chemical structure of the small molecule bound to protein. Only higher frequency entries are labelled for clarity. Note more than half of the subset is composed of the complexes with eight different molecules: ADP, NAD, NAP, ATP, AMP, FAD, SAH and COA.

This diverse set of molecules was classified as antibiotics (13 chemical structures), lipids (13 chemical structures), natural- product-like (72 chemical structures), nucleotides (104 chemical structures), peptide-like (16 chemical structures), steroids (37 chemical structures) and sugars (28 chemical structures). Figure 4.10 shows an example of chemical structures from each category for this set.

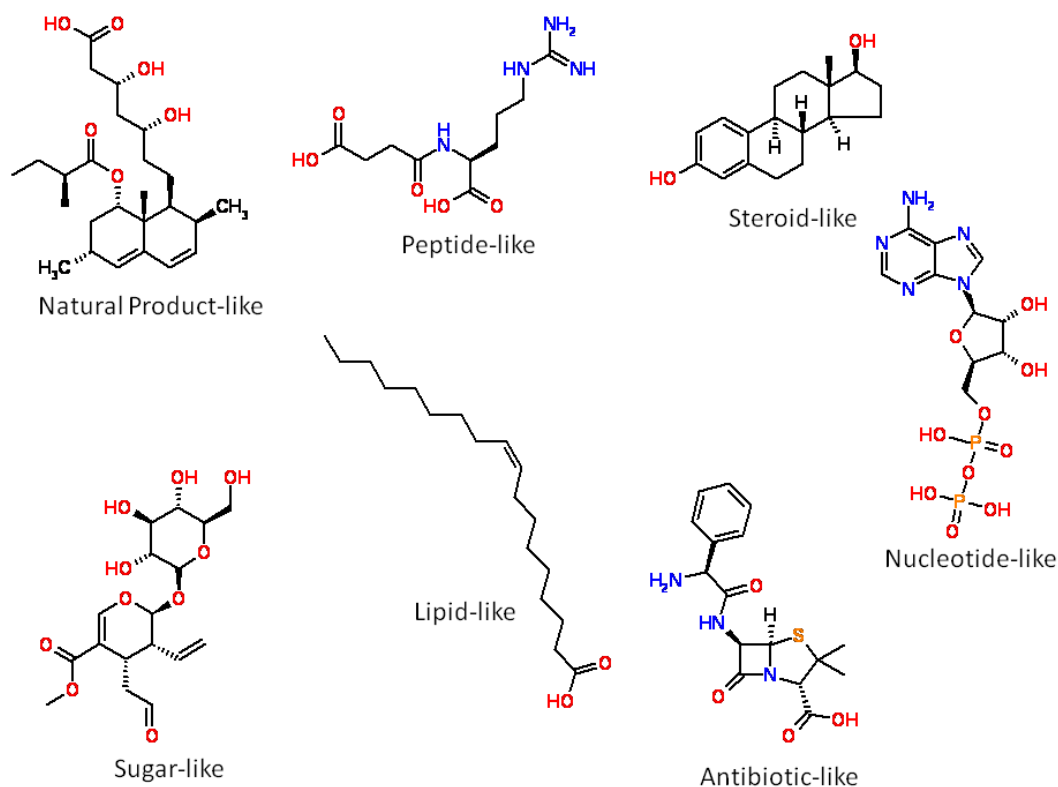


Figure 4.10. Examples of chemical structures in the natural molecules set. Labels correspond to the manual classification based in their structures and functions, so these molecules are categorised into natural product like, peptide like, steroid like, sugar like, lipid like, antibiotic like and nucleotide like.

4.3.1.3 Small peptides

This subset includes short peptides of up to eight residues. These residues can be standard and non-standard amino acids as well as any other residue type, as long as at least half of them are standard amino acids. Examples of molecules in this subset can be seen in Table 4.2.

PDB	Chain id	Residue list	Std_aa/ length
2IFR	B	ACE-PHE-LYS-PHE-TA2-ALA-LEU-ARG	6/8
1BZH	I	ASP-ALA-ASP-GLU-FLT-LEU-AEA	5/7 cyclic
2FNX	P	VAL-ILE-ALA-LYS	4/4
1CE1	P	GLY-THR-SER-SER-PRO-SER-ALA-ASP	8/8

Table 4.2. Examples of ligands in the small peptide set. Last column refers to the ratio of number of standard amino acids by the total residue length of the ligand.

4.3.1.4 Drug-like molecules

As discussed before, drug-likeness is not a precise definition. In order to avoid overlapping, molecules of this set have been selected from the PDB after extracting the small molecules from the previous sets. Therefore, this drug-like set comprises mainly synthetic man-made molecules. The molecular property thresholds applied here are somewhat loose, for instance the molecular weight cut-off is 900. The reason for these broad filters is to be able to compare like to like with the small molecules inhibiting protein-protein complexes (molecular weight range: 150-815Da). Molecules in this set have passed the following filters:

- Not in the ligands to remove set (section 4.2.1.3)
- Not in the small molecule inhibitors of protein-protein complexes, natural molecules or small peptides sets
- At least one carbon atom and one ring, composed only by carbon, nitrogen, oxygen, sulphur, halogen and chains no longer than six sp³-CH₂
- Not similar to nucleotide analogues or detergents

Figure 4.11 shows the chemical diversity of this set and the fine line between definitions, for example ligand MVB could be selected as natural molecule and ligand 8PP as lipid-like one.

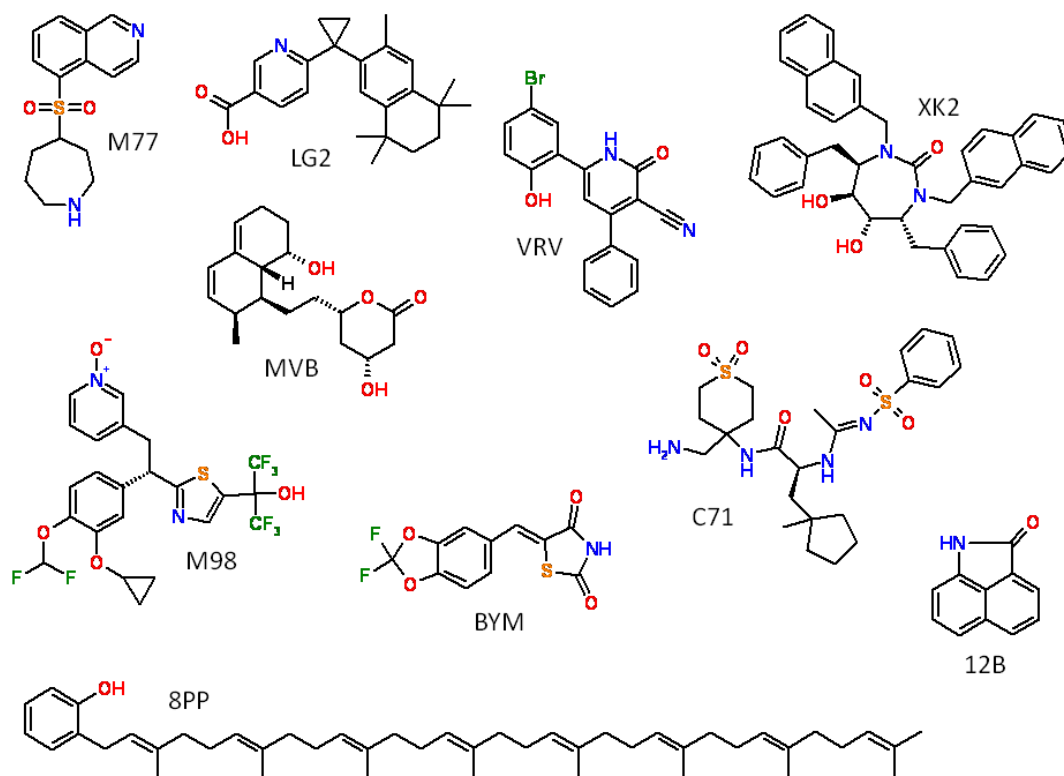


Figure 4.11. Examples of chemical structures in the drug-like subset. Molecules are labelled with their hetID (residue) identifier from the PDB. Ligand 8PP is depicted here as an extreme example of the result of the broad filters applied to select these molecules.

4.3.1.5 Approved and oral drugs

Molecules from this set were selected from the classification in DrugBank as approved drugs and from the classification in ChEMBL as oral drugs. However, molecules in the approved set can have any administration route, including oral. The same categorization applied to the natural molecule set was also done here. In this way, the drug set can be subdivided into antibiotics (25 chemical structures), lipids (two chemical structures), natural-product-like (29 chemical structures), nucleotides (six chemical structures), peptide-like (10 chemical structures), steroids (14 chemical structures), sugars (six chemical structures) and nota, none of the above (65 chemical structures). Figure 4.12 shows an example of chemical structures from each category for this set. It is worth noticing that the complexes studied are not necessarily the ligand drug with its intended target. For example, in PDB 2BXF

(Ghuman *et al.* 2005) Diazepam (Valium, positive allosteric modulator of GABA_A receptor) is bound to human serum albumin. Furthermore, the “approved drug” label also comprises molecules like Thiamin (vitamin B1, example of natural-product-like in Figure 4.12) or Ascorbic acid (vitamin C, example of sugar-like in Figure 4.12). There are also cases of molecules that were marketed but were later withdrawn, for example Bextra (Valdecoxib, example of nota in Figure 4.12). All these data are not easily accessible, either stored in a standardised manner, however molecules in this set were kept as models of small molecules that successfully made their way into the body with a therapeutic effect.

In terms of the size, it is worth remembering that molecules with less than 10 atoms have been removed from all sets. However, there are approved drugs that small. For example, guanidine with four atoms is an approved oral treatment of myasthenia (DrugBank ID DB00536), or dimethyl sulfoxide, also with four atoms, is a common solvent but also an approved topical analgesic (DrugBank ID DB01093). Nevertheless, the filter of a minimum of 10 atoms has been maintained even for this set, as such small molecules are more common as additives in the experimental solutions than as biologically relevant entities.

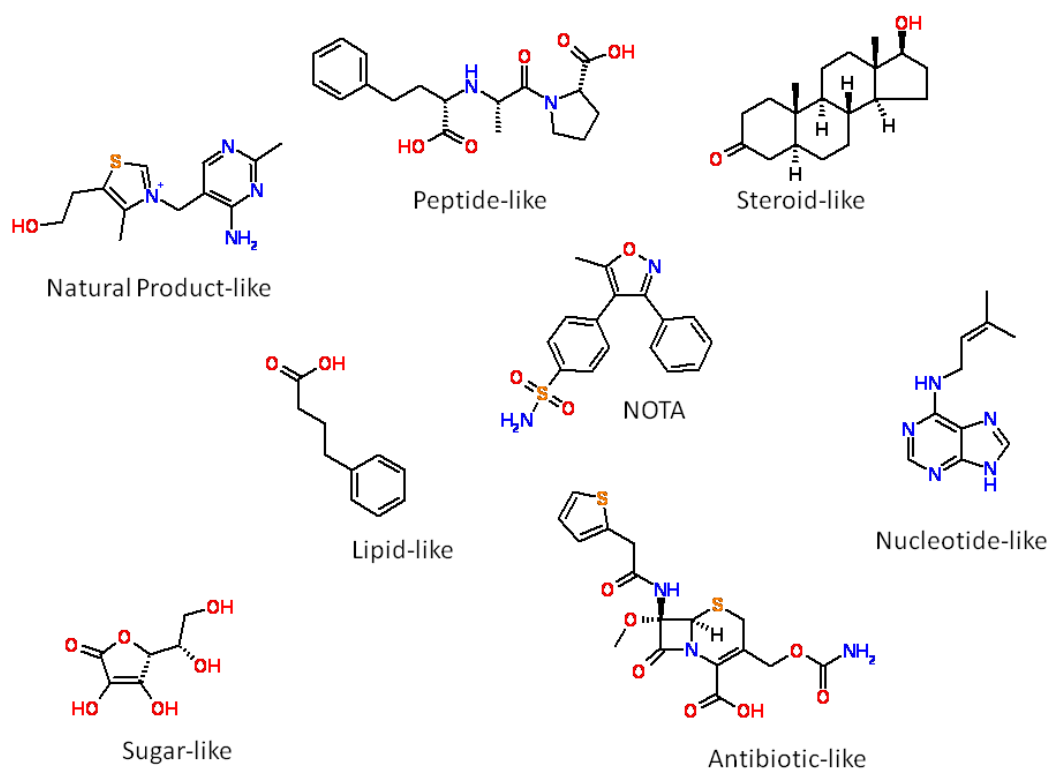


Figure 4.12. Examples of chemical structures in the approved and oral drugs set. Labels correspond to the manual classification based on their structures, so these molecules are categorised into natural product like, peptide like, steroid like, sugar like, lipid like, antibiotic like, nucleotide like and none of the above (NOTA).

4.3.1.6 Protein-protein sets

No further classification has been done in the protein sets. In this study, only protein interfaces are considered regardless of their function, or which constituents form the complex, for example antigen-antibody, enzyme-inhibitor or protein-receptor. The only categorization used refers to the lifetime of the complexes: obligate and transient dimers from the publicly available sets ((Zhu *et al.* 2006) and (Mintseris *et al.* 2005)). These were small sets (315 entries with both dimer classes), but were kept in order to capture any difference in binding pattern, such as transient complexes are more likely to be targeted by a small molecule drug. On the other hand, the general non-redundant set of protein-protein interfaces was considered from PICCOLO, from the quaternary structures predicted by PISA (Krissinel *et al.*

2007). These interfaces were further divided into hetero- (different proteins) and homo- (same protein interacting).

4.3.1.7 Resolution dependency

Crystal structures used in this study have a resolution of at least 3.5Å or better. Figure 4.13 shows that there is no dependency of the polar ratio of the atomic contacts with the resolution of the crystal structures. Furthermore, it also shows that the majority of the complexes studied have a resolution around 2Å, as structures with a better quality score, Qscore (section 2.1.5) have been prioritised.

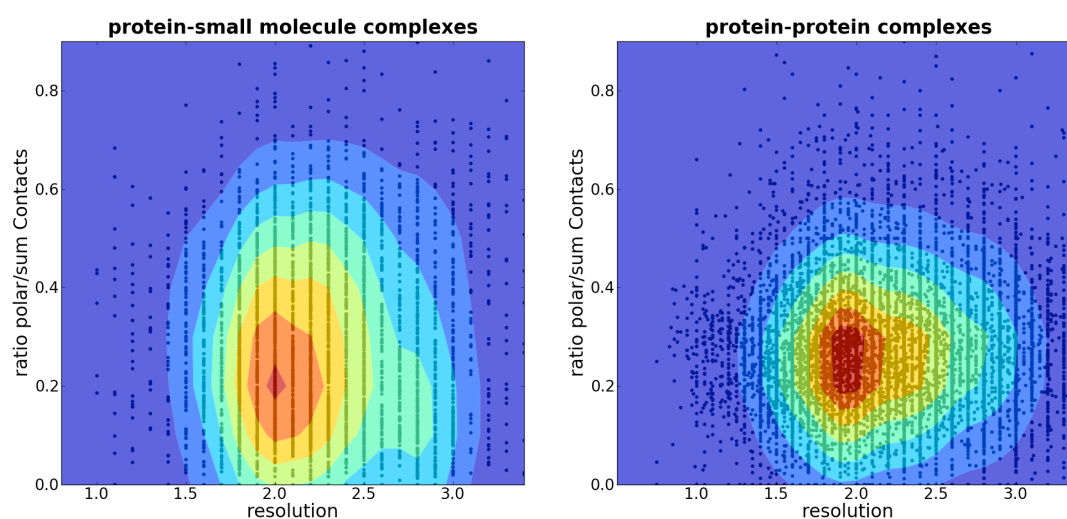


Figure 4.13. Resolution versus ratio of polar contacts as (polar/[polar + apolar]) for the protein-small molecule complexes (left) and for the protein-protein complexes (right). Contour levels show the density of points in the graphs, where red denotes high density and pale blue low density.

4.3.2 Polarity of the interactions

Following Olsson and co-workers (Olsson *et al.* 2008), who studied the binding between small molecules and proteins from the Scorpio database (ITC data) in terms of polar and apolar proportion of buried surface area upon binding, I have based comparisons between different sets of molecules on the extent of polar and apolar atomic contacts that they make. See 4.2.2

for the definition of these contacts. As seen in chapter 3, this discrete count of atomic interactions resembles the measurement of buried surface area used in other studies, providing a coarse description of the interfaces. Figure 4.14 shows the linear correlation of the buried surface area and the number of contacts for all the small molecules used in the analysis. Table 4.3 shows the r and P value for each subset and contact type. In all cases there was significant linear correlation between the surface area buried upon binding and the atomic contacts the small molecule made with the protein. For all cases, r value was 0.8 which shows a medium-strong correlation between the data (Townend 2002).

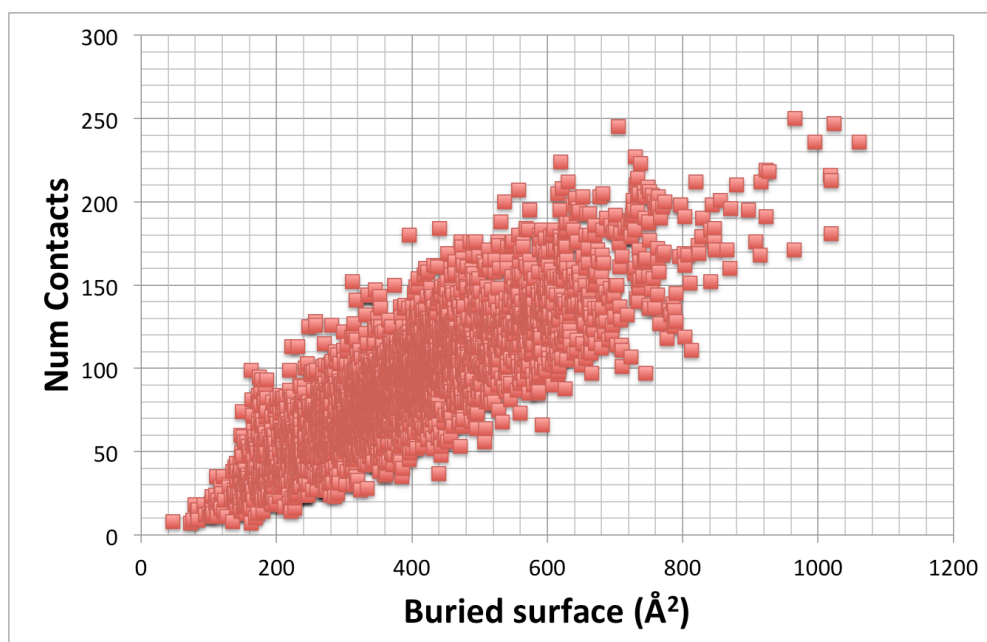


Figure 4.14. Scatter plot of buried surface area upon binding and the number of atomic contacts (polar and apolar) the small molecules made. Points are from all small molecule sets: drug-like, approved drugs, oral drugs, protein-protein interaction inhibitors, natural molecules and small peptides.

Subset	Contact type	r value	P value
Drug-like	all	0.82	0.00
	polar	0.85	0.00
	apolar	0.82	0.00
Approved drugs	all	0.79	3.86E-45
	polar	0.83	4.36E-53
	apolar	0.76	4.11E-39
Oral drugs	all	0.89	1.04E-67
	polar	0.85	1.37E-56
	apolar	0.90	1.54E-73
Protein-protein interaction inhibitors	all	0.84	2.03E-08
	polar	0.89	1.52E-10
	apolar	0.81	1.34E-07
Natural molecules	all	0.85	0.00
	polar	0.91	0.00
	apolar	0.81	0.00
Small peptides	all	0.79	0.00
	polar	0.84	0.00
	apolar	0.73	8.77E-93
All (Figure 4.14)	all	0.82	0.00
	polar	0.90	0.00
	apolar	0.82	0.00

Table 4.3. r and P values from linear correlation calculations between buried surface upon binding and number of atomic contacts small molecules make with proteins. P value has been rounded to zero when $P < 1E-100$.

4.3.3 More polar interactions in natural subsets

For each set of molecules, plotting the sum of contacts versus either the polar or apolar contacts generates the 'scissors plot' (see Methods for details). In these graphs, the openness of the trend lines gives the ratio of polar versus apolar contacts. Figure 4.15 shows the scissors plots for the

drug-like and PPI inhibitors (scissors open), natural small molecules (scissors closed) and small peptides and protein complexes (scissors half way). For the drug-like molecules, these plots show that molecular interactions are dominated by apolar contacts, whereas the polar contacts remain somewhat constant with the increase of ligand size (which correlates with sum of contacts). Similar conclusions were reached by Olsson and co-workers in their analysis of the SCORPIO database (Olsson *et al.* 2008). It is more pronounced in small molecules inhibiting protein-protein interactions as we have seen in chapter 2. On the other hand, natural small molecules, small peptides and protein complexes present a different trend where the polar interactions play a larger role. One aspect of this may be that evolutionary processes have produced a better fit than achieved by medicinal chemists. But more importantly, endogenous molecules have not been constrained to be absorbed or transported in the circulation or across membranes into cells of other living organisms. On the other hand, it is now recognised that medicinal chemists have tended to increase lipophilicity to gain potency (van de Waterbeemd *et al.* 2001; Leeson *et al.* 2007; Hann 2011). Interestingly, the lower part of the graphs, i.e. smaller molecular size (fragments), present a more balanced ratio between polar and apolar contacts. Firstly, this result agrees with Hann complexity model (Hann *et al.* 2001), where it is easier for a smaller molecule to match target features; and secondly, it also supports the strategy of fragment-based drug design where the initial fragments anchor in the site with specific interactions (Congreve *et al.* 2008) and deliver less lipophilic hits (Keserü *et al.* 2009). Natural molecules have a bimodal distribution as shown in Figure 4.15 (D and E) and Figure 4.16 (A); this is due to the presence of steroid-like molecules presenting an apolar profile, while the rest follow a polar trend. To evaluate whether the high proportion of polar contacts is because many of the natural molecules have phosphate groups, Figure 4.15 (E) shows the scissors plot for the subset of natural molecules without phosphorus. The graph has fewer points, but the trends are the same, and the bimodal distribution is maintained.

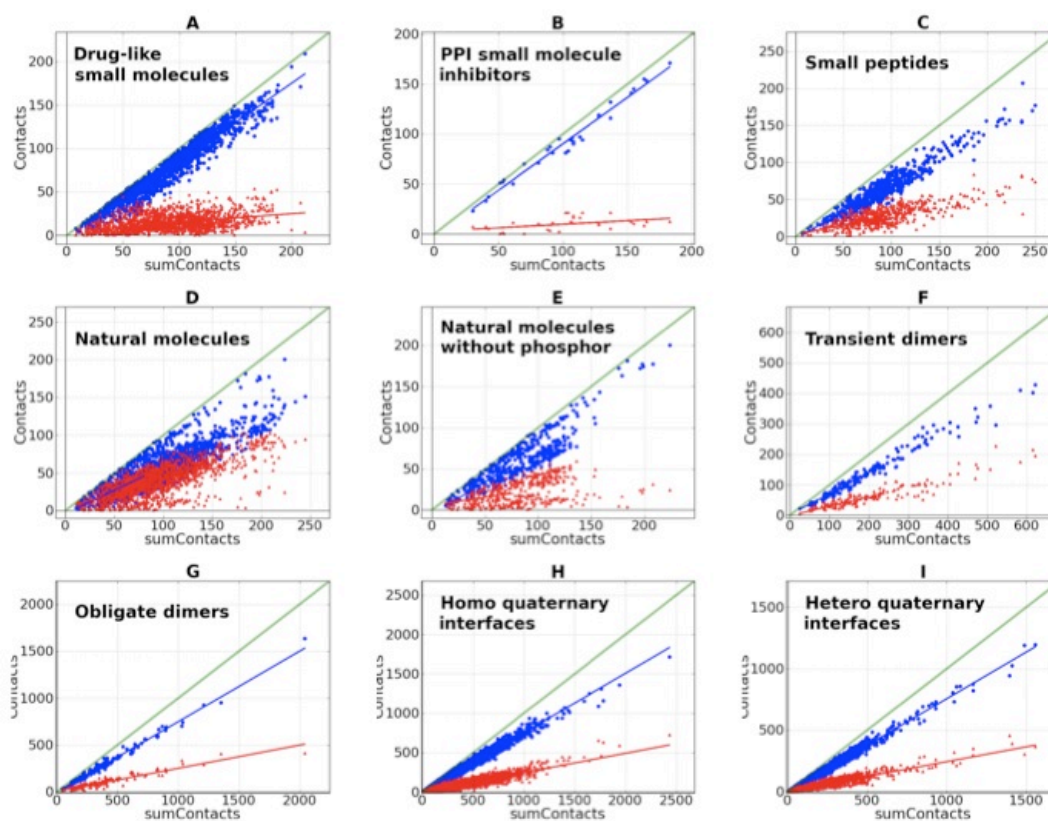


Figure 4.15. Scissors plots for the non-redundant-by-complex (table 1) sets of protein complexes. A: drug-like small molecules bound to proteins. B: Protein-protein interactions small molecule inhibitors bound to proteins. C: Small peptides bound to proteins. D: Natural small molecules bound to proteins. E: Natural small molecules without containing phosphor bound to proteins. F: Transient protein-protein dimers. G: Obligate protein-protein dimers. H: Homo protein-protein interfaces from quaternary structures. I: Hetero protein-protein interfaces from quaternary structures. Polar (red) and apolar (blue) contacts are scattered against sum of contacts. Details of the regression lines for each graph and contact type can be found in Table 4.4.

Subset	Type	Slope	Intercept	R value	P value	angle
Drug-like	polar	0.09	5.87	0.38	4.73E-53	5
Drug-like	apolar	0.91	-5.87	0.97	0.00	42
PPI inh	polar	0.07	2.57	0.44	1.45E-02	4
PPI inh	apolar	0.93	-2.57	0.99	2.93E-24	43
Small pep	polar	0.27	2.61	0.77	0.00	15
Small pep	apolar	0.73	-2.61	0.96	0.00	36
Nat mol	polar	0.37	4.04	0.73	0.00	21
Nat mol	apolar	0.63	-4.04	0.87	0.00	32
Nat mol -P	polar	0.20	5.44	0.46	1.15E-24	11
Nat mol -P	apolar	0.80	-5.44	0.91	0.00	39
Transient	polar	0.31	-1.92	0.92	2.41E-65	17
Transient	apolar	0.69	1.92	0.98	0.00	35
Obligate	polar	0.25	5.04	0.94	1.42E-75	14
Obligate	apolar	0.75	-5.04	0.99	0.00	37
Homo interf	polar	0.24	2.20	0.92	0.00	14
Homo interf	apolar	0.76	-2.20	0.99	0.00	37
Hetero interf	polar	0.24	3.17	0.92	0.00	13
Hetero interf	apolar	0.76	-3.17	0.99	0.00	37

Table 4.4. Linear regression details for each subset and contact type in Figure 4.15. Subsets: Drug-like, PPI inh (protein-protein interactions inhibitors), Nat mol (natural molecules), Nat mol -P (natural molecules that do not contain phosphor), Small pep (small peptides), Obligate (obligate dimers), Transient (transient dimers), Homo interf (homo quaternary interfaces), Het interf (hetero quaternary interfaces). Angle column denotes the angle that the regression line makes with the X axis, it is a translation of the slope into degrees. P value has been rounded to zero when $P < 1E-100$.

In order to define the statistical significance of these plots Multiple Linear Regression (MLR using OLS, Ordinary Least Squares) was used between the apolar regression lines of each set. In addition, the distribution of polar/apolar contact ratio (normalised as $\text{polar}/[\text{polar}+\text{apolar}]$) between sets was analysed with non-parametrical tests, as not all the sets have normal distribution of the contact ratio. This was done to minimise Type I error (concluding there is a significant difference when there is not) due to the heteroscedasticity of the residuals of the regression lines in the scissors plots. See Methods section for details. Table 4.5 summarises the comparisons across all data sets. Figure 4.16, Figure 4.17 and Figure 4.18 show comparison of the distribution of polar/sumContacts ratio for selected subsets.

	App drugs	Oral drugs	PPI Inh	Nat mol	Nat mol -P	Small pep
Drug-like	-0.03	0.00	0.06	-0.29	-0.14	-0.15
App drugs		0.04	0.09	-0.26	-0.11	-0.11
Oral drugs			0.05	-0.30	-0.15	-0.15
PPI inh				-0.35	-0.20	-0.20
Nat mol					0.15	0.15
Nat mol -P						0.00

	Obligate	Transient	Homo interf	Hetero interf	PPI inh by SM
Drug-like	-0.13	-0.16	-0.12	-0.12	-0.07
App drugs	-0.09	-0.12	-0.08	-0.08	-0.03
Oral drugs	-0.13	-0.16	-0.12	-0.12	-0.07
PPI inh	-0.18	-0.21	-0.17	-0.17	-0.12
Nat mol	0.17	0.13	0.17	0.18	0.22
Nat mol -P	0.02	-0.01	0.03	0.03	0.08
Small pep	0.02	-0.01	0.03	0.03	0.08
Obligate		-0.03	0.01	0.01	0.06
Transient			0.04	0.04	0.09
Homo interf				0.00	0.05
Hetero interf					0.05

Table 4.5. Differences in medians of the contact ratios (polar/[polar + apolar]) between the different sets of molecules (row - column). Table is divided in two for clarity. Subsets: Drug-like, App drugs (approved drugs including oral), Oral drugs, PPI inh (protein-protein interactions inhibitors), Nat mol (natural molecules), Nat mol -P (natural molecules that do not contain phosphorus), Small pep (small peptides), Obligate (obligate dimers), Transient (transient dimers), Homo interf (homo quaternary interfaces), Het interf (hetero quaternary interfaces), PPI inh by SM (protein-protein interfaces inhibited by small molecules). Values in bold denote significant differences in medians ($P < 0.05$). Note both subsets of PPI SM Inhibitors and PPI SM Inhibited are small (28 and 15 respectively); they are included for the exceptional insight these cases present rather than their statistical significance.

Values in Table 4.5 are the difference in medians of the ratio polar/sumContacts for each subset. Drug-like molecules bound to proteins present on average less polar contacts than the other sets, with the exception of the PPI inhibitors that have more apolar contacts. Approved and oral drugs analysed here present the same interaction profile as drug-like molecules. The group with more polar contacts on average is the natural molecules. When molecules containing phosphorus are removed from the natural set, the

average polar contacts decreases 15%, nevertheless this is the set that engages the most polar interactions. Amongst protein oligomers, the quaternary interfaces present the same profile for homo and hetero interfaces, which is similar to the subset of obligate dimers, whereas the transient dimers are slightly (3-4% on average) more polar (in agreement with previous findings (Nooren *et al.* 2003)) and more similar to the subset of small peptides. Interestingly, the small subset of protein-protein complexes inhibited by small molecules shows a trend that is similar to other protein complexes. However, the small molecules inhibiting them present a more apolar profile than the drug-like molecules.

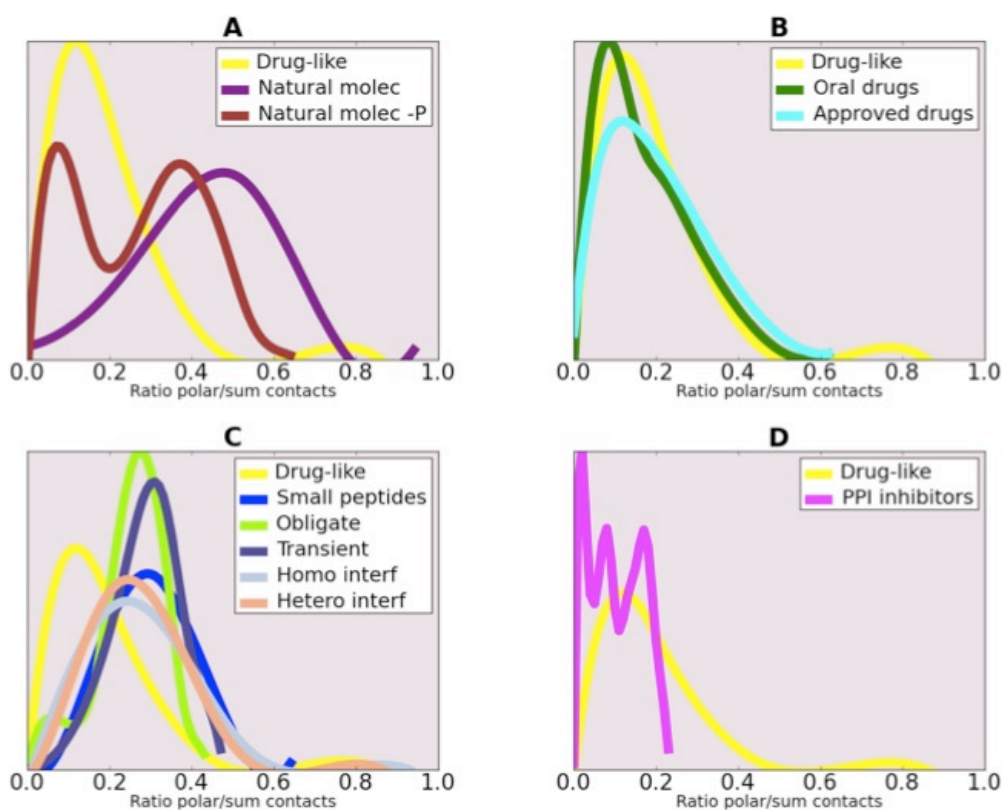


Figure 4.16. Normalised distributions of the ratio of polar contacts (represented by $\text{polar}/[\text{polar}+\text{apolar}]$), each chart compares drug-like against the others. A: drug-like versus natural small molecules with and without phosphor. B: drug-like versus approved and oral drugs. C: drug-like versus small peptides, obligate and transient protein-protein dimers, homo and hetero quaternary protein-protein interfaces. D: drug-like versus PPI inhibitors.

Figure 4.16 is a graphical representation of the data discussed in the previous paragraph. Drug-like molecules and drugs bound to proteins have, with the small molecules inhibiting PPI (in magenta, chart D), the most apolar interaction profile. All the other distributions are shifted to the right (more polar interactions) with respect to these. In chart A, Figure 4.16, the distribution of the natural molecules (in purple) is dominated by nucleotides with high content of phosphates, this distribution has 44% of polar contacts on average (median). When phosphor-containing molecules are removed from this set, the bimodal distribution seen in the scissors plots emerges again; natural molecules can engage few polar contacts (for example steroids) or many (for example heteroatom-rich molecules). In chart B, Figure 4.16 the distributions of the approved and oral drugs are virtually identical of the drug-like. In chart C, Figure 4.16 the small peptide and protein oligomer sets have similar distributions, less polar than natural molecules but more polar than synthetic drug-like molecules.

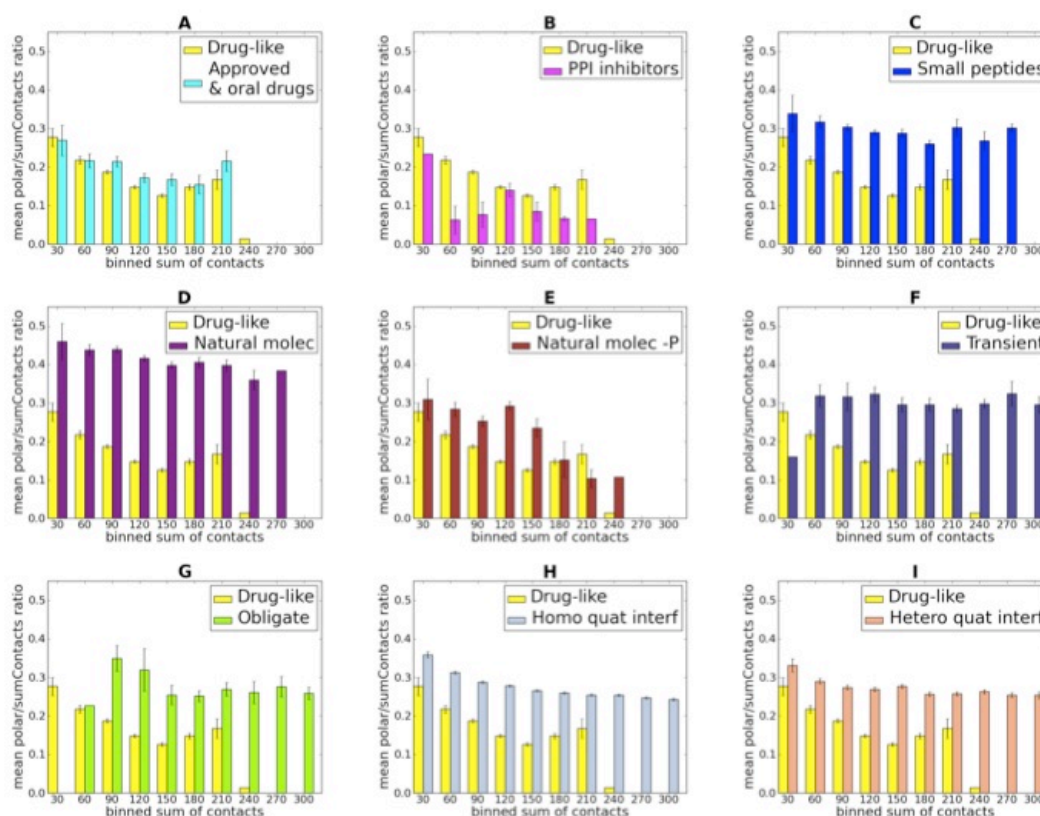


Figure 4.17. Comparisons of polar/sumContacts ratio means, binned by sum of contacts (polar+apolar), each chart compares drug-like against the others. A: drug-like versus approved, each and oral drugs. B: drug-like versus PPI inhibitors. C: drug-like versus small peptides. D: drug-like versus natural molecules. E: drug-like versus natural molecules without phosphor. F: drug-like versus transient protein-protein dimers. G: drug-like versus obligate protein-protein dimers. H: drug-like versus homo quaternary protein-protein interfaces. I: drug-like versus hetero quaternary protein-protein interfaces. Error bars denote the standard error of the mean.

Figure 4.17 looks at the same ratio of polar versus sum of contacts, but it is binned by sum of contacts. With this representation, it becomes clear that the molecules engaging more polar contacts, for example in the drug-like set have fewer contacts overall and they are generally smaller molecules. This effect is more pronounced in the small molecule inhibitors of protein interfaces. A similar situation occurs with natural molecules without phosphorus, the polar proportion of contacts decreases with ligand size. This becomes more evident in Figure 4.18, where the upper right quadrant of the nine charts is empty. In these graphs (Figure 4.18), the proportion of polar contacts (Y axis) decreases with molecular size (X axis as sum of contacts).

This result can be justified in terms of the Hann's complexity model (Hann *et al.* 2001), the chances of matching at the same time different polar interactions decreases with the number of interactions to match. Furthermore, the flexibility required to match many different specific interactions goes against spontaneous binding due to entropic penalty.

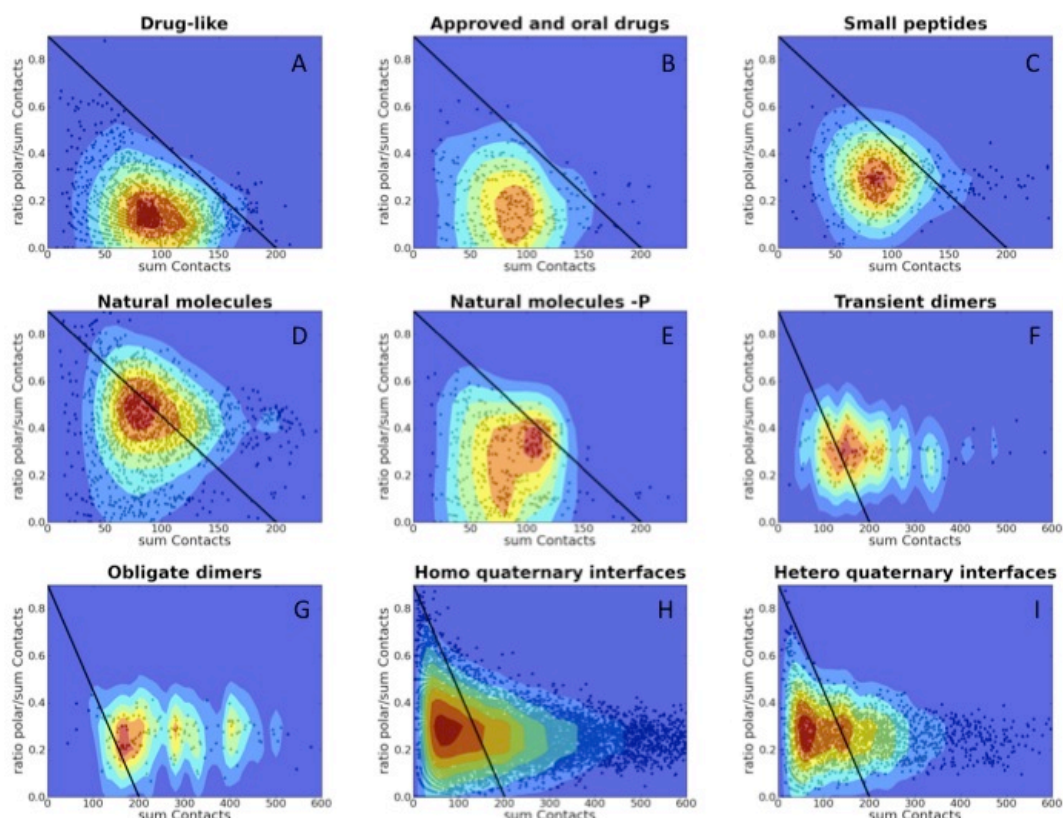


Figure 4.18. Ratio of polar/(polar+apolar) versus sum of contacts (polar+apolar). Contour levels show the density of points in the graphs, where red denotes high density and pale blue low density. The black line in all the graphs goes between 0.9 ratio to 200 sum of contacts to have the same reference to aid comparison between sets. A: drug-like small molecules bound to proteins. B: Approved and oral drugs bound to proteins. C: Small peptides bound to proteins. D: Natural small molecules bound to proteins. E: Natural small molecules without containing phosphor bound to proteins. F: Transient protein-protein dimers. G: Obligate protein-protein dimers. H: Homo protein-protein interfaces from quaternary structures. I: Hetero protein-protein interfaces from quaternary structures. For clarity, graphs for protein-protein complexes are plotted up to 600 contacts only.

The analyses shown in Table 4.5 and Figure 4.15, Figure 4.16, Figure 4.17 and Figure 4.18 demonstrate that natural molecules have a higher proportion of polar contacts than their synthetic counterparts. In order to pin

down these different interaction profiles, the following sections analyse the atomic composition of the ligands, in terms of heteroatom content and rotatable bonds, and the proportion of matched and unmatched atoms at the interfaces. Detailed analysis of the binding sites is discussed in chapter 5.

4.3.4 Atomic composition and molecular flexibility

For small molecules and small peptides, the number of heteroatoms and rotatable bonds are straightforward to calculate and interpret. For proteins, their interpretation is more difficult due to intramolecular hydrogen bonds and atomic occlusion from solvent. Therefore, this section discusses the interaction profile of small molecules in terms of their atomic composition and flexibility.

Natural small molecules and small peptides engage on average more polar contacts with their targets than synthetic molecules. Analysis of the content of heteroatoms (ratio of number of heteroatoms and total number of atoms) and rotatable bonds (ratio of number of rotatable bonds and total number of atoms) shows that the more polar interaction profile presented by natural molecules is due to a higher content of heteroatoms (19% more on average than drug-like molecules), arguably placed in the right conformation for interaction with the protein target. Whereas the lower content of heteroatoms in peptides in comparison with natural molecules (9% less on average) is compensated by greater flexibility (20% more on average) to match the more directionally constrained polar interactions. Table 4.6 and Figure 4.19 summarise these comparisons. The apolar interaction profile of synthetic molecules corresponds to rigid ligands with low content in heteroatoms. In contrast, natural molecules are also rigid but rich in heteroatoms, whereas small peptides are flexible with fewer heteroatoms. These observations are for the general trends, however it is worth noting that natural molecules can also be rigid and lipophilic, for example steroids like testosterone (see Figure 4.19, A and C). In fact, natural molecules cover the

whole range of polarity (0 to 0.9 in the polar/sumContacts scale), but the overall behaviour is predominantly polar, especially when compared with synthetic molecules.

Het/N_at	App drugs	Oral drugs	PPI inh	Nat mol	Nat mol -P	Small pep
Drug-like	-0.01	-0.02	0.01	-0.19	0.02	-0.10
App drugs		-0.01	0.02	-0.18	0.03	-0.09
Oral drugs			0.04	-0.16	0.04	-0.08
PPI inh				-0.20	0.00	-0.11
Nat mol					0.20	0.09
Nat mol -P						-0.12

Rot/N_at	App drugs	Oral drugs	PPI inh	Nat mol	Nat mol -P	Small pep
Drug-like	-0.01	-0.01	-0.07	-0.04	0.00	-0.24
App drugs		0.00	-0.07	-0.03	0.01	-0.24
Oral drugs			-0.07	-0.03	0.01	-0.24
PPI inh				0.03	0.08	-0.17
Nat mol					0.04	-0.20
Nat mol -P						-0.25

Table 4.6. Difference in medians of the ratios (number of heteroatoms/number of heavy atoms, upper table) and (number of rotatable bonds/number of heavy atoms, lower table). Differences are between the different set of small molecules (row – column). Subsets: Drug-like, App drugs (approved drugs including oral), Oral drugs, PPI inh (protein-protein interactions inhibitors), Nat mol (natural molecules), Nat mol -P (natural molecules that do not contain phosphor), Small pep (small peptides). Values in bold denote significant differences in medians ($P < 0.05$).

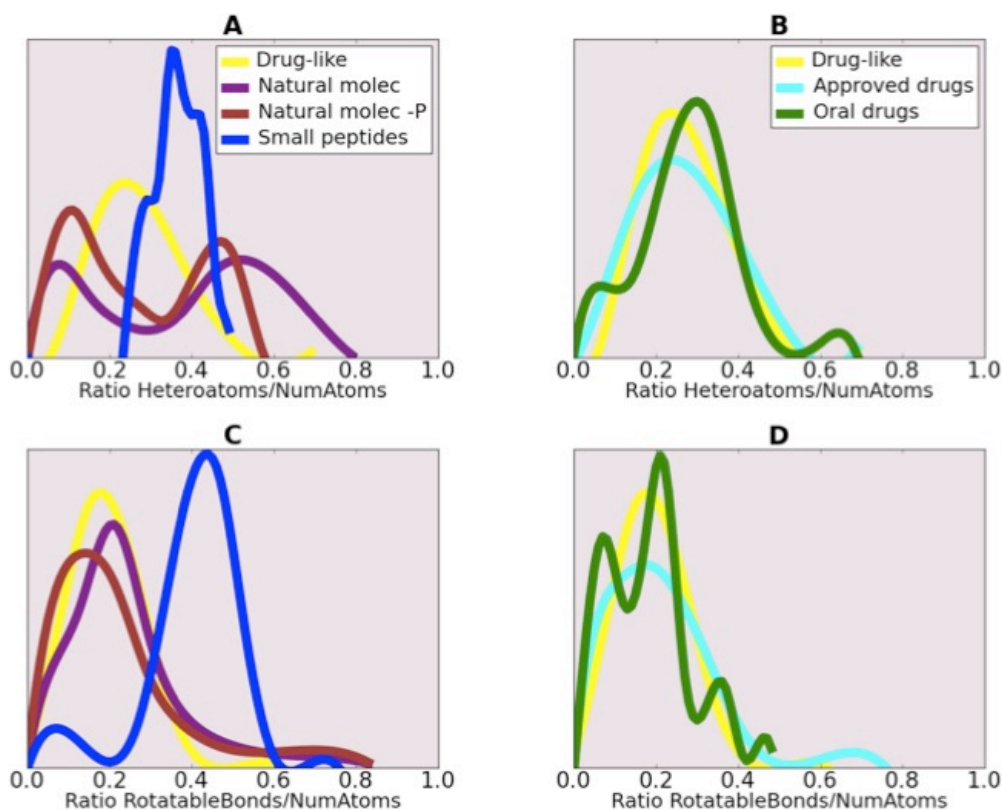


Figure 4.19. A: Distribution of the ratio of number of heteroatoms by number of heavy atoms for drug-like small molecules, natural molecules, natural molecules without phosphor and small peptides. B: Distribution of the ratio number of heteroatoms versus number of heavy atoms for drug-like small molecules, approved and oral drugs. C: Distribution of the ratio of number of rotatable bonds by number of heavy atoms for drug-like small molecules, natural molecules, natural molecules without phosphor and small peptides. D: Distribution of the ratio of number of rotatable bonds by number of heavy atoms for drug-like small molecules, approved and oral drugs.

4.3.5 Matched and unmatched atoms at the binding interfaces

In the previous section we have seen that higher content in heteroatoms for natural molecules leads to a more polar profile. This not the case for small peptides but may be compensated by their greater flexibility that facilitates more specific polar interactions, in particular hydrogen bonds. However, the key point is whether all these heteroatoms are making polar contacts or are unmatched. Figure 4.20 shows the mean of the percentage of buried atoms engaged in successful interactions and the same measure for the unmatched buried atoms. Small ligands, including drug-like up to small

peptides that are found in the left part of the figure, are more contact efficient than the protein to which they are bound, i.e. on average around 90% of the ligand atoms are matched in all sets, with natural molecules without phosphorus being the most efficient. In contrast with the 70-80% of the protein atoms matched, the small molecule atoms are more exposed and able to contact the protein, whereas the atoms in the protein can be less accessible. Furthermore, studies of hundred complexes of nine different ligands (Kahraman *et al.* 2007) found that binding pockets are on average three times bigger than the ligands they encapsulate; therefore in proportion more atoms in the protein will be at the periphery of the ligand (our cut-off here was 4.5Å) without making useful interactions. Another interesting result from this analysis is that synthetic molecules have a larger proportion of unmatched polar atoms (in both ligand and protein side) than the natural ones. In other words, if one wants to increase the polar contacts synthetic molecules make there is still room for improvement. Arguably, oral drugs need to restrain the polar signature to get distributed in the body, but Figure 4.19 shows that approved and oral drugs are not making the most of their polar composition. However, improving enthalpic contacts is not a trivial task, not only for the difficulty of designing geometries that match polar constraints, but also for the enthalpic and entropic penalties upon desolvation and the loss of conformational entropy (Ferenczy *et al.* 2010).

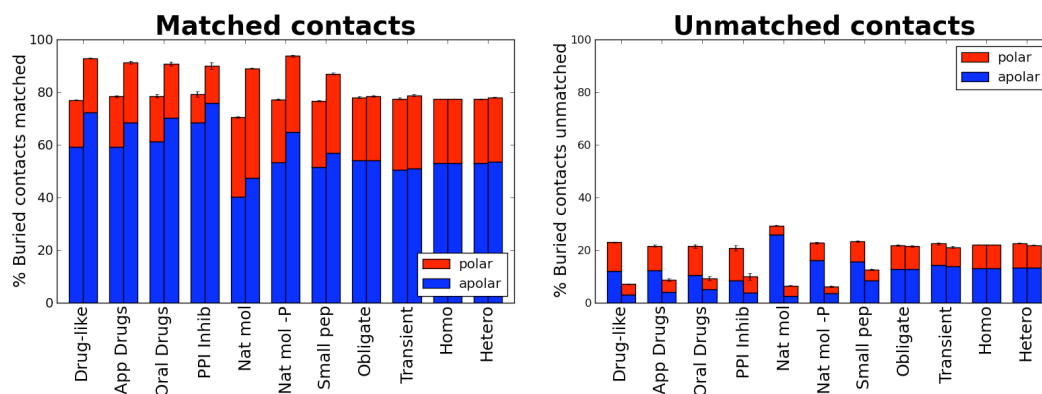


Figure 4.20. Mean of the percentage of buried atoms engaged in successful interactions (Matched contacts, left chart) and mean of the percentage of buried atoms without an appropriate partner in the other side of the interface (Unmatched contacts, right chart). The percentage is divided into polar (red) and apolar (blue) contribution. Each subset has two bars, one on the left for the atoms in the protein and one on the right for the atoms in the ligand or smaller protein in the case of protein complexes. Error bars denote the standard error of the mean. Subsets are ordered from left to right: Drug-like small molecules, approved Drugs, oral drugs, PPI small molecule inhibitors, natural molecules, natural molecules without phosphor, small peptides, obligate protein-protein dimers, transient protein-protein dimers, homo quaternary protein-protein interfaces and hetero quaternary protein-protein interfaces.

But, does nature make the most of its polar composition? Plotting the ratio of heteroatoms by number of atoms versus the ratio of polar interactions (as polar by sum of contacts), linear correlation (Figure 4.21) has been found for the natural-product-like subset in natural molecules. For these small molecules, the increase in polar features translates into more polar interactions with the protein. I note here that I have not analysed whether the proteins bound with these molecules are their putative partners, as that would be a one-to-one manual check. Nevertheless, this is a remarkable result, as it proves that it is possible in principle for a small molecule drug to engage many polar interactions with its partner.

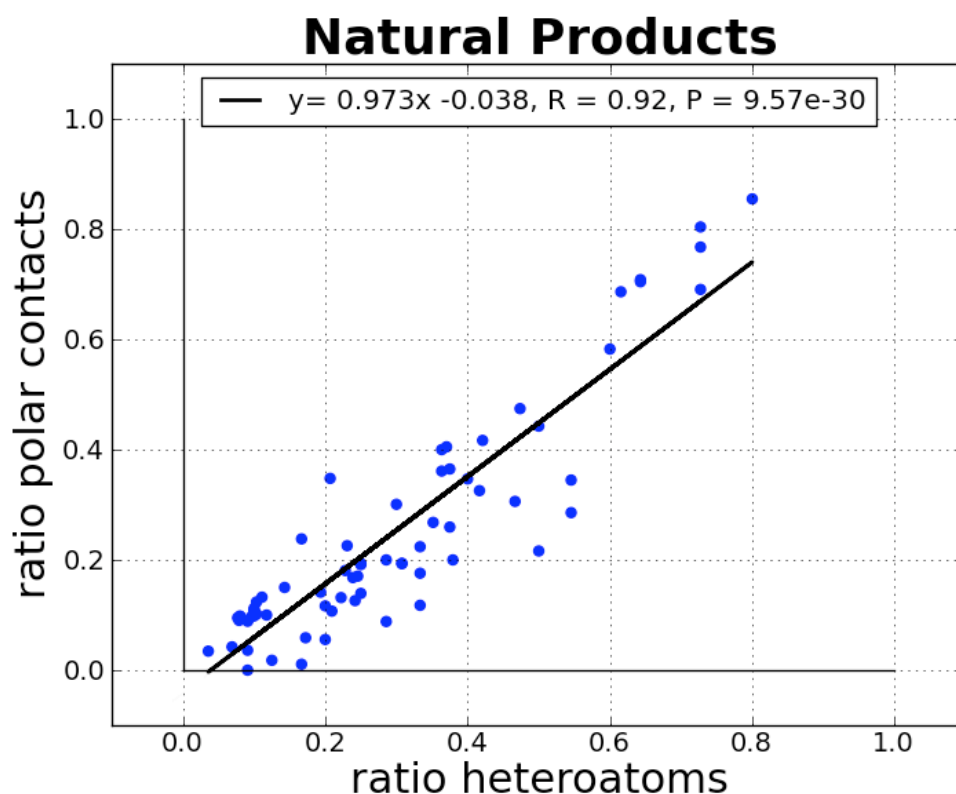


Figure 4.21. Linear correlation of the ratio of heteroatoms by number of heavy atoms versus the ratio of polar contacts by sum of contacts for the natural-product-like subset of the natural molecules set.

4.3.6 Drug-like complexes. Property versus interaction profile

Analysis of the distribution of the polar ratio across molecular weight, alogP , buried area upon binding and sum of contacts has been carried out for the synthetic drug-like molecules. Figure 4.22 shows these distributions colour-coded by SCOP family.

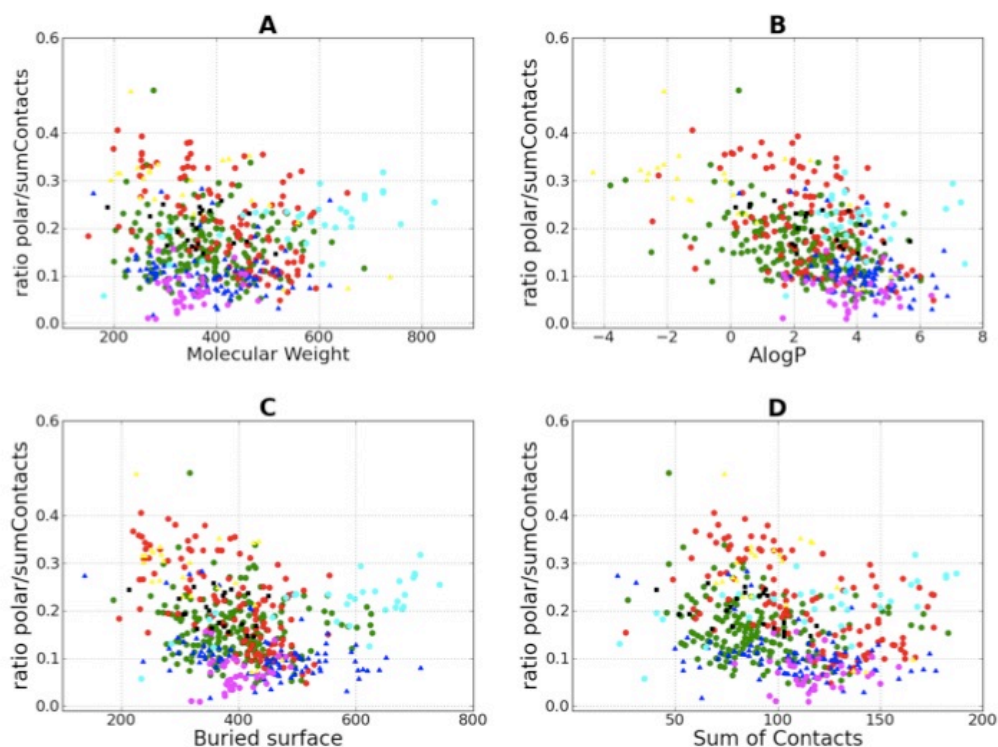


Figure 4.22. Ratio of polar/(polar+apolar) versus molecular weight (A), AlogP (B), buried area upon binding (C) and sum of contacts (D) for protein complexes with drug-like small molecules. Different colours denote SCOP families: Protein kinase catalytic subunit (green), nuclear receptor ligand-binding domain (blue), eukaryotic proteases (red), retroviral proteases - retropepsin (cyan), reverse transcriptase (magenta), Higher-molecular weight phosphotyrosine protein phosphatases (yellow), HSP90 N-terminal domain (black). For clarity, only SCOP families binding to more than 20 different ligands are shown.

Drug-like molecules bound to protein kinases (green dots in Figure 4.22) tend to have high alogP and hardly pass the threshold of 30% of polar contacts, not even the few that have alogP in the negative region. However it is also possible to have almost 50% of polar contacts (hetID 3C3 in 2CGW, (Foloppe *et al.* 2006)). In the case of nuclear receptor ligand-binding domain (NR-LBD, blue triangles in Figure 4.22), all the molecules have alogP > 1 and most of them do not have more than 15% of polar contacts, but as in kinases it is possible to have a more polar binding profile (34% of polar contacts for hetID 444 in 1UHL, (Svensson *et al.* 2003)). Eukaryotic proteases (red dots in Figure 4.22) bind to a wide range of molecules from 200MW up to 700MW with alogP between -2 to 6 with a varying percentage of polar contacts.

Bigger and more lipophilic molecules bind to retroviral proteases (cyan dots in Figure 4.22) with similar polar binding patterns as they do in eukaryotic enzymes, although polar fragments are also found. The proteins belonging to the reverse transcriptase SCOP family present similar characteristics to NR-LBD, and bind to molecules with $\text{alogP} > 1$, with none having more than 15% polar contacts. Most of the molecules bound to phosphotyrosine protein phosphatases (PTPP, yellow triangles in Figure 4.22) have around 30% of polar contacts with low alogP range (-2 to 2), although there are also three apolar binders (hetID 892 in 1T49, hetID BB3 in 1T48 and hetID FRJ in 1T4J, (Wiesmann *et al.* 2004)). However, these apolar molecules are inhibitors binding to an allosteric site. Finally, HSP90 domains (black squares in Figure 4.22) bind to molecules with a wide range of alogP (0-6) engaging between 15-25% of polar contacts.

Overall trends for drug-like molecules depend on their targets, but there is no correlation between lipophilicity (alogP) or molecule size (molecular weight) with the proportion of polar contacts in the bound complex. Although the highest polar profiles occur with lower alogP and small size molecules, one can see for instance that, for those with an alogP value of 4 drug-like molecules are in the range of 4% to 37% polar contacts.

4.3.7 Drug-like complexes. Affinity versus interaction profile

From the 1,206 distinct small molecules in the drug-like set, almost 700 have affinity data (K_d , K_i or IC_{50}) from the implementation of PDBBind (Wang *et al.* 2004) in CREDO. Unfortunately, not many natural molecules have affinity data, and comparison with how these molecules achieve high potency cannot be done with the current data available. However, there are 112 distinct small peptides with K_d , K_i or IC_{50} in CREDO. Transformation into free energy of ligand binding (Kcal/mol) for qualitative comparison was done with the thermodynamic law: $\Delta G = -RT \ln K_d$, where R is the gas constant ($1.9872E-03 \text{ Kcal mol}^{-1} \text{ K}^{-1}$), T is the temperature in Kelvin (taken as 300K,

ambient temperature) and K_d is the equilibrium dissociation constant of the binding event. When K_d was not available, IC_{50} or K_i were taken instead. Figure 4.23 shows there is no relation between the binding energy and the proportion of polar contacts the small molecules or small peptides made with their protein partners. Higher polar ratios occur only for drug-like weak binders with molecular weight below 300Da. As seen before, for the drug-like molecules, only weak small fragments can achieve many polar interactions. This is not the case for small peptides, where high polar contact ratio can be achieved across a wide range of affinities.

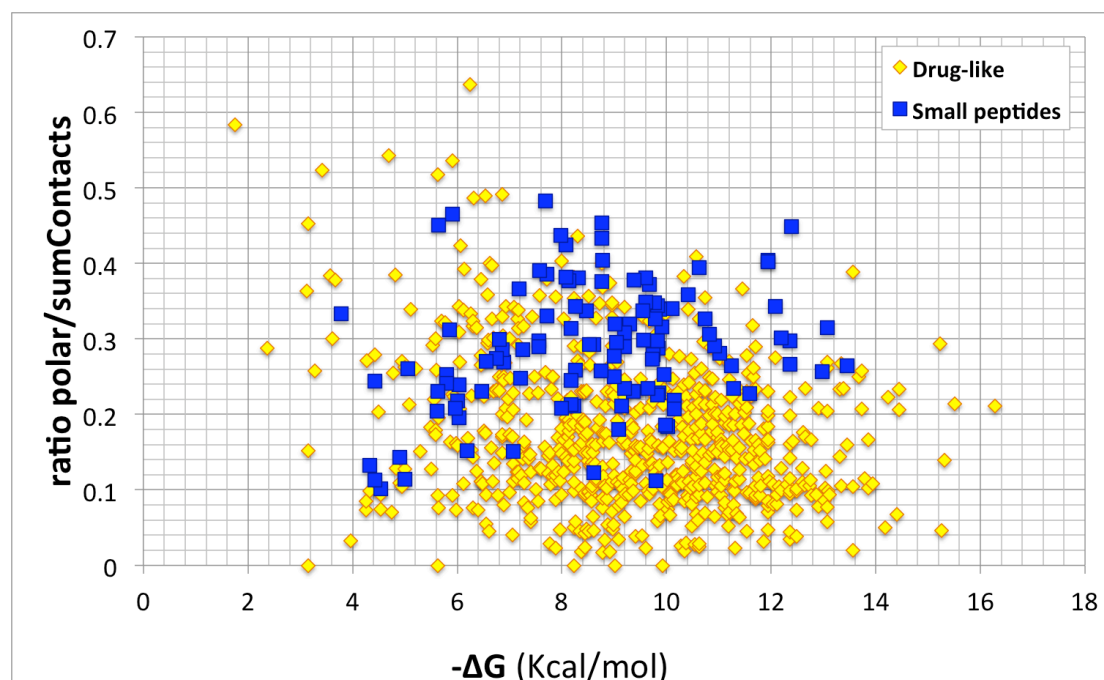


Figure 4.23. Free energy of ligand binding versus the polar ratio of contacts [polar/(polar+apolar)] for the drug-like set (yellow) and the small peptide set (blue).

Indeed, Figure 4.24 (A) shows that the most potent drug-like molecules have on average more atoms and higher alogP , whereas the average count of hydrogen bond acceptors and donors remains constant across the whole range of potency. This result is in agreement with the much discussed general trend in drug discovery of gaining potency by adding lipophilicity to the small molecules; see for example (Leeson *et al.* 2007). In

the set studied here, this translates (Figure 4.24, B) into a lower ratio of heteroatom content and a lower ratio of polar interactions.

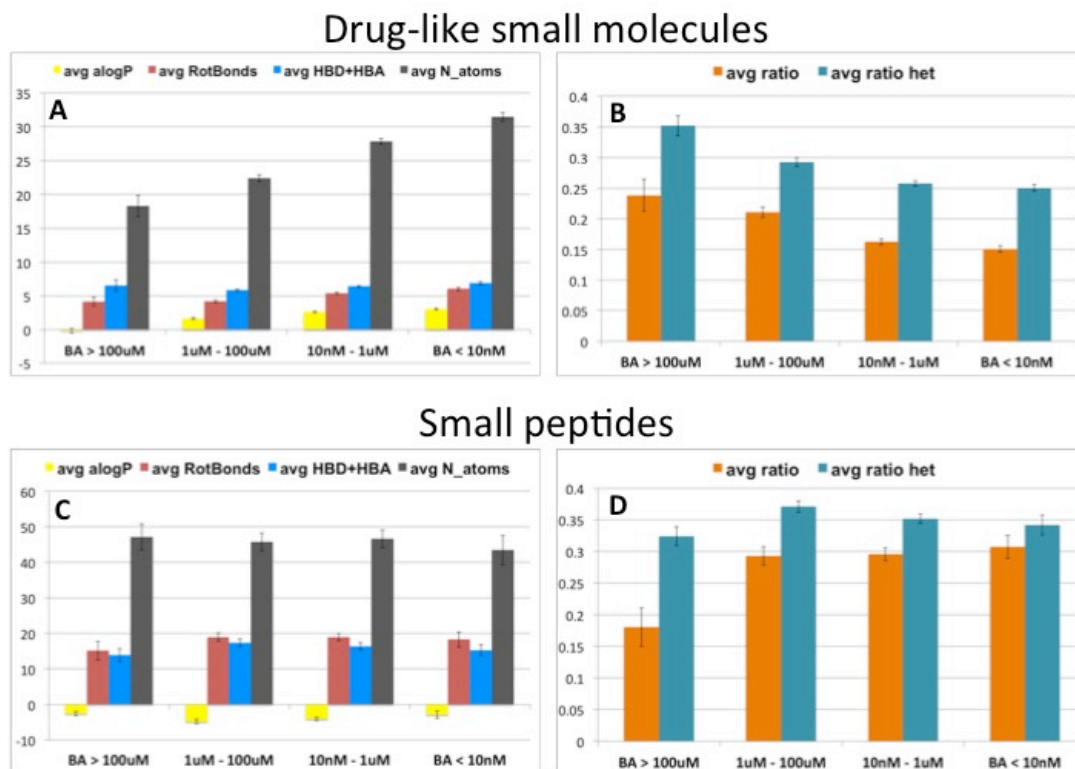


Figure 4.24. Binned binding affinity (BA) data for drug-like small molecules (A and B) and for small peptides (C and D). Bars in A and C denote the average of molecular properties for each affinity bin: alogP (yellow), rotatable bonds (red), sum of hydrogen bond donors and acceptors (blue) and number of atoms (black). Bars in B and D denote the average of the ratio of polar contacts [polar/(polar+apolar)] (orange) and the average of the ratio of heteroatom content [num heteroatoms/num atoms] (cyan). Error bars are the standard error of each sample.

Figure 4.24 (C and D) shows that small peptides have on average the same BA property profile regardless of their potency. From this result, it is clear that small peptides do not achieve tight binding through increase of lipophilicity, furthermore the proportion of polar contacts and heteroatom content is maintained across the whole range of affinities with the exception of weak binders where the polar ratio is lower. However, there are only seven complexes in this category. The important point to highlight here is that peptides manage to increase their affinity for their receptors maintaining

specific interactions, arguably through their flexibility. Note that the average of rotatable bonds in small peptides is four-fold higher than drug-like molecules. In fact, small peptides studied here are bigger than drug-like molecules, as no size limit was applied to select the small peptide set, whereas drug-like molecules larger than 900Da were omitted. Plotting the free energy of binding versus the number of atoms for both subsets (Figure 4.25) confirms that there is no correlation between the number of atoms and the free energy of binding for small peptides. However, they are less efficient than small drug-like molecules as they use more atoms to achieve the same affinity. Furthermore, the values for binding affinities are confined in the range of what can be measured (tens of millimolar that translates into ~ 2 Kcal/mol to picomolar that translates into ~ 16 Kcal/mol). In this way, peptides are able to sample binding energies between 4 Kcal/mol to 14 Kcal/mol regardless their size, which translates into the flat bar representation in Figure 4.24 (D).

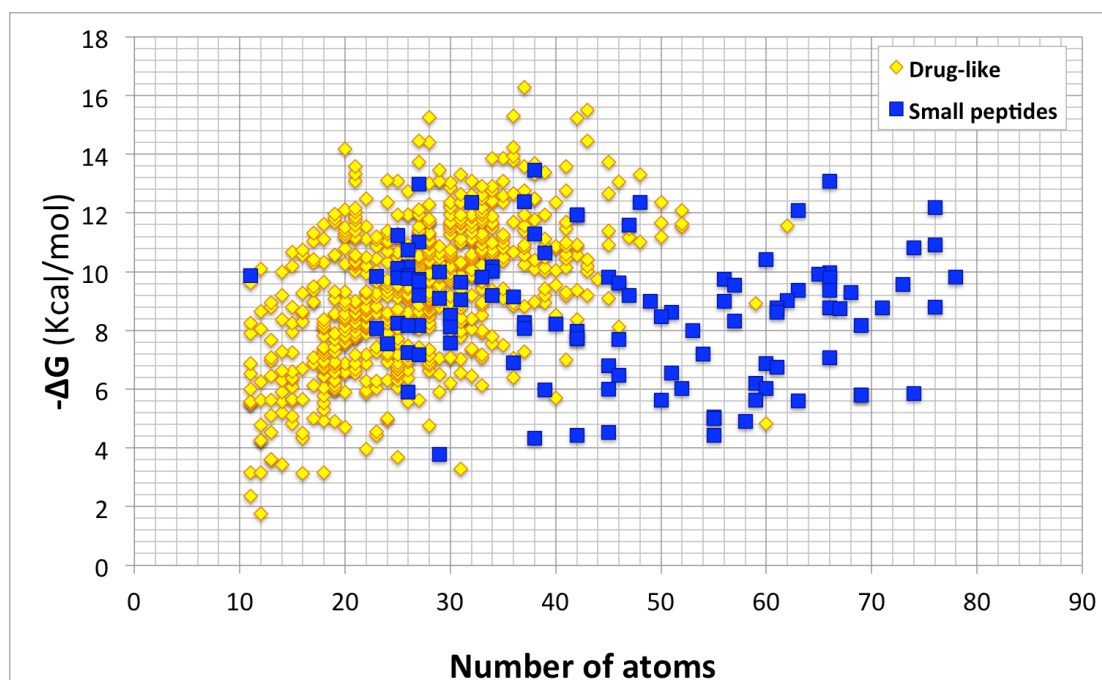


Figure 4.25. Free energy of ligand binding versus the number of atoms of the ligand. Drug-like set is plotted in yellow, and small peptide set in blue.

In Figure 4.25 there are three drug-like molecules with circa 60 atoms. The most potent is a symmetric cyclic urea HIV-1 protease inhibitor with 4nM affinity (1BWB, (Ala *et al.* 1998)). The weakest, with 300uM affinity is the detergent deoxy-bigchaps bound to IGF-1 through the steroid-like head, the two polar tails of the molecule are floating in the solvent (1IMX, (Vajdos *et al.* 2001)). The third molecule binds to calmodium with an affinity of 3uM; this is a big complex non-planar molecule, which binds to residues from the N- and C-terminal domains of calmodulin and induces a major conformational change (1XA5, (Horváth *et al.* 2005)).

4.3.8 Menagerie of small molecules for the same target

In this section, four examples of specific protein targets have been selected because they bind to small molecules from the different sets studied so far: drug-like molecules, approved and oral drugs, small peptides and natural molecules. For each protein, all small molecules bind to the same site. In this way we can visualise the property and interaction profile for the different sets of molecules binding to the same target. AlogP and % of polar contacts have been chosen to map these profiles, see Figure 4.26.

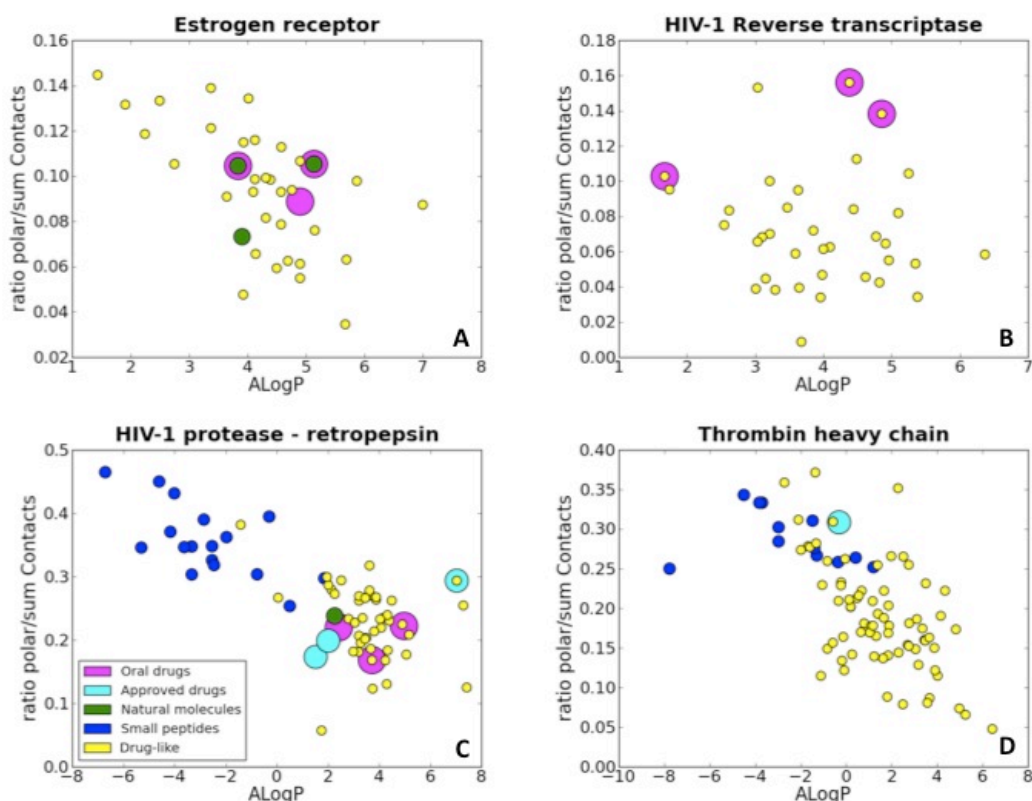


Figure 4.26. Ratio of polar/(polar+apolar) versus AlogP for four different proteins. A: Estrogen receptor (from NR-LBD SCOP family). B: HIV-1 Reverse transcriptase from Ribonuclease H SCOP family. C: HIV-1 Protease from retroviral proteases SCOP family. D: Thrombin heavy chain from the eukaryotic proteases SCOP family. Colour coding refers to the subsets, which the small molecules belong to: Oral drugs (magenta), Approved drugs (cyan), Natural molecules (green), Small peptides (blue) and Drug-like (yellow).

4.3.8.1 Estrogen receptor

Human estrogen receptor (ER) belongs to the NR-LBD SCOP domain family. As discussed before, this domain binds mainly to lipophilic molecules with low a ratio of polar contacts. Here we can see similar characteristics (Figure 4.26 A), for instance the natural product Estradiol is an approved oral drug with drug-like properties at alogP 3.8 with 10% polar contacts (DrugBank ID DB00286). Another natural molecule for this target that is also an approved oral drug is Diethylstilbestrol (DrugBank ID DB00255) with 11% of polar contacts and alogP 5.1. The drug-like molecule Raloxifene (DrugBank

ID DB00481) is another oral drug with similar profile, 9% polar contacts and alogP 4.9.

4.3.8.2 HIV-1 Reverse transcriptase

The reverse transcriptase domain of the HIV-1 Gag-Pol polyprotein belongs to the Ribonuclease H SCOP domain. Drug-like molecules binding to this protein (Figure 4.26 B) have a range of alogP between 1 and 7 with a modest proportion of polar contacts (1% to 16%). Interestingly, the three approved oral drug molecules have, within this range, the highest ratio of polar contacts: Efavirenz (16% of polar contacts, alogP 4.4, ChEMBL ID CHEMBL308954), Etravine (14% of polar contacts, alogP 4.8, DrugBank ID DB00625) and Delavirdine (10% of polar contacts, alogP 1.7, DrugBank ID DB00705).

4.3.8.3 HIV-1 Protease – retropepsin

The protease domain of the HIV-1 Gag-Pol polyprotein belongs to the retroviral proteases SCOP family. For this protein, drug-like molecules, approved and oral drugs, natural molecules and small peptides bind to the same site. As seen in Figure 4.26 (C), there is one cluster of small peptides in the polar corner, low alogP and more than 30% polar contacts. Remarkably, all the approved drugs analysed here have an oral administration route. Furthermore, all these six molecules are long and flexible with a range of lipophilicity (alogP from 1.5 to 7), and have 17%-29% of the contacts being polar. See Table 4.7 for their chemical structures.

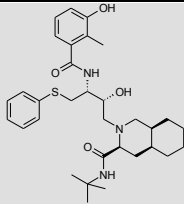
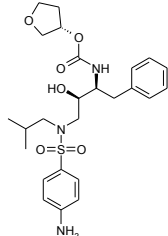
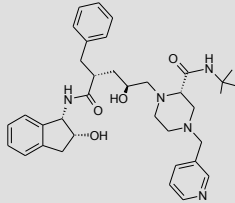
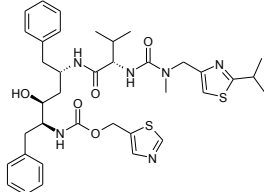
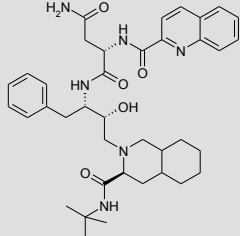
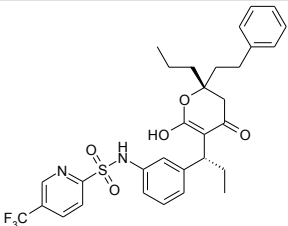
	hetID	PDB	Ratio	alogP	DrugBankID
	1UN	3ELO	0.17	3.7	DB00220 Nelfinavir
	478	1HPV	0.22	2.4	DB00701 Amprenavir
	MK1	1SDV	0.17	1.5	DB00224 Indinavir
	RIT	1HXW	0.22	5.0	DB00503 Ritonavir
	ROC	3EKQ	0.20	2.0	DB01232 Saquinavir
	TPV	1D4Y	0.29	7.0	DB00932 Tripanavir

Table 4.7. Chemical structures of the six oral drugs structural characterised in the PDB for the HIV-1 Protease. HetID is the residue identifier for the ligand in the PDB. Ratio refers to the ratio of polar contacts as [polar/(polar+apolar)].

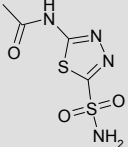
4.3.8.4 Thrombin

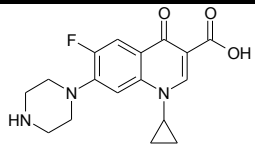
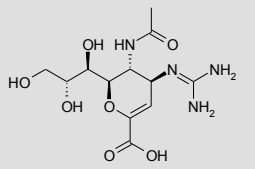
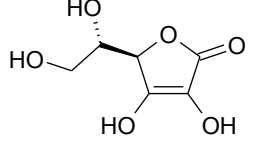
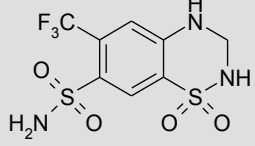
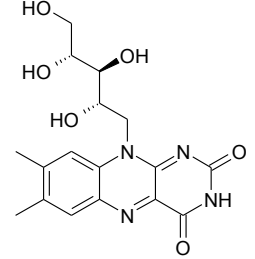
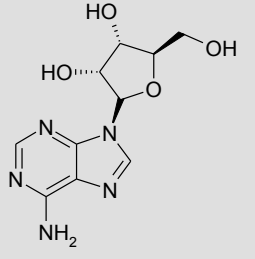
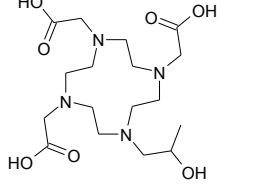
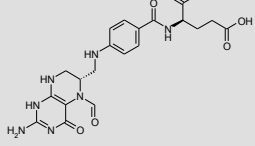
Thrombin heavy chain protein is a member of the eukaryotic protease SCOP family. It has drug-like binding properties with apolar interactions (high *a*logP and low polar ratio) as well as small peptides and peptidomimetics drug-like molecules in the more polar part of the graph (Figure 4.26 D). The only approved drug characterised in the PDB is Argatroban (DrugBank ID DB00278), with more polar contacts than the average drug-like molecules for this target. However, this drug is intravenous.

4.3.8.5 Approved and oral drugs

As seen in Figure 4.16 (B), approved and oral drugs have on average the same interaction profile as drug-like molecules. In this section, the four targets binding to different types of molecules show that oral drugs present a wide range of lipophilicity and ratio of polar interactions. In other words, it is possible to achieve more specific interactions without compromising the molecular profile of the drug leads. The case of HIV-1 protease, probably the most reported success of structure-based drug design (Wlodawer *et al.* 1998), proves that oral drugs can be long, complex and flexible molecules.

Here I report a summary (Table 4.8) of approved and oral small molecule drugs that have more than 40% of polar contacts. The low lipophilicity of these molecules is noteworthy.

	HetID (PDB)	Ratio	<i>a</i> logP	Drug type	DrugBank ID
	AZM (3HS4)	0.63	-1.3	Approved oral	DB00819 Acetazolamide

	HetID (PDB)	Ratio	alogP	Drug type	DrugBank ID
	CPF (1T9U)	0.61	-1.3	Approved oral	DB00537 Ciprofloxacin
	ZMR (2HTQ)	0.60	-5.0	Approved inhalation	DB00558 Zanamivir
	ASC (1F9G)	0.58	-1.9	Approved nutraceutical oral	DB00126 Ascorbic acid (vitamin C)
	HFZ (3ILU)	0.53	0.0	Approved oral	DB00774 Hydro- flumethiazide
	RBF (3DDY)	0.52	0.1	Approved nutraceutical oral	DB00140 Riboflavin (vitamin B2)
	AND (1UAY)	0.48	-1.9	Approved intravenous	DB00640 Adenosine
	DO3 (2QMI)	0.48	-8.9	Approved intravenous	DB00597 Gadoteridol
	FON (3GEH)	0.46	-5.0	Experimental Vitamin B complex	DB03256 Folinic acid

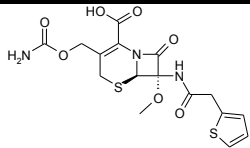
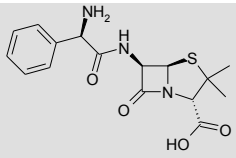
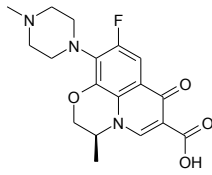
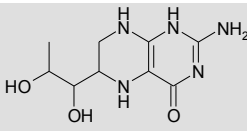
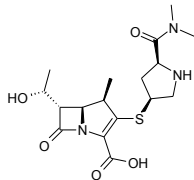
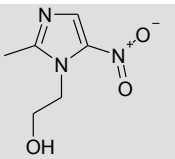
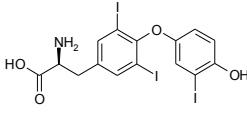
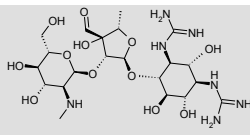
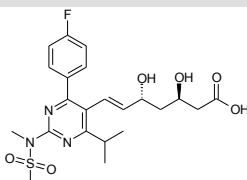
	HetID (PDB)	Ratio	alogP	Drug type	DrugBank ID
	CFX (1I2W)	0.46	-1.6	Approved antibiotic intravenous	DB01331 Cefoxitin
	AIC (2RDD)	0.46	-2.3	Approved oral antibiotic	DB00415 Ampicillin
	LFX (3K9F)	0.44	-1.4	Approved oral antibacterial	DB01137 Levofloxacin
	H4B (2DTT)	0.43	-1.1	Approved nutraceutical oral	DB00360 Tetrahydro- biopterin
	MER (1H8Y)	0.42	-4.9	Approved intravenous antibiotic	DB00760 Meropenem
	2MN (1W3R)	0.41	-0.2	Approved oral	DB00916 Metronidazole
	T3 (2PIV)	0.41	1.5	Approved oral	DB00279 Liothyronine
	SRY (3HAV)	0.40	-7.7	Approved intramuscular antibiotic	DB01082 Streptomycin
	FBI (1HWL)	0.40	0.9	Approved oral	DB01098 Rosuvastatin

Table 4.8. Approved and oral small molecule drugs that engage more than 40% polar contacts with their bound protein. HetID is the residue identifier for the ligand in the PDB. Ratio refers to the ratio of polar contacts as [polar/(polar+apolar)].

4.3.9 Small molecule inhibitors of PPI

In the introduction of this chapter, a question was left open: is the size and lipophilicity of small molecules inhibiting protein-protein interactions a requirement that small molecules need to fill in order to bind to protein interfaces? Although these molecules are on average lipophilic with few polar features, Figure 4.20 (B) shows that they have polar atoms unmatched in the binding site. Using TIMBAL database, I have extracted the seven cases where there is structural information for both the small molecule-protein and the protein-protein complexes. In all cases studied (see Table 4.9), the protein interface has more available polar contacts than the small molecule uses to bind to it. Figure 4.27 shows Bcl-XL binding to both BAD (magenta) and the Abbott compound ABT-737 (cyan), dotted lines represent the polar contacts each molecule does with Bcl-XL. This picture highlights the common pattern for synthetic molecules: fewer anchor points (understood as more constrained polar contacts) and more hydrophobic interactions that usually boost potency. Small molecule inhibitors of protein-protein interactions do not take advantage of the available polar contacts in the interfaces and only few of them are engaged. For these seven cases, comparison of the interacting residues in the target protein highlights that small molecules tend to use more aromatic and less charged residues than the protein partner.

Target	PDB p-p	ratio p-p	Affinity	PDB p-sm	ratio p-sm	Affinity	refs
IL-2	1Z92 (A:B)	0.35	10nM (Kd)	1PY2 (A)	0.21	60nM (IC50)	(Rickert <i>et al.</i> 2005) (Thanos <i>et al.</i> 2003)
Bcl-XL	2BZW (A:B)	0.19	6nM (Kd)	2YXJ (B)	0.08	0.6nM	(Lee <i>et al.</i> 2007)
MDM2	1YCR (A:B)	0.14	600nM (Kd)	1T4E (A)	0.03	67nM (Kd)	(Kussie <i>et al.</i> 1996) (Grasberger <i>et al.</i> 2005)
XIAP	1G3F (A:B)	0.22	-	1TFT (A)	0.12	-	(Liu <i>et al.</i> 2000) (Oost <i>et al.</i> 2004)
ZipA	1F47 (B:A)	0.10	21.6uM (Kd)	1Y2F (A)	0.00	12uM (Kd)	(Mosyak <i>et al.</i> 2000) (Rush <i>et al.</i> 2005)
TNF	1TNF (AB:C)	0.30	-	2AZ5 (C&D)	0.12	13uM	(Eck <i>et al.</i> 1989) (He <i>et al.</i> 2005)
S100B	1DT7 (A:X)	0.34	-	3GK1 (A)	0.12	-	(Rustandi <i>et al.</i> 2000) (Charpentier <i>et al.</i> 2009)

Table 4.9. Examples of polar/sumContacts ratio for proteins that bind to both protein partners (ratio p-p, left) and drug-like molecules (ratio p-sm, right). The PDB code includes the interacting chains, for example 1TNF(AB:C) denotes chain A and B interacting with chain C of the TNF trimer, whereas 2AZ5 (C&D) denotes chains C and D interacting with the small molecule. When available, affinity measure and units is specified in table. See Figure 5.24 to Figure 5.30 for a graphical representation of these examples.

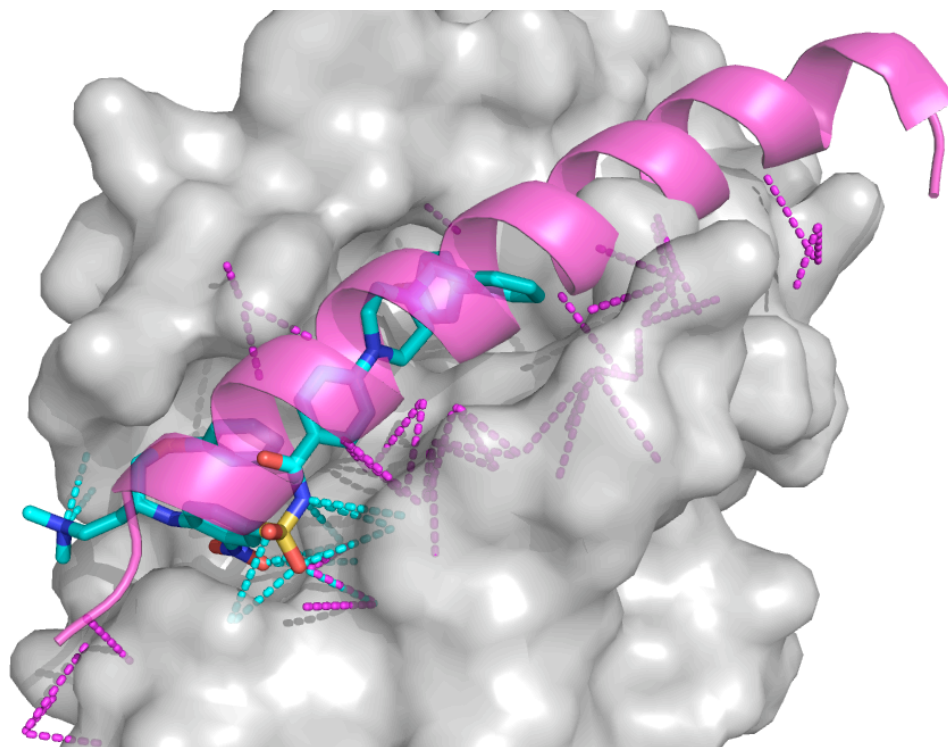


Figure 4.27. Bcl-XL bound with one of its putative partners (BAD, in magenta, PDB 2BZW) and with small molecule inhibitor (ABT-737, in cyan PDB 2YXJ). Only polar contacts are shown for clarity. Colour of the contacts (in dotted lines) is the same as the molecules making them. Synthetic molecule only uses a fraction of the polar contacts available for the natural counterpart.

4.3.10 Natural molecules and small peptides

Table 4.10 contains selected examples for proteins where there are structural data for both complexes, i.e. the protein target to the natural molecule or bound to a small peptide, as well as bound to a drug-like molecule. Figure 4.28 shows an example for each of these two classes.

Classical kinase inhibitors compete with endogenous nucleotides. There is a dramatic decrease in the polar/apolar ratio for kinase inhibitors compared with the ratio for the ADP, for example Abl kinase and MK2 in Table 4.10. However, these endogenous ligands do not need to cross the cell membrane. Indeed, the more polar drug-like kinase inhibitor of MK2 is reported to be inactive in cellular assays (Wu *et al.* 2007). It seems that drug-like molecules mimic natural ligands by engaging just a few of the available polar contacts

and raise affinity by increasing the lipophilicity, as too much polarity is not good for permeability (high probability of good rat bioavailability when $PSA \leq 140$ and $rotBonds \leq 10$, (Veber *et al.* 2002)). This might hold true for classical drug targets where much has been done to optimise molecular recognition and absorption. However, many authors ((Hann 2011), and references therein) associate high lipophilicity of the compounds entering drug development (amongst other factors) for the high failure rate of clinical candidates.

A comparison of small peptides and drug-like molecule binding to the same site demonstrates a smaller number of polar interactions but the difference is not as dramatic as for endogenous ligands. Table 4.6 shows that small peptides are not as polar (by the count of heteroatoms) as the endogenous ligands but they are much more flexible. When compared with drug-like molecules, we can understand the ability of small peptides to engage polar contacts by their flexibility, where drug-like molecules tend to be rigid scaffolds to minimise entropy lost upon binding.

Target	PDB p-nat	Ratio p-nat	Affinity	PDB p-DL	Ratio p-DL	Affinity	refs
Visfatin	2G96 (A&B)	0.29	53uM (Ki)	2G97 (A&B)	0.07	0.15uM (Ki)	(Kim <i>et al.</i> 2006)
Abl Kinase	2G2I (A)	0.39	-	2HZI (B)	0.06	70nM (IC50)	(Levins on <i>et al.</i> 2006; Cowan-Jacob <i>et al.</i> 2007)
MK2	1NY3 (A)	0.61	-	2PZY (A)	0.28	34nM (IC50)	(Under wood <i>et al.</i> 2003; Wu <i>et al.</i> 2007)
HIV Protease	1A94 (D&E)	0.37	14nM (Ki)	1D4Y (A&B)	0.29	8pM (Ki)	(Thaisri vongs <i>et al.</i> 1996; Wu <i>et al.</i> 1998)
Phospho lipase A2	2O1N (A)	0.20	-	2B17 (A)	0.13	620nM (Ki)	(Singh <i>et al.</i> 2006)
Alpha-Thrombin	1NY2 (2)	0.33	1.75mM (Ki)	1BCU (H)	0.20	0.53mM (Kd)	(Conti <i>et al.</i> 1998; Pillai <i>et al.</i> 2007)

Table 4.10. Examples of polar/sumContacts ratio for proteins that bind to both natural molecules, including small peptides (ratio p-nat, left) and drug-like molecules (ratio p-DL, right). The PDB code includes the interacting chains, for example 2G96(A&B) denotes chains C and D interacting with the small molecule. When available, affinity measure and units is specified in table. The three first target proteins bind to natural molecules and drug-like ones. The last three target proteins bind to small peptides and drug-like ones.

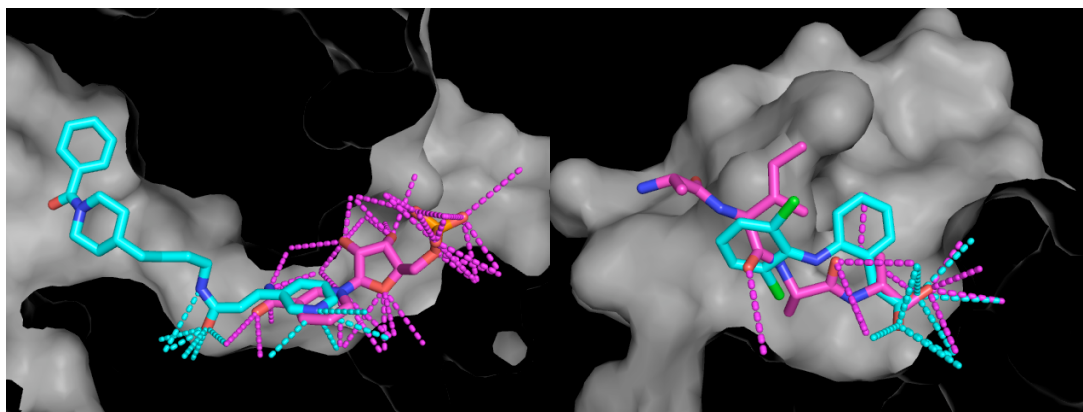


Figure 4.28. Examples of natural molecules (magenta) and drug-like molecules (cyan) binding to the same protein target. Only polar contacts are shown for clarity. Colour of the contacts (in dotted lines) are the same as the molecules making them. LEFT: Visfatin with Nicotinamide Mononucleotide (2G96) and FK-866 (2G97). RIGHT: Phospholipase A2 with a tetrapeptide (2O1N) and Diclofenac (2B17). In both cases synthetic molecules only use a fraction of the polar contacts available for the natural counterparts.

4.4 Discussion

Comparisons between the different sets of molecules consistently show more polar interactions between natural molecules (protein with protein, small peptides and natural molecules) than protein with synthetic drug-like molecules. Drug-like molecules are dominated by apolar contacts, specially the subset of molecules inhibiting protein-protein complexes. This is in accordance with the ITC data studied by Olsson *et al.* (Olsson *et al.* 2008), which demonstrated that synthetic molecules binding to proteins have greater entropic contributions than natural molecules.

On the other hand, it has been shown that small fragments present a more balanced signature with higher polar / apolar ratio than the average drug-like molecules. In fact, fragment hits are usually very polar and water-soluble (Congreve *et al.* 2008) (as they need to be in the high concentration format assay) and initial data show that they tend to present favourable enthalpy of binding (Ladbury *et al.* 2010). Being small and polar, fragments have a minor contribution from water displacement and therefore favourable enthalpic interactions have to overcome the entropic rigid body penalty (Ladbury *et al.* 2010). Indeed drug-like fragments have a higher proportion of polar contacts than bigger drug-like molecules. Similar results have been found by Ferenczy and Keseru (Ferenczy *et al.* 2010), analysing thermodynamic and structural data from public available databases. These authors found that, for maximal affinity compounds, binding is enthalpically driven for small ligands and entropically for larger ones.

Regarding protein-protein interfaces, quaternary interfaces (homo and heterogenic) have the same proportion of polar contacts as obligate dimers, whereas the transient dimers are slightly more polar than previously found (Nooren *et al.* 2003) and more similar to the small peptides subset. Overall, protein complexes present more polar interactions than synthetic drug-like molecules. Analysis of the protein complexes successfully inhibited by small

molecules, shows that these interfaces do not differ from other protein-protein interfaces, however the small molecules inhibiting them are at the apolar end of the already lipophilic drug-like spectrum. Furthermore, it was shown in chapter 2 that this type of molecules did not increase on average the numbers of hydrogen bond features with increase of molecular size, i.e. small efficient binders have on average the same number of hydrogen bond donors and acceptors than bigger less efficient molecules.

Having arrived at this conclusion I can argue that there are two plausible scenarios: (i) the proportion of polar contacts can not be improved for drug molecules due to the characteristics of druggable binding sites and the molecular property profile required of a drug, or (ii) improving the polar nature of drugs is hard but doable, although medicinal chemistry settings are not optimised for it.

Analysis of four individual protein targets with a menagerie of small molecules exemplifies the fact that, although there is a range of polarity and other properties specific for each target, it is possible to develop oral drugs with higher content in matched polar atoms. Moreover, it has been shown that there is no correlation between logP and the proportion of polar contacts in the binding mode. This result is encouraging, as it shows that the second scenario is likely, and it is possible to increase specific interactions without lowering too much the lipophilicity of the molecules.

Indeed, current drug discovery practises are being scrutinised by thermodynamic studies (Freire 2008; Olsson *et al.* 2008; Ferenczy *et al.* 2010; Ladbury *et al.* 2010). Retrospective analyses of two different targets, HIV-1 protease inhibitors and statins by Freire (Freire 2008) indicated that binding enthalpy gets better whilst improving already marketed drugs. This process, however has taken more than ten years with the current drug development settings. There is an emerging and also controversial (Erlanson 2011) viewpoint, suggesting that ligand interactions that enhance the enthalpic

contribution to binding are critical to optimise drugs and promotes the early use of thermodynamic assays in drug discovery (Freire 2009; Ferenczy *et al.* 2010; Ladbury 2010). Following the ligand efficiency metric (Hopkins *et al.* 2004), enthalpic efficiencies have been defined (Ferenczy *et al.* 2010; Ladbury *et al.* 2010) to guide prioritisation and modification of compounds towards a more balanced thermodynamic signature. Also, definition of ligand lipophilicity efficiency (LLE) (Leeson *et al.* 2007) and ligand efficiency indices, including polarity of molecules (Abad-Zapatero *et al.* 2010), for monitoring increase affinity without increasing much lipophilicity. Engineering polar contacts is an arduous task, as structural information is not always available, but thermodynamic assays may help medicinal chemists deliver less hydrophobic leads. It has also been suggested that increasing enthalpic contributions should be done from the starting small hits, as maximal enthalpy negatively correlates with ligand size (Ferenczy *et al.* 2010).

Furthermore, the more polar profile of natural molecules and small peptides is due to interplay between flexibility (measured by number of rotatable bonds) and number of heteroatoms to give the general interaction profile. Drug-like and natural molecules have similar numbers of rotatable bonds whereas small peptides are much more flexible. In contrast, drug-like molecules and small peptides have less heteroatoms than natural molecules. Therefore the more polar interaction profiles for natural products and endogenous molecules are due to more heteroatoms in the right constrained conformation. In comparison, small peptides engage more polar contacts due to their flexibility that allows them to reach specific interactions. Finally, looking at the proportion of these heteroatoms being matched I found that drug-like molecules have a larger proportion of unmatched polar atoms than the natural molecules, especially oral drugs and inhibitors of protein complexes. It seems that synthetic molecules are not making the most of their polar composition. However, that is easier said than done. Even when the protein target is known and structurally characterised, the design of polar

interactions is far from trivial and involves more than achieving the required atomic geometry (Freire 2009).

4.5 Conclusions

Molecular recognition is a complex event. It depends on the location and concentration of the molecules involved, their plasma or tissue distribution as well as physiological conditions. Structural dynamic fluctuations, protonation states and tautomerisms are important. In fact, atomic interactions are just one of the many factors involved in molecular recognition. However, in the case of drugs, they are key to the association of binding affinity with molecular properties, which in turn will impact in the ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) profile of the synthetic candidates. Currently, there is a consensus in the drug research community to try to keep these molecular properties within a "safe" range of drug-like space, particularly in keeping lipophilicity low. In this chapter I have analysed the atomic contacts between different sets of molecules, divided into natural and synthetic ones. The results presented here show that natural complexes typically engage more polar interactions than synthetic molecules bound to proteins. These drug-like molecules also have a higher proportion of unmatched heteroatoms than the natural sets and probably for this reason, show no correlation between logP and proportion of polar contacts, suggesting there is room to improve specific interactions without changing drastically the molecular properties of drug-like compounds. Nevertheless, the ratio of polar versus apolar contacts is greater when the size of the synthetic molecules is smaller. In other words, synthetic small fragments seem to anchor in sites with more specific interactions than the average size drug molecule. It has been discussed in recent conferences and meetings in the field, that one should aim to improve affinity of fragments before adding molecular weight in order to maximise the interactions with the original site hot spot. In this way, the starting point will have a path to grow to with more chances to succeed as a drug. For drug-like molecules in general, but in particular for the inhibitors of protein-protein interactions, I conclude that efforts should be invested to maximise polar contacts to better resemble the interaction patterns that natural molecules present as well as to minimise

promiscuity and poor ADMET profile. For all the reasons discussed here, it seems important to undertake this challenging task as early as possible in the discovery process, not only because it is the more feasible but also because it should ultimately reduce the costs of delivering safe drugs to the market.

5.1 Introduction

We have seen in chapter 4 the interaction profiles of protein-small molecule and protein-protein complexes with emphasis on the properties of the small molecules. In brief, they demonstrate that protein complexes and natural molecules tend to interact with higher ratios of polar to non-polar contacts than drug-like molecules. This chapter is concerned with the structural characteristics of the binding sites for these complexes, with the aim of highlighting differences, if any, between binding sites for each type of molecule studied in chapter 4.

5.1.1 Other studies classifying interfaces and cavities

Characterization of binding interfaces is crucial for the understanding and prediction of molecular recognition and it is not surprising that it has been the focus of many studies from different disciplines, for example binding dynamics, distinction of crystal contacts from biological relevant interactions in the X-ray structures, protein-protein docking scoring functions, homology modelling of protein complexes, protein engineering, quaternary structure generation, druggability target assessment, insight into toxicology issues due to promiscuous binding sites and prediction of function for orphan proteins. I summarise here other attempts to predict druggability and characteristics of protein-protein interfaces, including the subset that is known to be inhibited by small molecules.

5.1.1.1 Pocket detection and druggable interfaces

Two factors define a druggable protein target. First its modulation has therapeutic effect and second it is able to bind to a small drug-like molecule (Hopkins *et al.* 2002). Druggability predictors usually refer to the latter, recently redefined as “ligandability” (Edfeldt *et al.* 2011), mostly when the 3D structure of the target or a close analogue is known. These methods identify

and score pockets (or cavities) at the protein surface in terms of their likelihood of accommodating a small drug molecule. We can classify available tools by the algorithms that detect cavities and the scoring schemes that rank them (Perot *et al.* 2010). Comparative studies of the most commonly used tools are also available (Oda *et al.* 2009; Schmidtke *et al.* 2010) as well as servers that generate consensus solutions from several predictors, see for example (Zhang *et al.* 2011). Overall, the classification of protein-binding sites in terms of its druggability is centred on the identification and description of the available pockets. However, the definition of what is a pocket is not trivial and consequently has not been yet standardised (Fuller *et al.* 2009). Indeed, different pocket detection programs will give different results, and matching the ligand putative binding site is not always guaranteed (Capra *et al.* 2009). In addition, druggability scores are also biased by the training set, which usually includes only a few negative cases, and have low prediction power for new targets like protein-protein interactions. Nevertheless, an open source repository of druggable and un-druggable proteins is maintained to help to improve druggability scores (<http://fpocket.sourceforge.net/dcd>). For instance, analysis of these structures reveals that in addition to shape and hydrophobicity of the cavities, polar groups have an important role in molecular recognition and should be considered in the druggability predictions (Schmidtke *et al.* 2010).

Regarding residue propensity at the drug-like binding interfaces, Soga *et al.* (Soga *et al.* 2007) found that drug binding sites are richer in aromatic residues and Met, and are depleted in Pro, Lys, Gln and Ala. This study considered a 41-member, non-redundant set of proteins complexed with drug-like molecules and compared the residue composition at the binding interface (defined as residues within 4.5Å of the drug-like ligand) with the residue composition at the surface of a non-redundant set of 756 protein complexes.

5.1.1.2 Protein-protein interfaces

Numerous studies have analysed protein-protein interfaces deriving typical ranges for several interface properties, see for example (Nooren *et al.* 2003; Ofran *et al.* 2003; Janin *et al.* 2007; Keskin *et al.* 2008; Yan *et al.* 2008). Naturally, these ranges depend on the data analysed and the definition of what constitutes an interface, as well as the classification used to divide protein complexes into different types. Here, I briefly summarise the findings of Richard Bickerton from his analysis of the PICCOLO database (Bickerton 2009; Bickerton *et al.* 2011).

The non-redundant set of protein interfaces studied in this chapter is the same as Bickerton generated for his research. The findings highlight the similarity between the interface core and the protein core, and the interface periphery and the exposed protein surface. The core of the interface is more hydrophobic than the interface periphery; it is enriched with hydrophobic residues (Ile, Val, Leu, Phe, Met and Ala) and depleted of polar and charged residues (Asp, Gln, Asn, Glu, Lys and Arg). Between obligate and transient dimers, the obligate interfaces are more hydrophobic than the transient ones. In terms of pairing preferences, hydrophobic interactions, hydrogen bonds, salt bridges and disulphide bonds are important in macromolecular recognition. Hydrophobic residues favour other hydrophobic and avoid polar and charged residues. Aromatic residues prefer other aromatic or hydrophobic residues, although they also often interact with ions through the CH and the π system. Prolines interact significantly more with aromatic than other residue types. Positive charged residues (Arg, Lys and His) favour negative charged ones (Glu and Asp) but Arg-Arg, His-His and Arg-His are also common due to aromatic interactions, pi-cation and hydrogen bonds (with the main chain atoms) due to the versatile capability of these side chains. Regarding the number of contacts normalized by interface area, protein-protein complexes have on average 4% of the total contacts as hydrogen bond, 11% as ionic (including some of the hydrogen bonds), 10% as pi-cation,

10% as aromatic and 39% as hydrophobic. These average ratios are slightly different for transient and obligate dimers: 4% hydrogen bond, 8% ionic, 11% pi-cation, 11% aromatic and 40% hydrophobic for obligate dimers and 4% hydrogen bond, 12% ionic, 10% pi-cation, 8% aromatic and 36% hydrophobic for transient complexes.

5.1.1.3 Protein-protein interfaces inhibited by small molecules

Fuller and co-workers analysed the interfaces of several non-redundant sets of protein complexes (Fuller *et al.* 2009), including 134 protein-small molecule, 97 pairwise non-obligate hetero protein-protein complexes, 50 protein-marketed drugs and 24 small molecule inhibitors of protein-protein interactions. Their analysis was based on pocket identification using the program Q-SiteFinder (Laurie *et al.* 2005). The authors found that classical small molecules bound to proteins tend to occupy a single large pocket, whereas protein-small molecule inhibitors of protein-protein interactions target several smaller pockets in the same fashion as protein-protein complexes. Furthermore, they found that the pockets in protein-protein interfaces are often preformed in the free monomer and bound state, although there is an increase in the pocket volume upon binding, suggesting some degree of site adaptability, at least from the side chain atoms. Interestingly, this study showed that all ligands targeting the IL-2/IL-2Ra interaction bound not only to residues in the interface but also to residues that are not in direct contact between the two proteins. With respect to protein complexes with small molecules, Fuller *et al.* found that marketed drugs are the group that fill most efficiently the available volume in the active site pocket.

The 2P2I resource (Bourgeas *et al.* 2010) is a hand-curated database of the structures of protein-protein complexes with known inhibitors. Only targets with structural information for both the protein-protein complex and the protein-inhibitor complex are included in the database. In the first release,

there were 17 protein-protein and 56 protein-small molecule inhibitors and with these data the authors analysed the characteristics of the protein interfaces that I report here. Although a recent update (Morelli *et al.* 2011) has removed some of the entries (2P2I has now 12 protein-protein complexes with 39 non-redundant protein-small molecules), the original analysis gave a general overview of the protein-protein interfaces. Contrary to the views often expressed that some degree of site adaptability has to occur at the interface of protein complexes in order to bind to a small molecule (Wells *et al.* 2007) and in agreement with Fuller *et al.* (Fuller *et al.* 2009), Bourgeas and co-workers found that the root mean square deviation (rmsd) of the alpha-carbons of the bound protein complexes, the monomer and the monomer bound to a small molecule were in the same range as the resolution of the crystal structures, $1.12 \pm 0.4\text{\AA}$ on average. The authors analysed the structural data from these interfaces and compared them with representative heterodimeric protein-protein complexes. A classification of these protein-protein complexes was proposed, based on the number of continuous segments at the interface. Class I includes complexes with a few segments, three on average, and usually one of the partners is a small peptide or can be replaced by one. These complexes are also richer in elements of secondary structure at the interface, are more ordered and present lower affinity (in the micromolar range); they are also the complexes with the higher number of small molecule inhibitors. Class II complexes are usually formed by two globular proteins. They have more continuous segments, eight on average, and a higher proportion of unstructured elements at the interface and have affinities in the nano or sub-nanomolar range. In comparison with transient heterodimers, protein complexes with known inhibitors have on average, smaller interface size, similar geometric shape but fewer pockets, more hydrogen-bonds, fewer salt bridges (with the exception of IL2/IL2Ra) and fewer charged residues.

A recent report by Kozakov *et al.* (Kozakov *et al.* 2011) analysed druggability of ligand-hot spots at the interfaces of 15 protein-protein

complexes. The authors defined hot spots computationally, by solvent mapping the protein surface with 16 different chemical probes. For each probe, they clustered minimised docked poses and retained the lowest energy ones, authors called these 'probe clusters'. Then, these low energy poses were clustered again. Hot spots were defined as the consensus sites where multiple probes converged. These consensus sites were further expanded by side-chain flexibility for selected residues close to the original hot spot. In this way, druggable ligand-hot spots at protein-protein interfaces appear to be concave pockets with a "mosaic-like" pattern of hydrophobic and polar functionality, which are able to bind at least 16 probe clusters and one or two neighbouring hot spots. The authors concluded that these sites therefore have the ability to bind to hydrophobic drug-like molecules with some polar functionality. However, at least to my view, this ability is in part due to the methodology used to identify these hot spots. First, consensus sites were ranked by the number of probe clusters bound to them, where these probe clusters can be from different chemical probes. Secondly, the selection of close residues to explore flexibility was restricted to residues that have a least 75% of the total hydrophobicity calculated for all surface residues. Nevertheless, the authors successfully classify protein-protein targets as well as classical targets (in the supplementary information of the publication) with this approach.

The focus of this chapter is on the structural characteristics of the binding sites for the different types of complexes studied in chapter 4: protein-protein, protein-natural molecules, protein-small peptides and protein-synthetic small molecules (drugs and drug-like). Starting with a discussion of binding site definitions and the assessment of pocket detection algorithms, I analyse the residue propensity at the interfaces, the proportion of main chain as well as polar atoms at the interfaces, the depth of the protein contacting atoms and the density of contacts for each type of complexes.

5.2 Methods

5.2.1 Subsets definitions

Protein complexes studied here are the same as those described in chapter 4. See 4.2.3 for details. They are: small molecules protein-protein interactions inhibitors, small natural molecules, small peptides, drug-like small molecules, approved and oral drugs, obligate and transient dimers and protein-protein interfaces from quaternary assemblies. However, as this chapter focuses on the binding sites, I have used datasets filtered for UniProt and for SCOP family redundancy for the protein-small-molecule sets. See table 4.1 and section 4.2.1.5 for details.

5.2.2 Definition of binding interfaces and binding pockets

Conceptually, the binding interface is the region between binding partners. In this sense, a simple distance cut-off between atoms in the different binding entities is sufficient to define the interface. However, this definition is biased by what constitutes the binding entities and potentially can leave out areas capable of binding (see Figure 5.1 for an example). In addition, it is not suitable for un-bound structures.

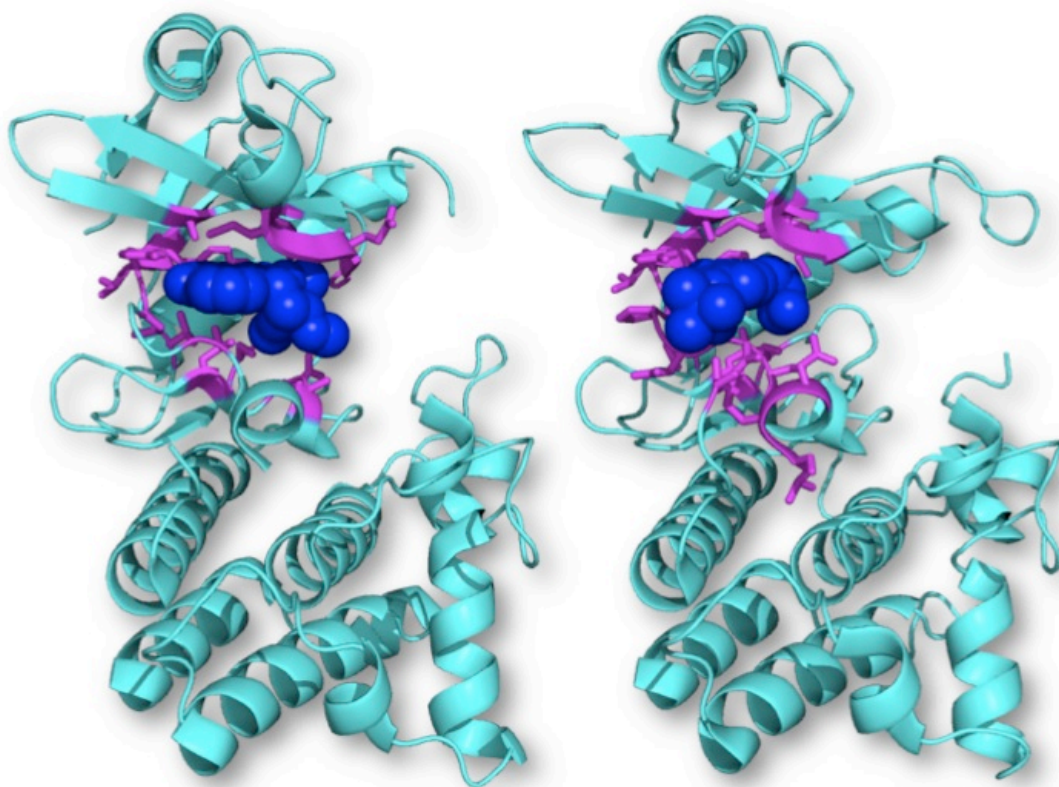


Figure 5.1. Structures of human IRAK-4 bound to different small molecules. LEFT: Staurosporine, 2NRY. RIGHT: benzimidazole inhibitor, 2NRU. Cyan cartoon represents the kinase domain; residues within 4.5Å of each ligand are displayed in magenta with stick representation of the side chains. The frontal loop of the five non-contacting residues is not shown for clarity.

An alternative approach, used extensively for small molecule binding is to use a cavity or pocket detection algorithm that will identify regions where a small molecule can bind. Thus, binding interfaces in this context are synonymous with the identified binding pockets, but this definition is not unbiased as pocket detection methods depend on the technique and parameters used (Fuller *et al.* 2009). For structures that are different from of the training set used to calibrate these parameters, the pockets that actually recognise ligands can be very different from what one would intuitively define as a pocket (Capra *et al.* 2009). Furthermore, in order to use these algorithms in an automated manner for broad datasets, they need to be accessible and able to run stand-alone from a workstation. In the course of the analysis presented here, I have used three pocket detection programs: ConCavity (Capra *et al.* 2009), Fpocket (Le Guilloux *et al.* 2009) and ghecom (Kawabata

2010) and compared them with the binding interfaces generated by a 4.5Å cut-off distance between binding partners. ConCavity output is limited to a residue-based score of the likelihood of a particular residue to belong to a binding site. The program does not give geometric properties of the sites and it will be not considered further.

5.2.3 Residue propensity plots

For each subset protein redundancy has been removed using the UniProt identifier (see Table 4.1). For these interfaces, the total number of residues within 4.5Å of the ligand or the other protein is recorded. Then, the percentage of each amino acid (or amino acid type) per interface (%res_i) is the total number of amino acids i-res divided by the total number of all amino acids at the interface:

$$\%res_i = \frac{\sum res_i}{\sum res}$$

In this way, we can compare compositions of the different sets by plotting the mean and standard error of these %res_i values for all 20 standard amino acids. The natural occurrence of each amino acid does not need to be taken into account when comparing interfaces, as natural abundances for each interface are the same for all. Comparison for all sets of molecules for all 20 standard amino acids is very content-rich, difficult to represent and to interpret. For this reason, comparisons of amino acid types rather than individual amino acids are also discussed here. There are several ways to divide the natural amino acids into different types. This division depends on the objective of the study. For further classification of these propensities several side chain properties or amino acid types have been grouped together:

- **Polarity**
 - Charged: Arg, Lys, Asp, Glu and His
 - Polar: Asn, Gln, Ser, Thr, Tyr and Trp
 - Hydrophobic: Ala, Gly, Cys, Val, Pro, Ile, Leu, Met and Phe
- **Size**
 - Small (4-7 heavy atoms): Ala, Cys, Gly, Pro, Thr, Val and Ser
 - Medium (8-10 heavy atoms): Asn, Asp, Gln, Glu, Ile, Leu, Lys, Met and His
 - Bulky (11-14 heavy atoms): Arg, Phe, Trp and Tyr
- **Flexibility**
 - Constrained: Pro
 - Free: Gly
 - Rigid (0-1 rot bonds): Ala, Cys, Ser, Thr and Val
 - Medium (2-3 rot bonds): Asn, Asp, Gln, Glu, Ile and Leu
 - Flexible (4-5 rot bonds): Arg, Lys and Met
 - Aromatic (2 rot bonds + aromatic ring): His, Phe, Trp and Tyr

5.2.4 Depth of the protein atoms at the interface

The ghecom program (Kawabata 2010) that detects pockets in protein structures was used to calculate the depth of the protein atoms at the binding interface. The idea behind this program is that a pocket is a region where a small spherical probe can enter but a big one cannot. The radius of the smallest big (inaccessible) sphere - $R_{inaccess}$ - gives a measure of the shallowness of the pocket. Kawabata has improved the performance of his original program phecom (Kawabata *et al.* 2007), by using mathematical morphology from set theory. For each atom of the protein, the $R_{inaccess}$ values of the surrounding spheres is averaged by harmonic mean (a special case of the power mean):

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}}$$

These values give a measure of the location of the atoms within the pocket. The program uses several probes with radii from 1.87Å to 10Å. Therefore a small value for Rinaccess means the atom is located deep within the pocket, whereas a Rinaccess of 10Å, means the atom is in a convex area or at the limits of the cavity. See Figure 5.2 for a schematic representation of the Rinaccess calculation.

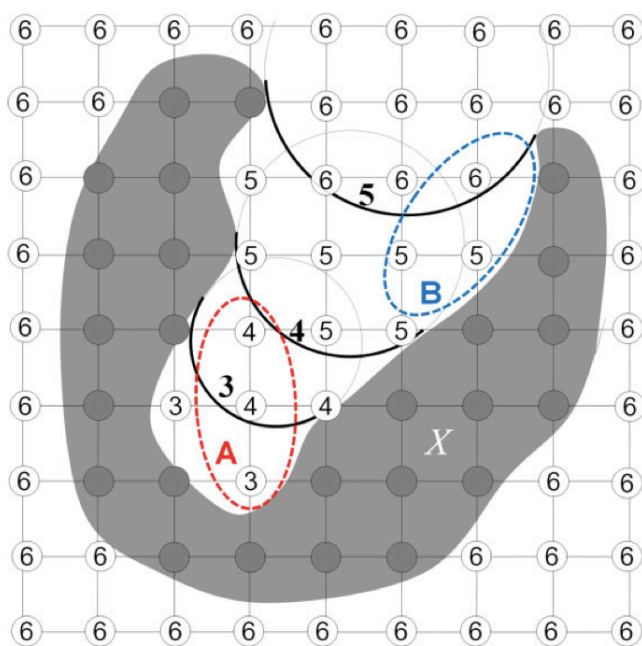


Figure 5.2. The concept of Rinaccess calculation. Reprinted from (Kawabata 2010). Three spherical probes are used: 3Å, 4Å and 5Å. Grid representation captures the smallest of the larger spheres that cannot access the grid point; the number represented is the radius of the sphere plus the grid resolution. Red and blue shapes represent different ligands bound in different regions of the pocket. The average of grid values per ligand gives a measure of the depth where the ligand is bound.

5.2.5 Size of the protein in protein-protein complexes

The size of the protein is taken from the PICCOLO database (Bickerton *et al.* 2011), as the number of standard residues composing the chain. Most non-standard amino acids are not accounted for, as the database does not consider them for the interaction pairs between chains.

5.2.6 Statistical treatment

The significance of the comparisons between distinct sets has been assessed by comparing medians of the calculated parameters. Because the distributions of the parameters analysed are not always normal, the non-parametric method of Kruskal-Wallis, implemented in the stats module in scipy (Jones *et al.* 2001 -), has been used for all comparisons. A difference is labelled as significant if the P value is lower than 0.05.

5.3 Results and discussion

5.3.1 Pocket detection algorithms

Fpocket output is information rich; not only does it provide geometric properties of the cavities identified, but also chemical characteristics, including a druggability predictor trained on small-pocket drug-like molecule binders (Le Guilloux *et al.* 2009). However, manual inspection of several examples indicated that the pockets found do not always match binding sites, as the region occupied by the binding partner. Therefore, it would be difficult to compare pocket properties across the subsets of molecules studied here. For example, in the case of human IRAK-4, Fpocket predicts correctly the binding site for Staurosporine. But to cover the binding site of the benzimidazole inhibitor, merging of two predicted pockets is needed (see Figure 5.3) Although this may be a meaningful result, as this inhibitor stretches itself to occupy several pockets, this manual fine-tuning is not possible for widespread comparisons and Fpocket was not used further.

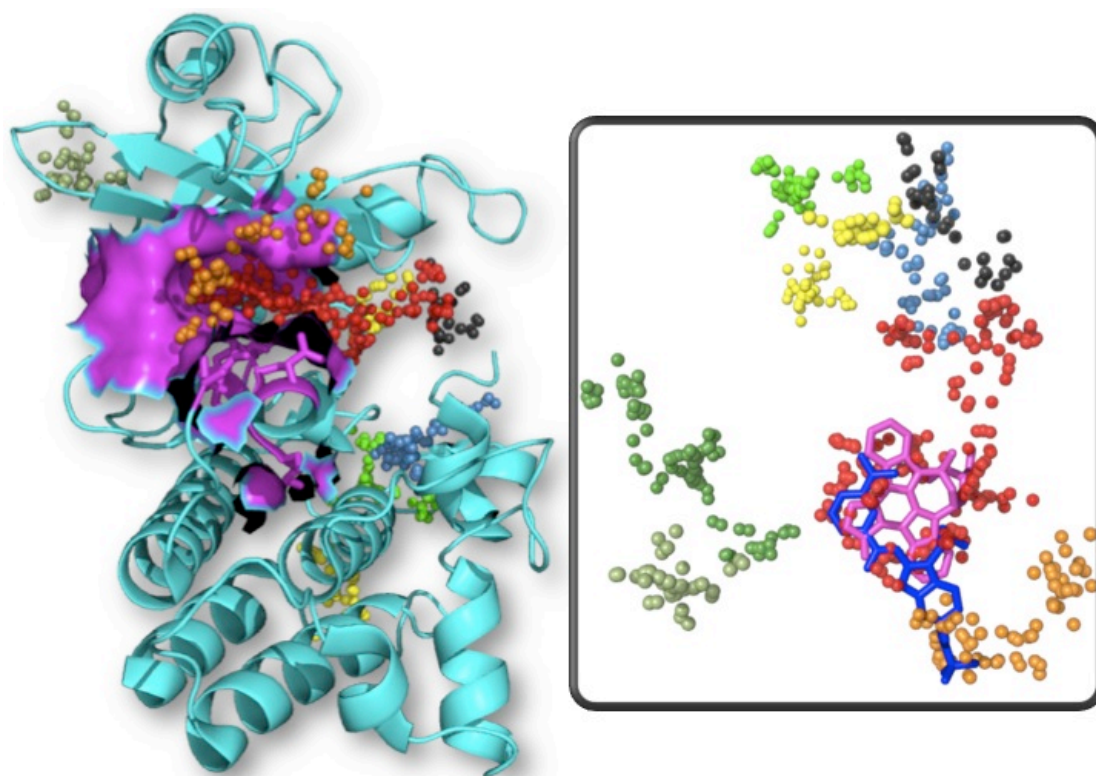


Figure 5.3. LEFT: IRAK-4 kinase domain (2NRU) bound to benzimidazole inhibitor (not shown). The magenta region represents the residues in contact with the inhibitor; coloured spheres represent the pockets predicted by Fpocket, where each colour represents a different pocket. RIGHT: Protein-based overlay of the pocket prediction from Fpocket shown on the left with Staurosporine from 2NRY. The benzimidazole inhibitor is represented by blue sticks, Staurosporine with magenta sticks. Note the binding mode for benzimidazole inhibitor is covered by pocket 1 (red) and pocket 4 (orange).

The program ghecom (Kawabata 2010), an evolved version of the original phecom (Kawabata *et al.* 2007), gives atomic detail of the pockets found on the protein surface. For all examples analysed, this program gives the more consistent prediction of pockets. The cavities described by ghecom are usually larger than the ones generated by other programs (see Figure 5.4 and Figure 5.5) but this characteristic is also what makes the output robust, in the sense that all surface atoms are explored and described. However, it is also the reason why this program is not considered further to define binding interfaces, as it is too sensitive and not specific. Nonetheless, calculations of

atom accessibility performed by this program are used to study the depth of the atoms at the interface.

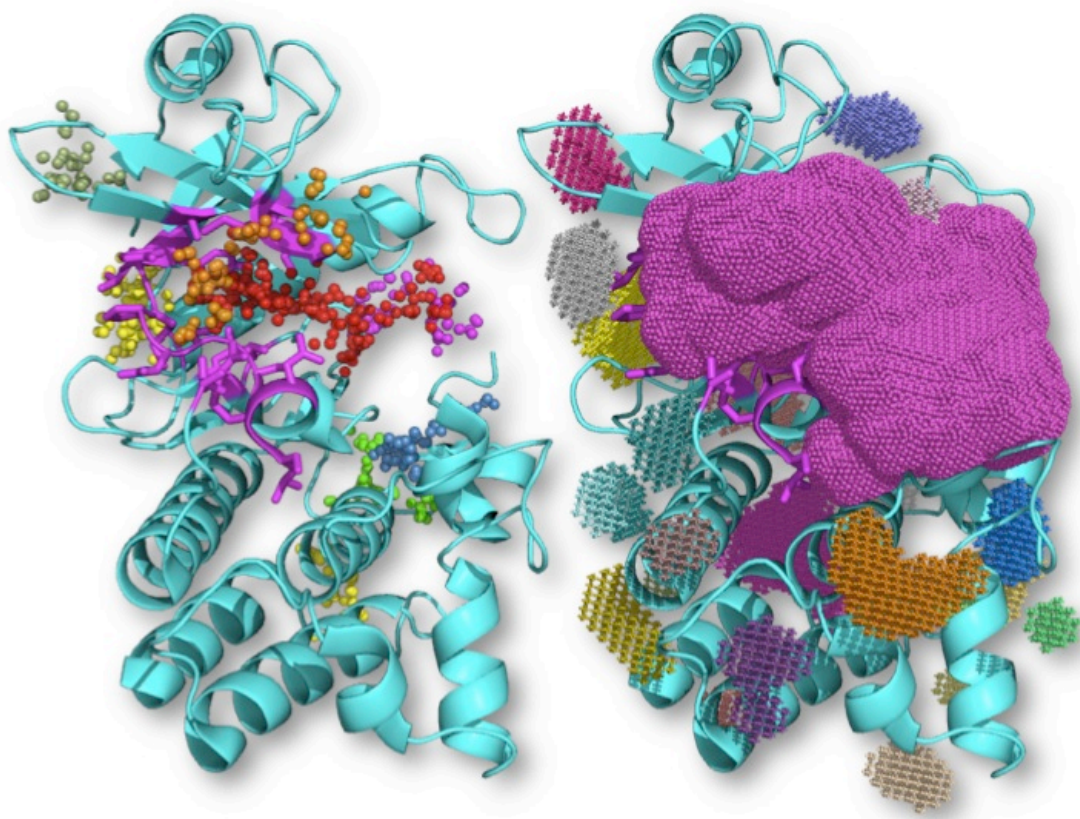


Figure 5.4. Comparison of pocket detection by Fpocket (LEFT) and ghecom (RIGHT) for the human IRAK-4 (2NRU). Ghecom gives one single large pocket in the ATP binding site (magenta cloud), whereas Fpocket gives several different ones. The small coloured clouds on the right picture are the additional pockets found by ghecom.

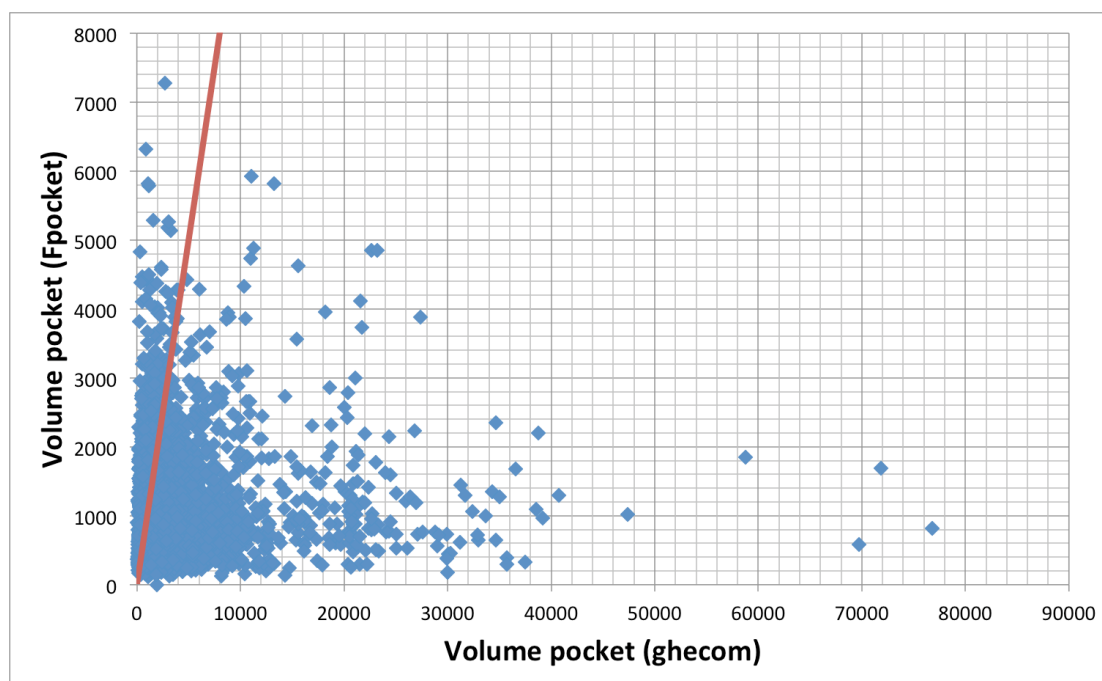


Figure 5.5. Comparison of the volumes for the pockets identified by Fpocket (Y axis) and ghecom (X axis) programs for the small molecule data sets (Drug-like, drugs, natural molecules and small peptides). These points represent the volume of the pocket that matched the ligand bound. Red straight line represents the line of slope one to aid comparison. One quarter (23%) of the pockets have greater volume for Fpocket than ghecom.

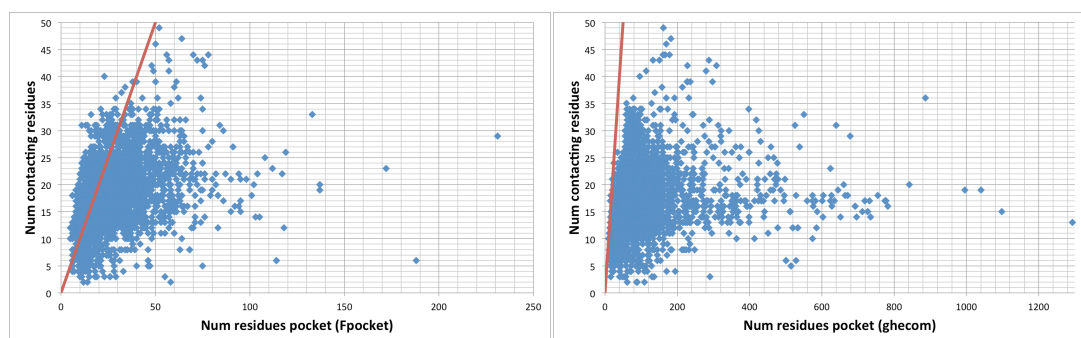


Figure 5.6. Comparison of the number of residues forming a pocket from Fpocket (left) and ghecom (right) predictions versus the number of contacting residues from the binding partner (buried residues). Red straight line represents the line of slope one to aid comparison. For Fpocket, 23% of the predicted pockets enclose fewer residues than the residues buried upon binding, on average 77% of the binding site is covered for these cases. For ghecom this proportion is less than 0.1%, and for these few cases more than the 90% of the binding site is covered by the prediction.

In conclusion, taking into account that all molecules studied in this analysis are complexed molecules, i.e. they are all bound to proteins, I define the binding interface as the region containing the atoms that are within 4.5\AA

distance of any atom of the binding partner. In this way, comparisons between subset-binding interfaces will refer from now on to the contacting atom (buried) between partners. In the case of, for example, different molecules binding to the same protein, differences at the interfaces will reflect the way that each molecule interacts.

5.3.2 Residue propensity

This section investigates the residue propensity of the binding sites. Assessment of the protein redundancy in the sets has been carried out by comparing the residue propensities of each set of molecules with distinct UniProt identifiers versus distinct SCOP families. Figure 5.7 to Figure 5.13 show the propensities for both levels of redundancy for all protein-small molecule complex sets. The residue propensities do not vary much between the two levels of redundancy, thus the analyses presented here use the subsets with distinct UniProt identifiers, unless otherwise stated.

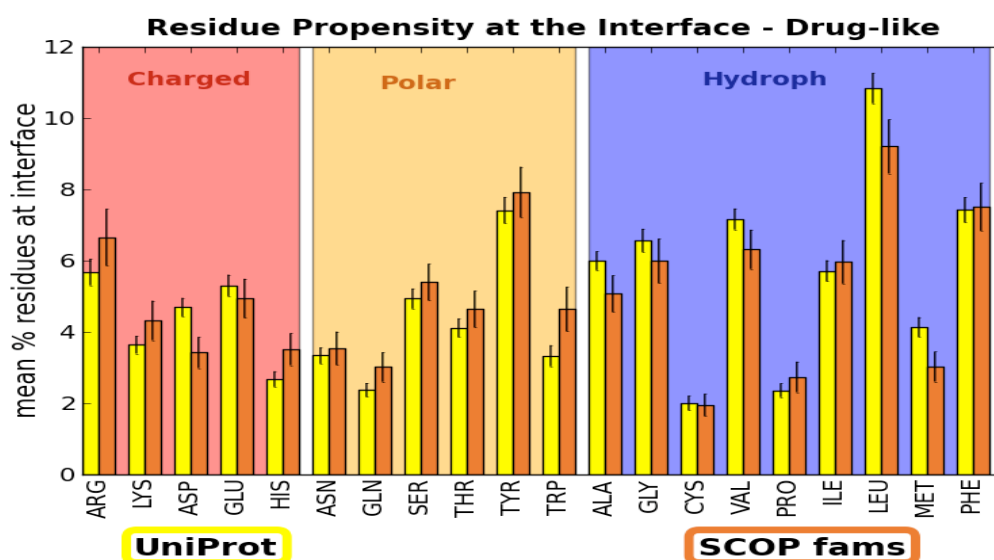


Figure 5.7. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (yellow) and SCOP family (orange) for the drug-like set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

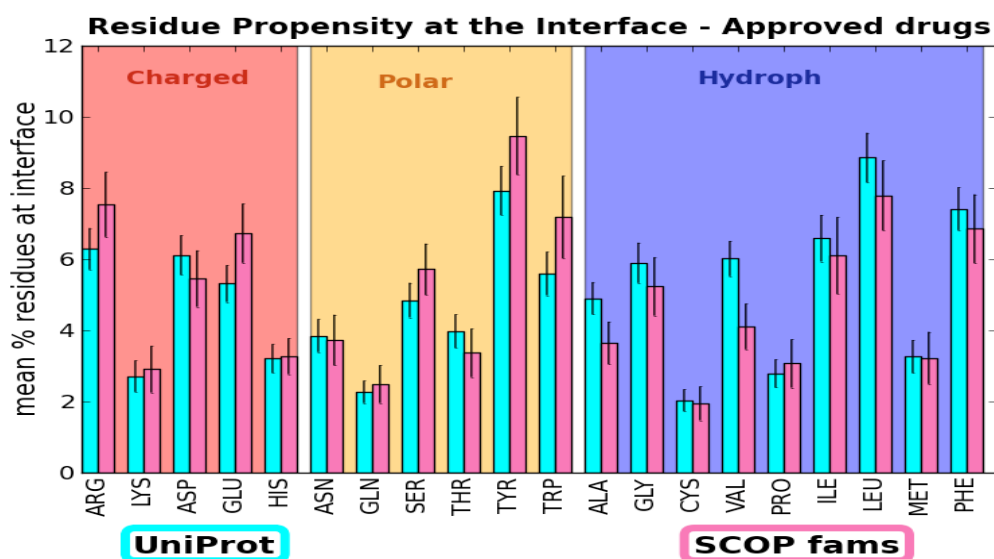


Figure 5.8. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (cyan) and SCOP family (magenta) for the approved drug set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

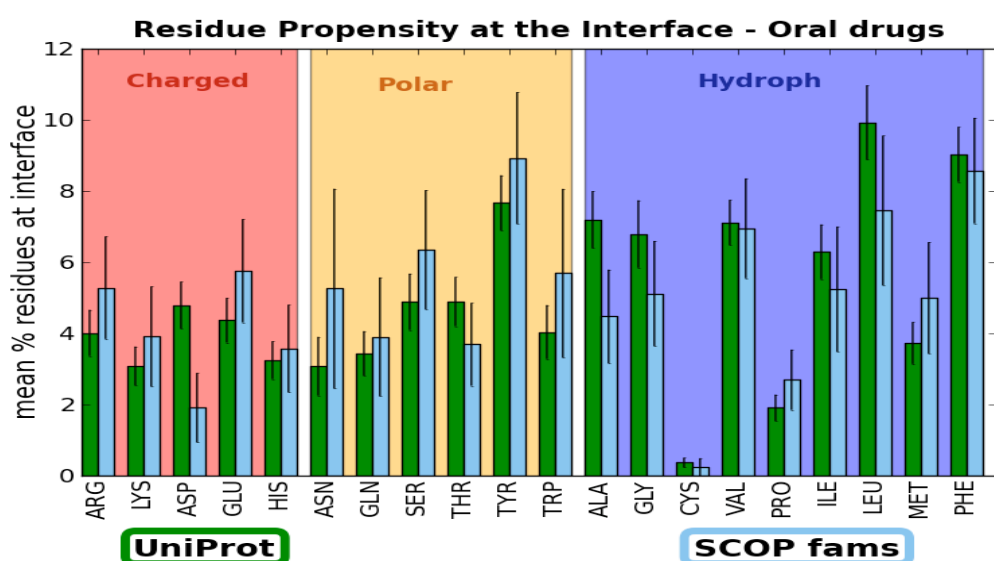


Figure 5.9. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (green) and SCOP family (light blue) for the oral drugs set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

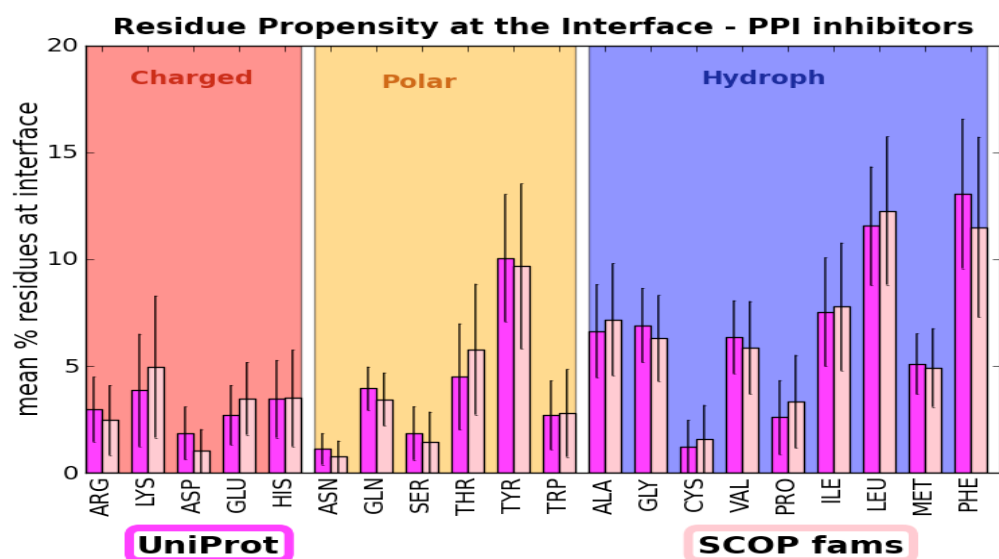


Figure 5.10. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (magenta) and SCOP family (light pink) for the small molecule PPI inhibitors set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

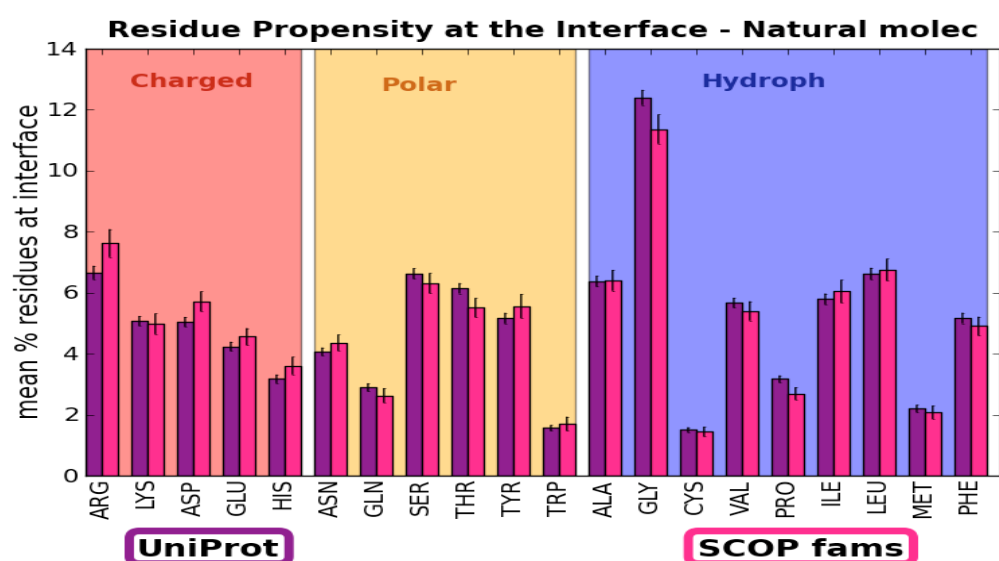


Figure 5.11. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (purple) and SCOP family (bright pink) for the natural molecules set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

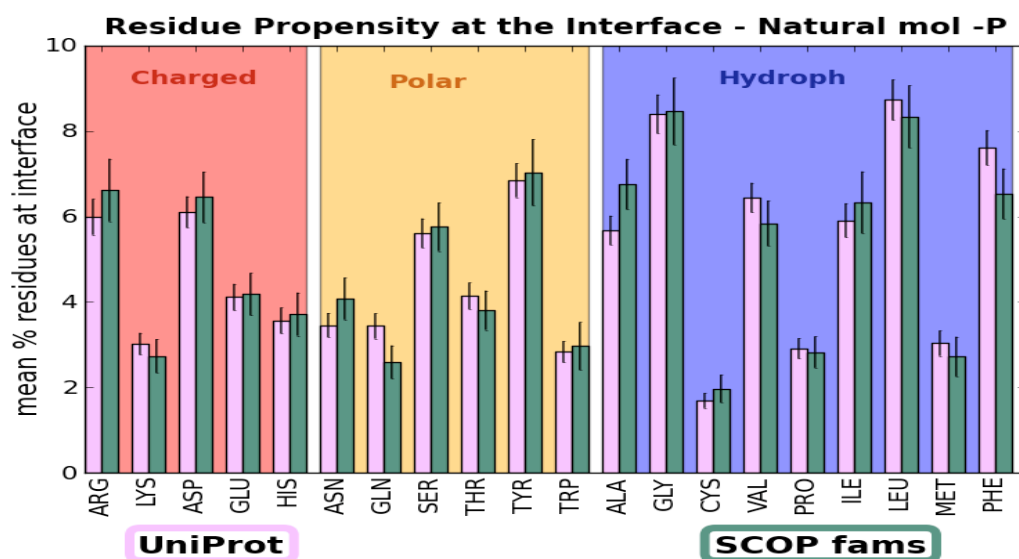


Figure 5.12. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (pale pink) and SCOP family (pale green) for the natural molecules not containing phosphorus set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

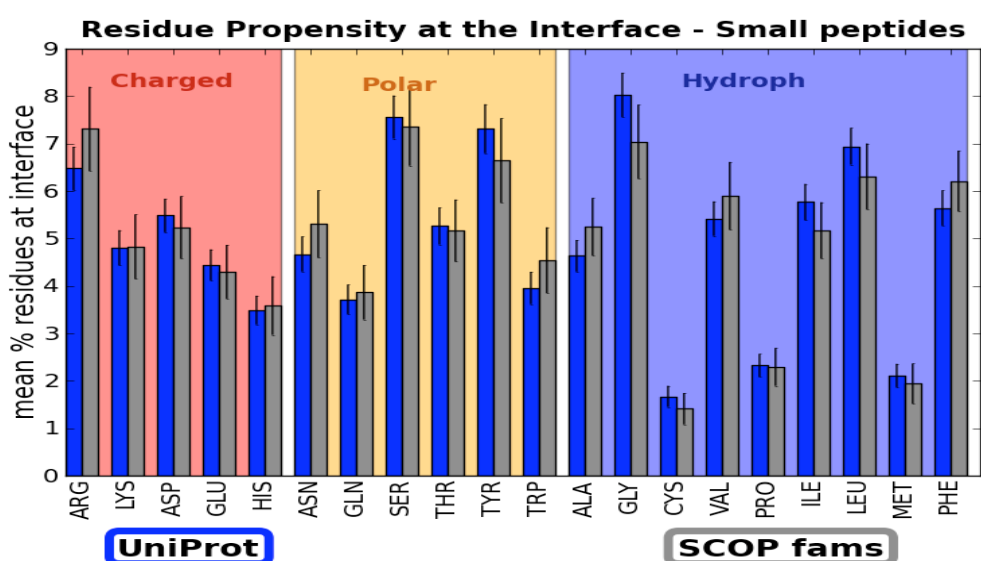


Figure 5.13. Comparison of residue propensities at the binding sites for the two levels of protein redundancy, UniProt (blue) and SCOP family (grey) for the small peptides set. Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

Regarding the protein-protein complexes, residue propensity can be studied for all chains of the assembly. In this work, I use PICCOLO interaction data that it is structured into interacting pairs of chains. For analogy with the protein-small molecules complexes I choose to represent only the propensity of the long chain of the protein-protein interaction. In fact, this is an arbitrary choice, as the multiple chains interacting in an assembly do not have analogies with small molecule binding sites, but it eases the representation and interpretation of the graphs. Figure 5.14 to Figure 5.17 show that there is virtually no difference in the residue propensities of long and short chains of the protein-protein complexes, with the exception of the transient hetero- and homo-dimers. This difference might be due to the high proportion of structures in this subset where one globular domain interacts with a shorter peptide.

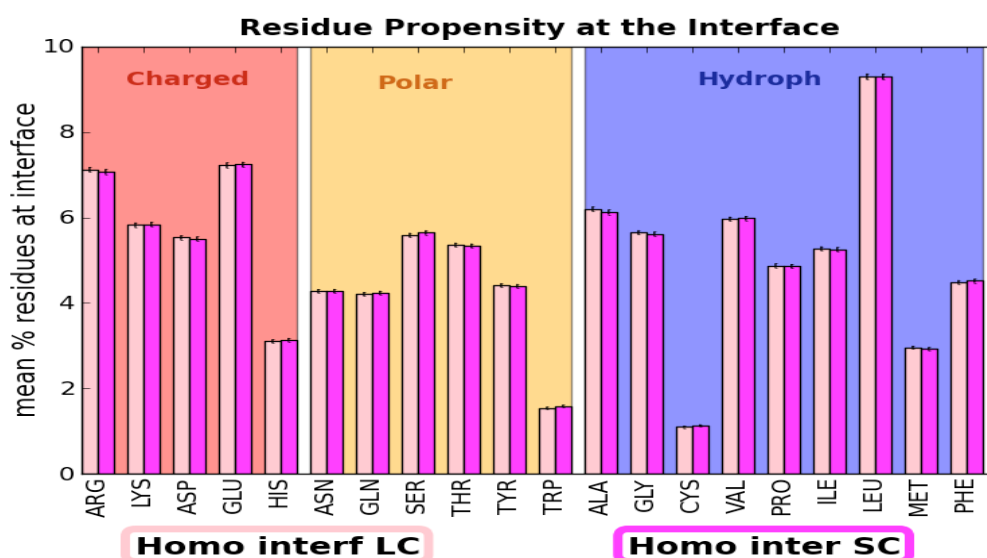


Figure 5.14. Comparison of residue propensities for long chain (LC, pale pink) and short chain (SC, magenta) of the homo quaternary interfaces of the protein complexes. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

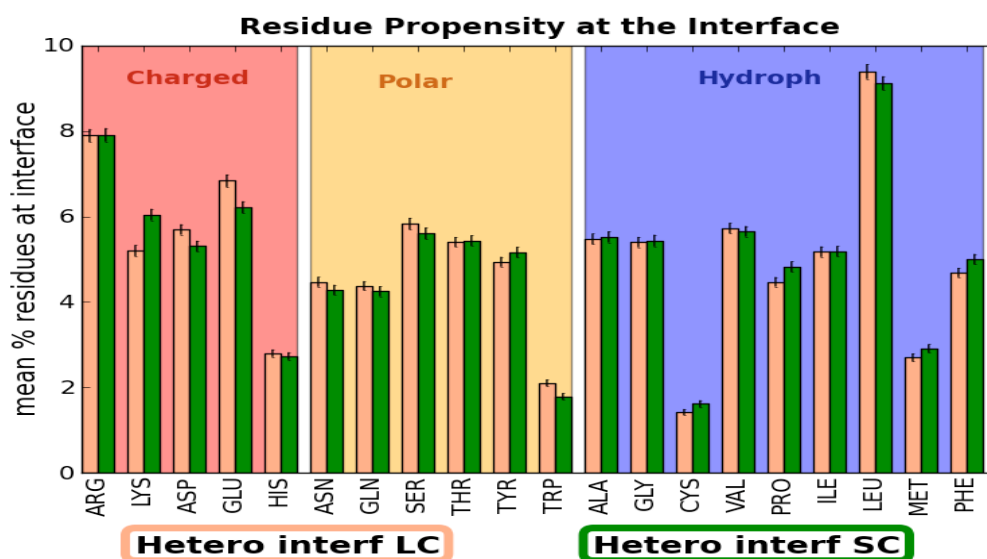


Figure 5.15. Comparison of residue propensities for long chain (LC, pale orange) and short chain (SC, green) of the hetero quaternary interfaces of the protein complexes. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

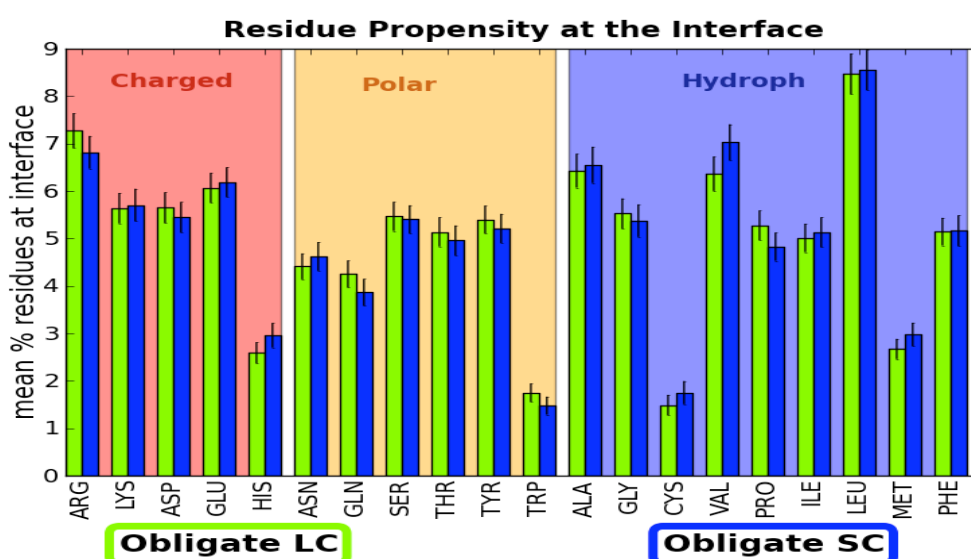


Figure 5.16. Comparison of residue propensities for long chain (LC, bright green) and short chain (SC, blue) of the obligate protein dimers. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

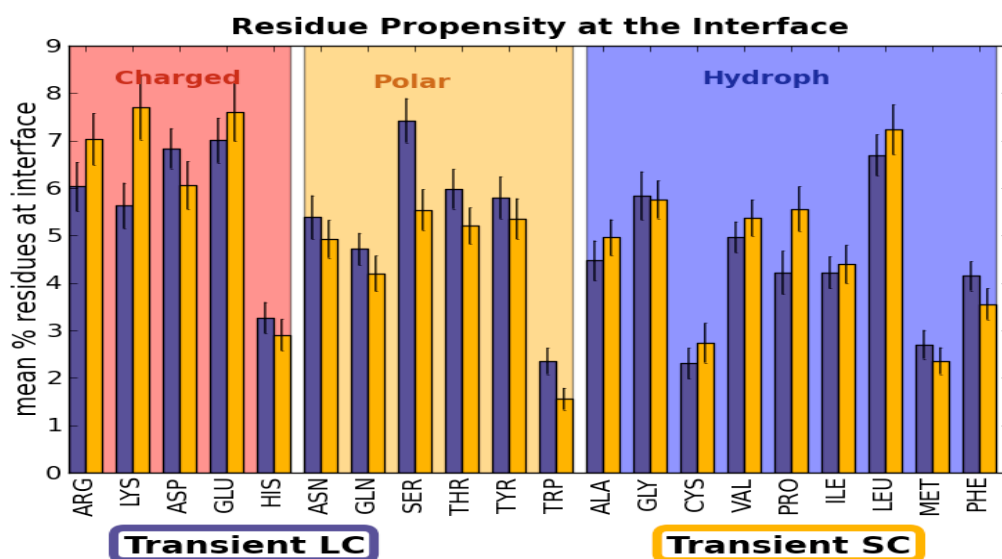


Figure 5.17. Comparison of residue propensities for long chain (LC, bright green) and short chain (SC, blue) of the transient protein dimers. Bar heights represent the mean of the percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

5.3.2.1 Charged, polar and hydrophobic

This classification allows comparison of binding sites with respect to the binding profile they present, in terms of polar and apolar interactions discussed in chapter 4. The classification considers the side chains only; therefore the charged residues are Arg, Lys, Asp, Glu and His, the polar residues are Asn, Gln, Ser, Thr, Tyr and Trp, and the hydrophobic residues are Ala, Gly, Cys, Val, Pro, Ile, Leu, Met and Phe.

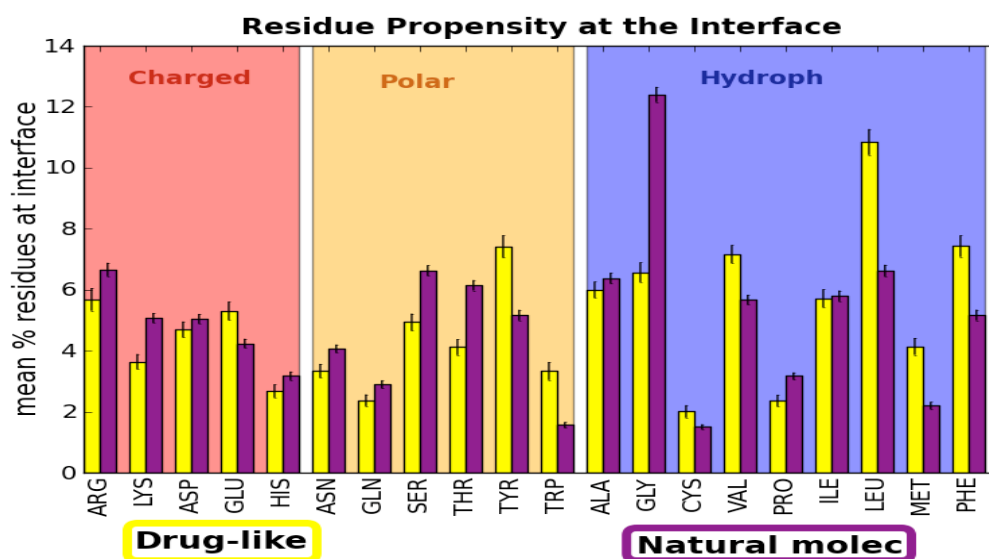


Figure 5.18. Comparison of residue propensities at the binding sites for drug-like (yellow) versus natural molecules (purple). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

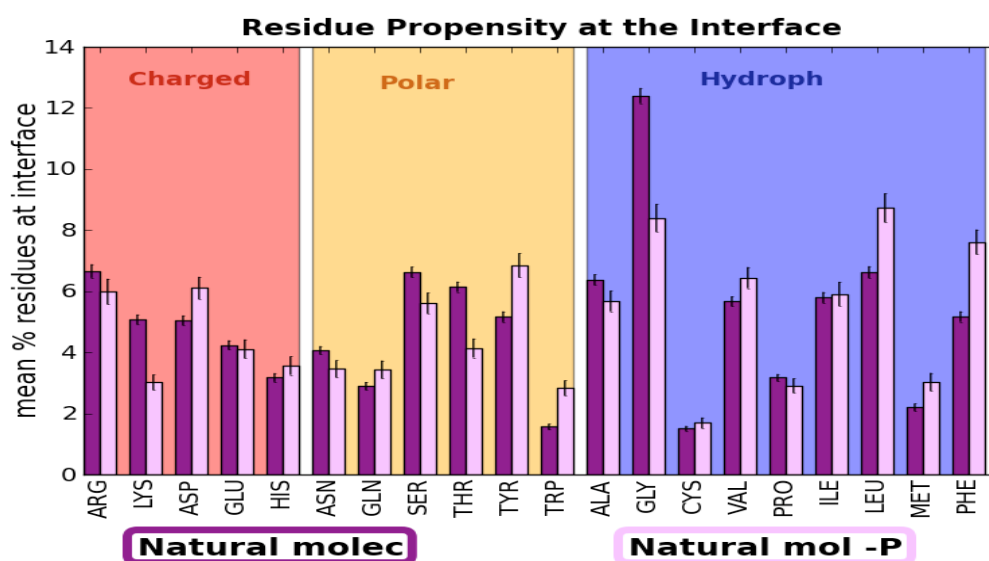


Figure 5.19. Comparison of residue propensities at the binding sites for natural molecules (purple) versus natural molecules without phosphorus (pale pink). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

Comparison of the residue compositions between drug-like and natural molecules highlights the more hydrophobic and aromatic character of the

drug-like binding sites (Figure 5.18), as found previously by Soga *et al.* (Soga *et al.* 2007). Natural binding sites have on average more non-aromatic and polar residues. Interestingly, natural molecules also interact more with glycines than the synthetic molecules. This may reflect the preference of these molecules for binding through main chain NH and CO. But also the presence of glycine-rich loops that often recognise phosphate groups via hydrogen bonds with the main chain nitrogen atoms (Gherardini *et al.* 2010). Indeed, comparison of the residue composition of the natural-binding sites versus the binding sites of natural molecules that do not contain phosphorus (Figure 5.19) shows an increase of glycine content for binding sites containing phosphate molecules. However, the proportion of glycines in natural binding sites is higher than in synthetic-molecule binding sites regardless of the phosphorus content, as mentioned before probably reflecting a tendency of these types of molecules to interact with main chain atoms. For classical drug targets, drug-like molecules typically bind to endogenous small molecule binding sites. The difference between natural and drug-like binding sites reflects either that the sites compared are very diverse between sets or that drug-like molecules avoid the regions where the natural molecules bind, for example the phosphate-binding region of ATP in protein kinases.

In comparison with protein-protein interfaces, drug-like molecules also bind to more hydrophobic sites, whereas the content of charged residues is greater in protein-complexes (Figure 5.20). Obligate protein complexes have on average more hydrophobic residues at the binding interface and less polar and charged residues compared to transient complexes (Figure 5.21).

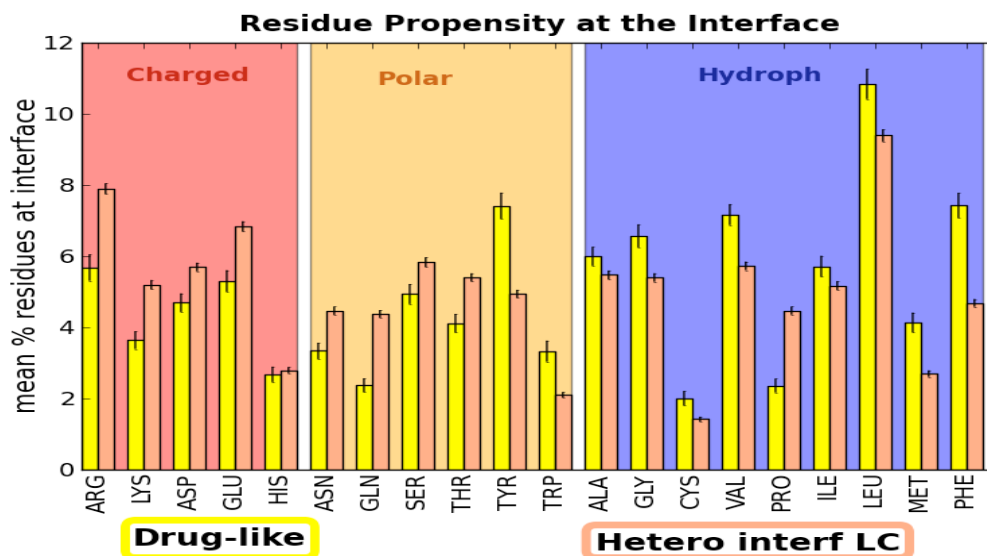


Figure 5.20. Comparison of residue propensities at the binding sites for drug-like (yellow) versus protein-protein quaternary hetero interfaces (pale pink). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

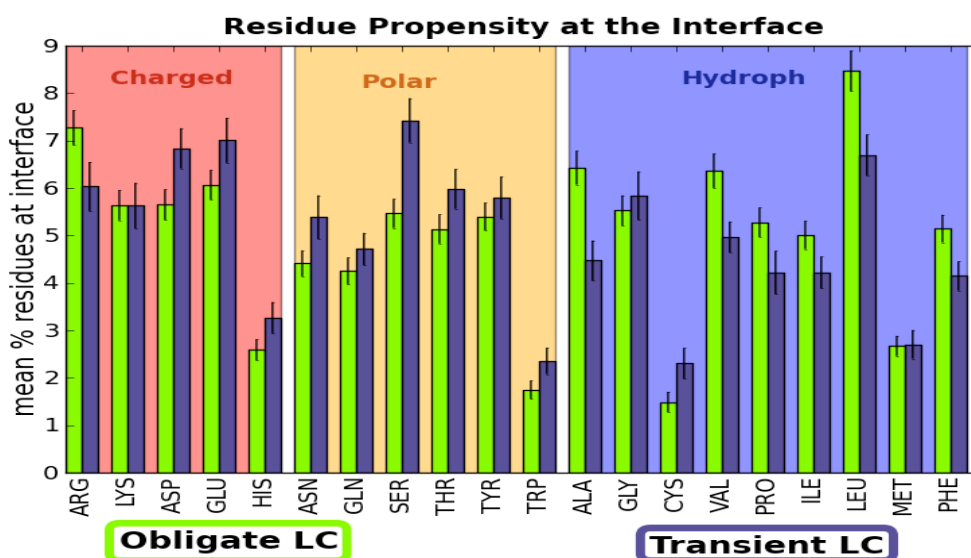


Figure 5.21. Comparison of residue propensities at the binding sites for obligate dimers (bright green) versus transient dimers (dark blue). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

Figure 5.22 shows the proportions of charged, polar and hydrophobic residues for all subsets. The proportion of charged residues in protein-protein complexes is significantly greater ($P < 0.05$) than that of the small molecule subsets. In contrast, the proportion of hydrophobic residues for small molecule inhibitor complexes with proteins is significantly greater than that for complexes with small peptides and proteins. Small peptide complexes are similar to transient dimers, having a greater proportion of polar residues than other subsets. In particular, transient dimers have more polar and fewer hydrophobic residues than obligate dimers.

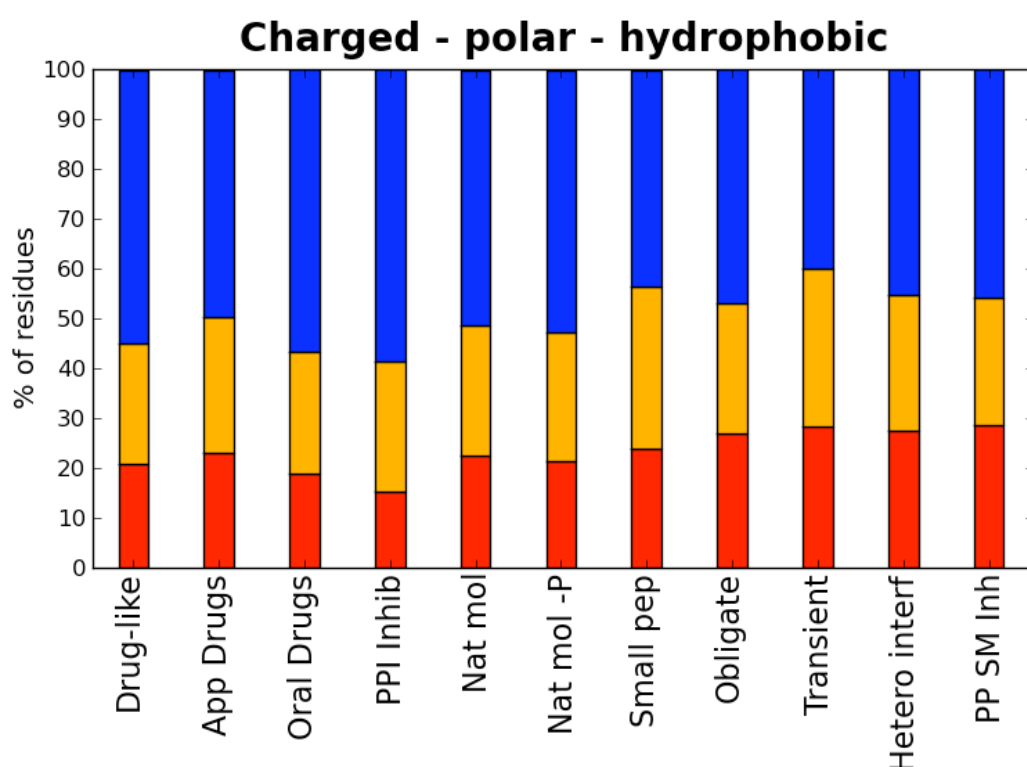


Figure 5.22. Average proportion of charged (red), polar (orange) and hydrophobic (blue) residues at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero- quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only the long chain is considered.

5.3.2.1.1 Protein-protein complexes inhibited by small molecules

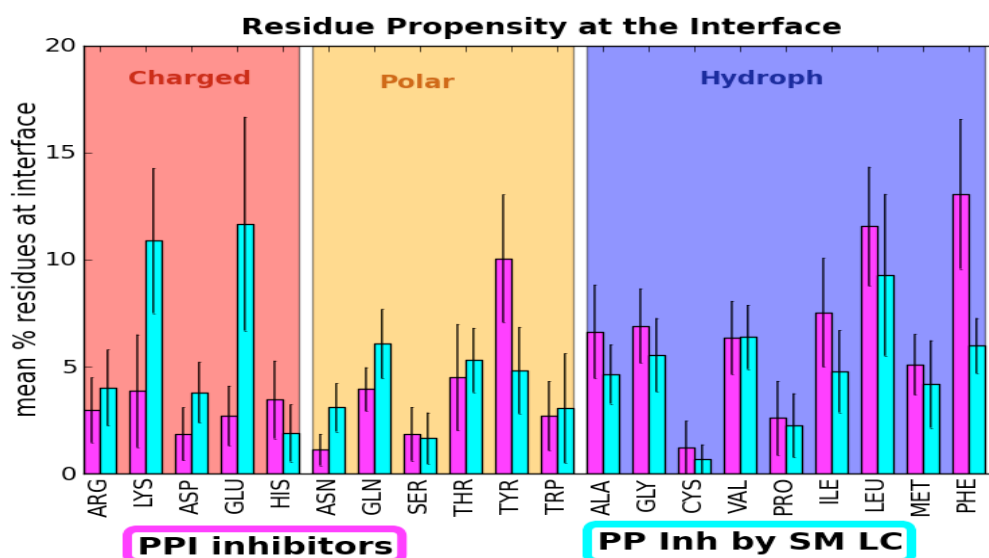


Figure 5.23. Comparison of residue propensities at the binding sites for small molecule protein-protein inhibitors (magenta) versus protein-protein complexes inhibited by them (cyan). Note these subsets are small (9 and 7 complexes respectively). Bar heights represent the mean percentage of each residue at the interface. Error bars denote the standard error of the mean. The background colour represents whether the residue is charged (red), polar (orange) or hydrophobic (blue).

Regarding protein-protein complexes inhibited by small molecules, Figure 5.23 shows the comparison of the residue propensities for the protein interfaces that have been independently structurally determined bound to both partners: the protein partner and the small molecule inhibitor. These cases are S100B, IL-2, MDM2, ZipA, XIAP, Bcl-XL, Bcl-2, and TNF alpha. There are nine distinct UniProt protein-small molecule complexes and seven protein-protein complexes. The high standard error bars denote the variability and the small size of the sets. However, it is clear from this comparison that the small molecules avoid contact with the available charged and polar residues in favour of interacting with the hydrophobic ones. Indeed, small molecules occupy only a portion of the protein-protein binding interface, and they tend to maximise the hydrophobic contacts rather than the polar ones. This may be a result of the small molecules binding at the hot spots of the interfaces, especially in the standard medicinal chemistry settings where the

pursuit of affinity is prioritised. However, these molecules seem to be missing the specific contacts that would confer them selectivity towards these interfaces. However, hydrogen bond matching at an open interface might require a degree of flexibility that it is harder to design and successfully achieve, and it could explain the low content of hydrogen bonds in the first successful small molecule inhibitors of protein-protein interactions. Figure 5.24 to Figure 5.30 show a graphical representation of these binding modes.

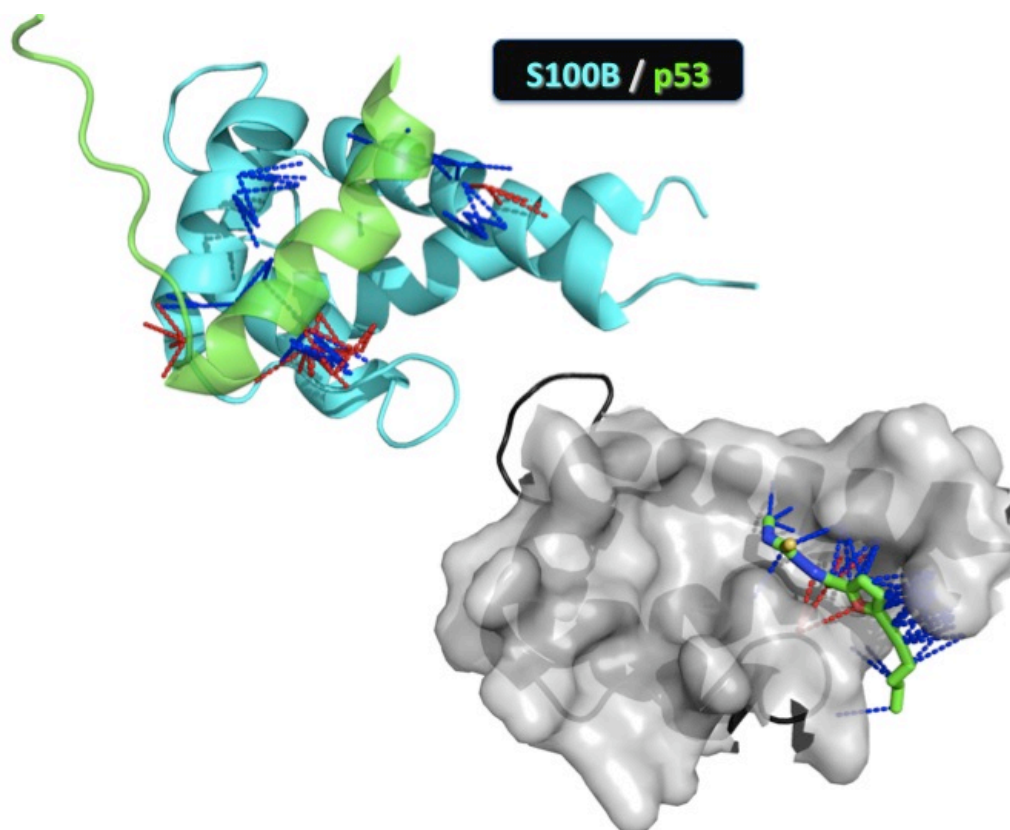


Figure 5.24. **S100B**. Upper left: 1DT7, S100B (cyan) with the C-terminal negative regulatory domain of p53 (green). Lower right: 3GK1, S100B (dark grey) with small molecule inhibitor (green). The surface covers the S100B residues that are within 4.5Å of p53. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.

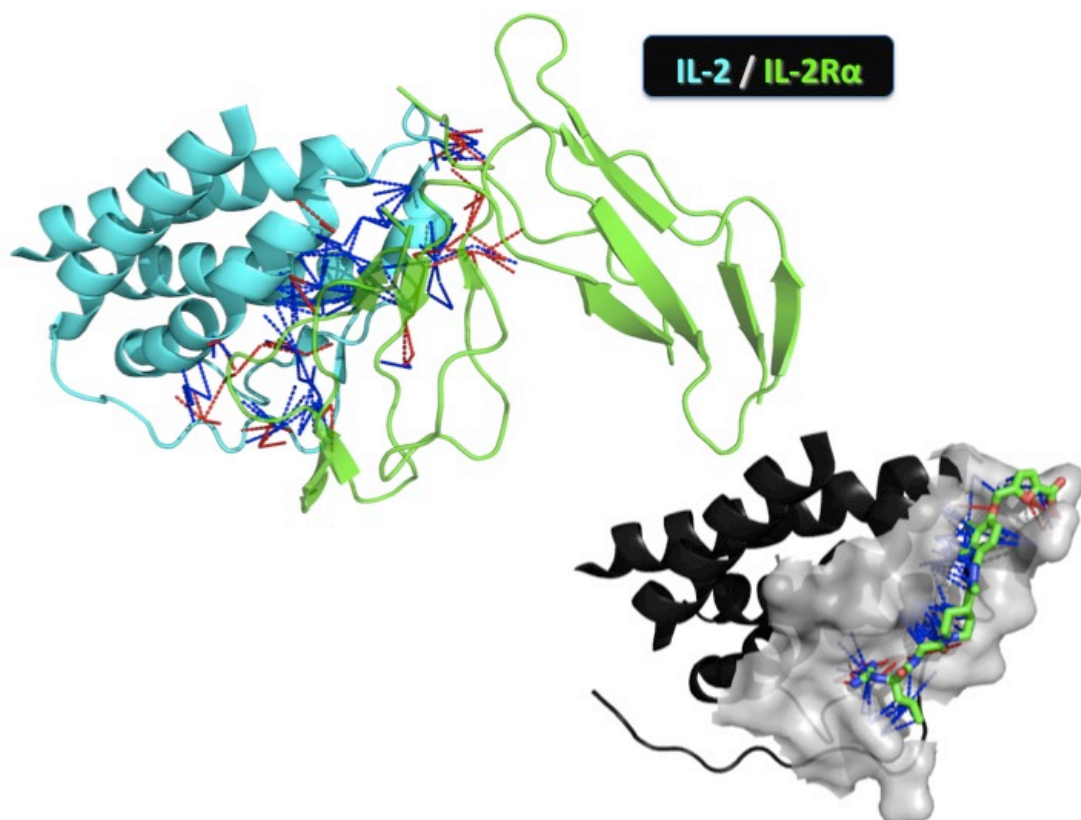


Figure 5.25. **IL-2**. Upper left: 1Z92, IL-2 (cyan) bound to IL-2R alpha subunit (green). Lower right: 1PY2, IL-2 (dark grey) with a Sunesis small molecule inhibitor (green). The surface covers the IL-2 residues that are within 4.5Å of the IL-2R α . For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.

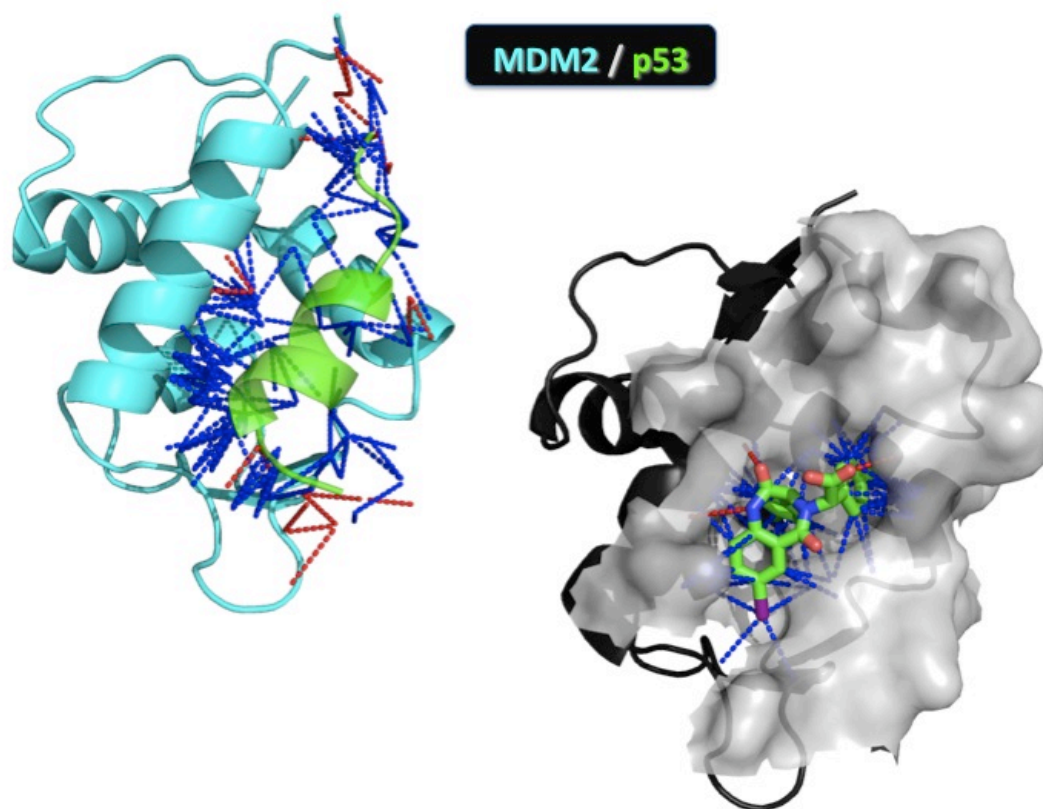


Figure 5.26. **MDM2**. Upper left: 1YCR, MDM2 (cyan) bound to the transactivation domain of p53 (green). Lower right: 1T4E, MDM2 (dark grey) with a benzodiazepine inhibitor (green). The surface covers the MDM2 residues that are within 4.5Å of the p53. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.

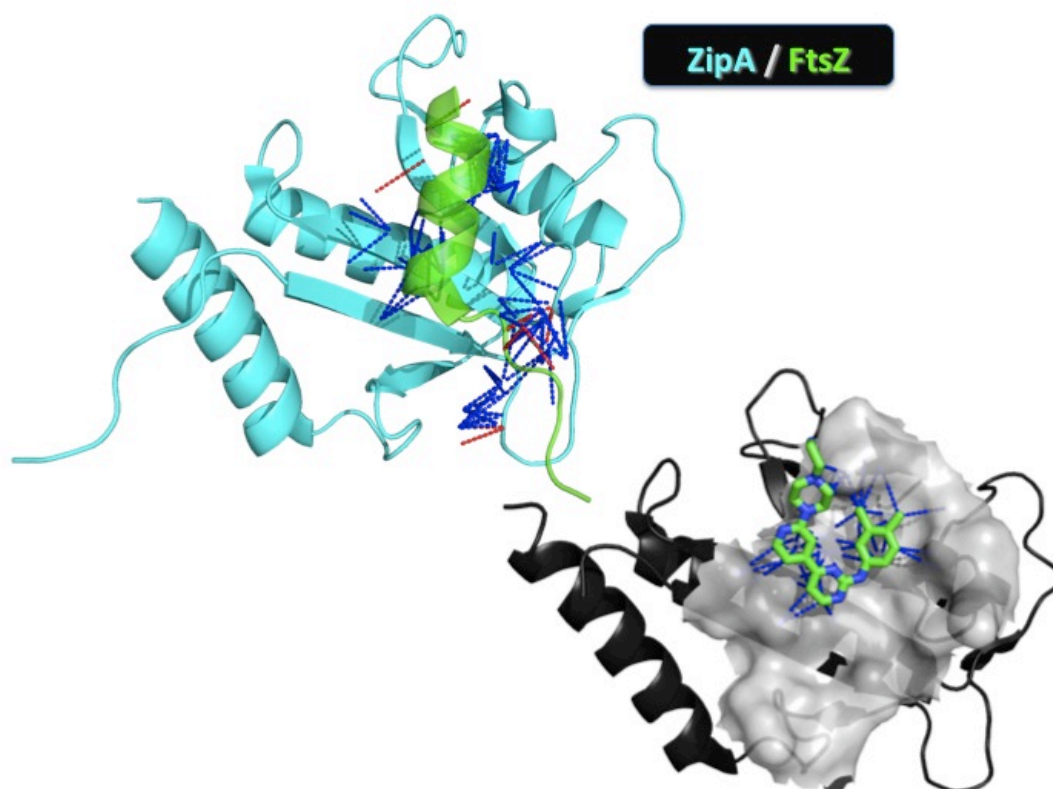


Figure 5.27. **ZipA**. Upper left: 1F47: ZipA (cyan) bound to a fragment of FtsZ (green). Lower right: 1Y2F: ZipA (dark grey) with an aminopyrimidine inhibitor (green). The surface covers the ZipA residues that within 4.5Å of the FtsZ. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note the small molecule does not engage a single polar contact.

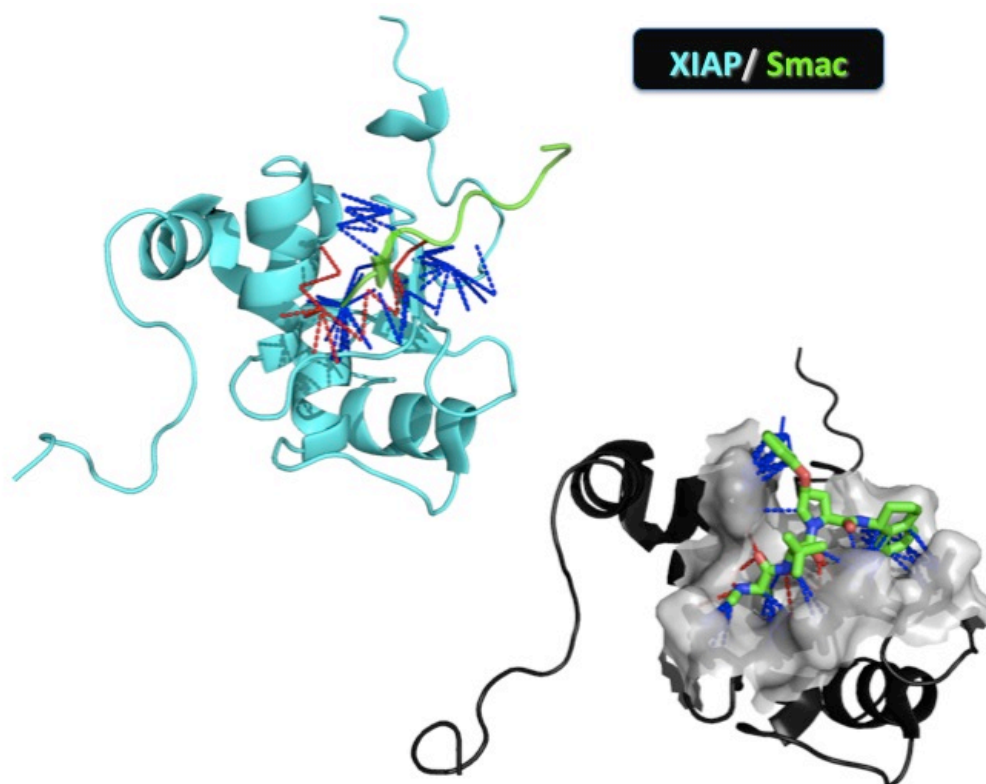


Figure 5.28. **XIAP**. Upper left: 1G3F, BIR3 domain of XIAP (cyan) bound to an active nine-residue peptide derived from Smac (green). Lower right: 1TFT, XIAP (dark grey) with a small molecule inhibitor (green). The surface covers the XIAP residues that within 4.5Å of the Smac fragment. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines.

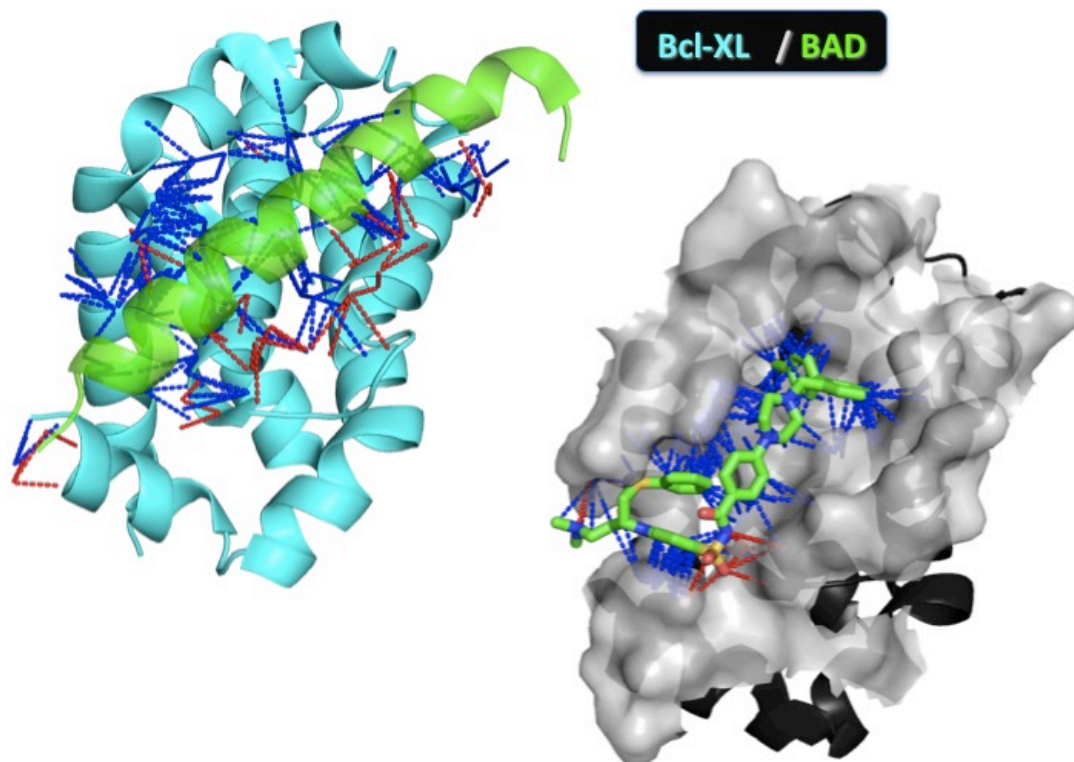


Figure 5.29. **Bcl-XL**. Upper left: 2BZW, Bcl-XL (cyan) bound to BAD (green). Lower right: 2YXJ, Bcl-XL (dark grey) with the Abbott compound ABT-737 (green). The surface covers the Bcl-XL residues that within 4.5Å of BAD. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note that the small molecule only engages polar contacts at the bottom of the picture and it is bound to Bcl-XL mainly though apolar contacts.

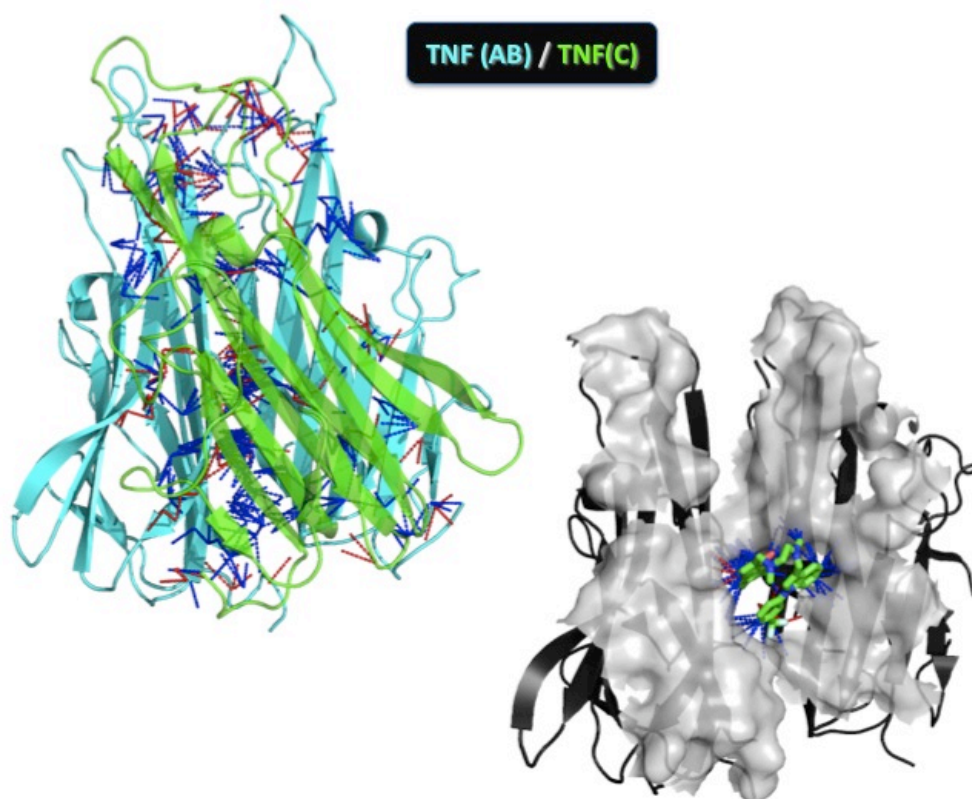


Figure 5.30. **TNF**. Upper left: 1TNF, TNF alpha trimer, two chains are coloured in cyan and the third in green. Lower right: 2AZ5, two chains of the TNF trimer (dark grey) bound to a small molecule (green) that accelerates subunit dissociation. The surface covers the residues in these chains that are within 4.5Å of the third chain. For both complexes polar contacts are red dotted lines and apolar are blue dotted lines. Note small molecule binds to an area where there are no interactions in the trimer.

5.3.2.2 Small, medium and bulky

Here, residues are grouped by the number of their heavy atoms. In this way, small residues (4-7 heavy atoms) are Ala, Cys, Gly, Pro, Thr, Val and Ser. Medium (8-10 heavy atoms) are Asn, Asp, Gln, Glu, Ile, Leu, Lys, Met and His. Bulky residues (11-14 heavy atoms) are Arg, Phe, Trp and Tyr. This classification gives a rough measure of the exposure of the main chain atoms. If a site is composed of many bulky side chains, in principle the main chain atoms will be more occluded from interacting with the binding partner. In this respect, natural molecules have a significantly higher proportion of small side chains and a lower percentage of bulky residues. Figure 5.31 shows the proportion of these residue types for all subsets. Small molecule

protein-protein inhibitors have a significantly greater proportion of bulky residues in comparison with natural molecules and protein complexes, but the difference with the other sets of synthetic molecules (drugs and drug-like) is not significant.

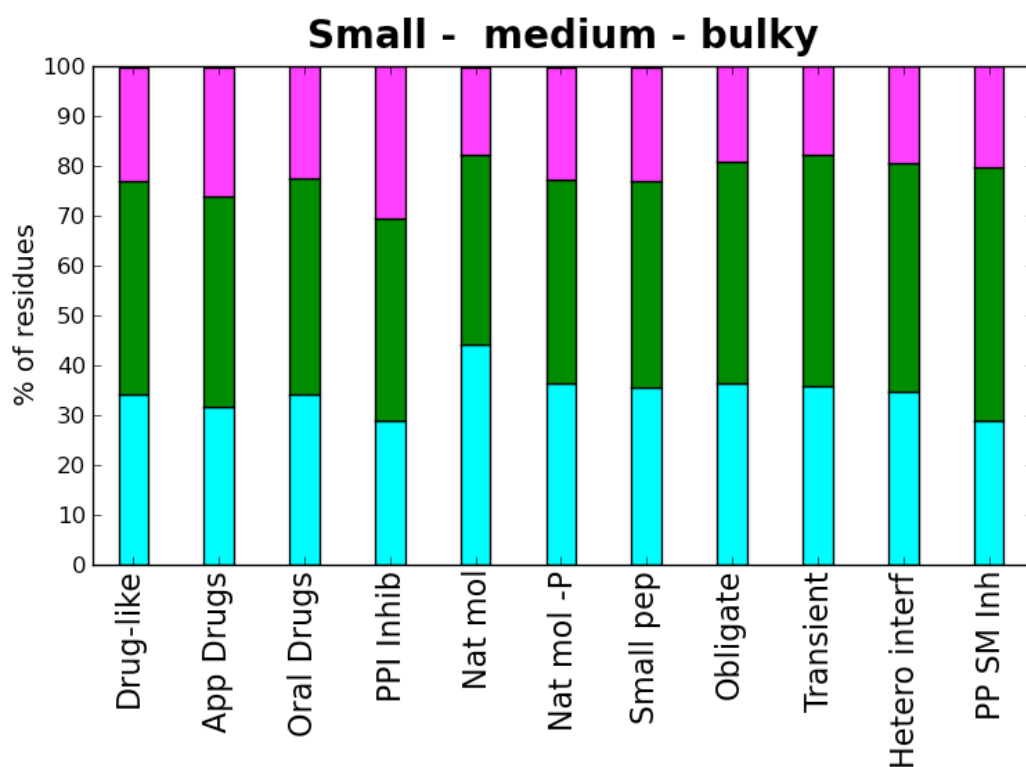


Figure 5.31. Average proportions of small (cyan), medium (green) and bulky (magenta) residues at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only the long chain is considered.

5.3.2.3 Constrained, free, rigid, medium, flexible and aromatic

In order to have an estimate of the “softness” or adaptability of the binding interfaces, residues are classified by the number of rotatable bonds in the side chain. Proline and Glycine are separated into constrained and free groups respectively. Rigid residues are those with none or one rotatable bond

in the side chain, they are Ala, Cys, Ser, Thr and Val. Medium flexible residues have 2-3 rotatable bonds and small functional groups and are Asn, Asp, Gln, Glu, Ile and Leu. Flexible residues have 4-5 rotatable bonds; they are Arg, Lys and Met. Finally all aromatic residues have two rotatable bonds and an aromatic ring; they are His, Phe, Trp and Tyr. Figure 5.32 shows the proportion of these residue types across all subsets. Proline content is significantly greater at protein-protein interfaces, especially for obligate dimers, and natural molecules. Natural molecules have a greater proportion of glycines at their binding interfaces than other molecular subsets. Aromatic content is greater for synthetic molecules, small peptides and natural molecules without phosphorus, compared to protein complexes and natural molecules. Protein-protein complexes have a significantly greater proportion of flexible residues than small molecule interfaces, suggesting these complexes might have a greater ability to adapt to the binding partner than the preformed pockets where small molecules usually bind.

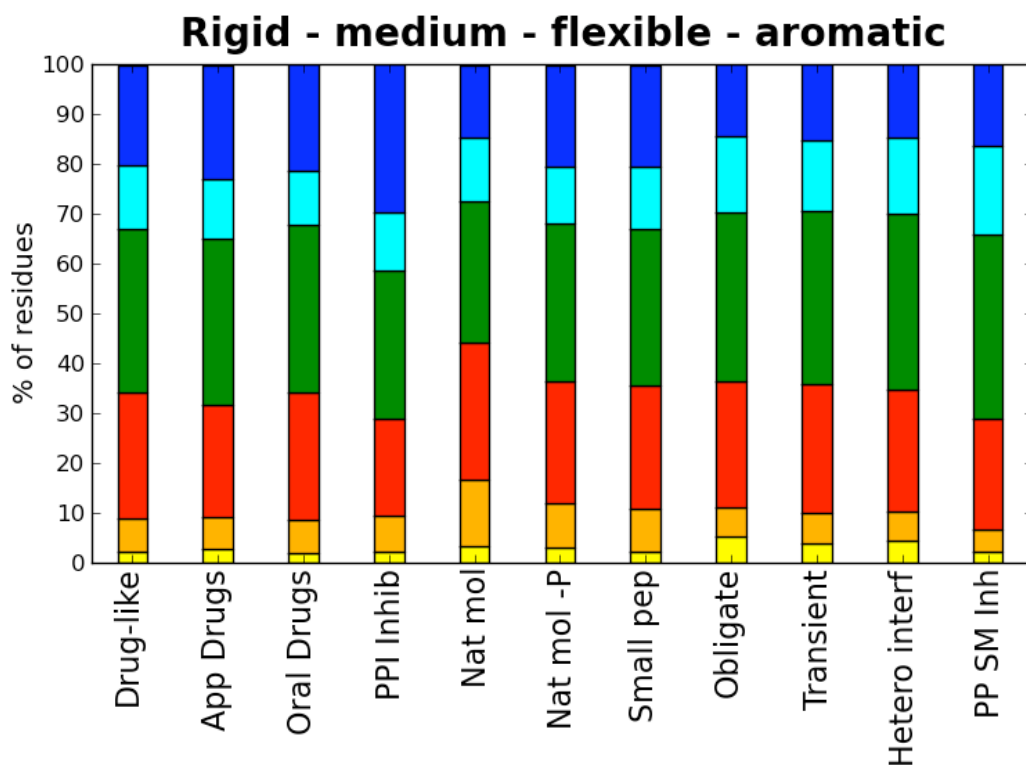


Figure 5.32. Average of the proportion of constrained (yellow), free (orange), rigid (red), medium (green), flexible (cyan) and aromatic (blue) at the interfaces for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, only long chain is considered.

5.3.3 Proportion of main chain atoms at the binding interfaces

In order to have an indication of the flexibility and robustness to mutation of the protein side, the main chain atoms have been counted in the binding sites of each subset. Figure 5.33 shows the average proportion of main chain atoms over the total number of atoms at the binding interface (left panels) and the average of the proportion of main chain atoms over total number of atoms that are matched at the interface (right panels). For protein-small molecule complexes, both levels (UniProt and SCOP) of protein redundancy are assessed (upper versus lower panels in Figure 5.33), although there are changes in the absolute average numbers, the trends for

each subset are maintained when the numbers are filtered by SCOP families. In the case of protein-protein subsets, main chain atoms have been counted for both chains, labelled as long and short chain depending on the number of residues in the chain. No statistical difference has been found between long and short chain for any of the subsets and the ratios studied.

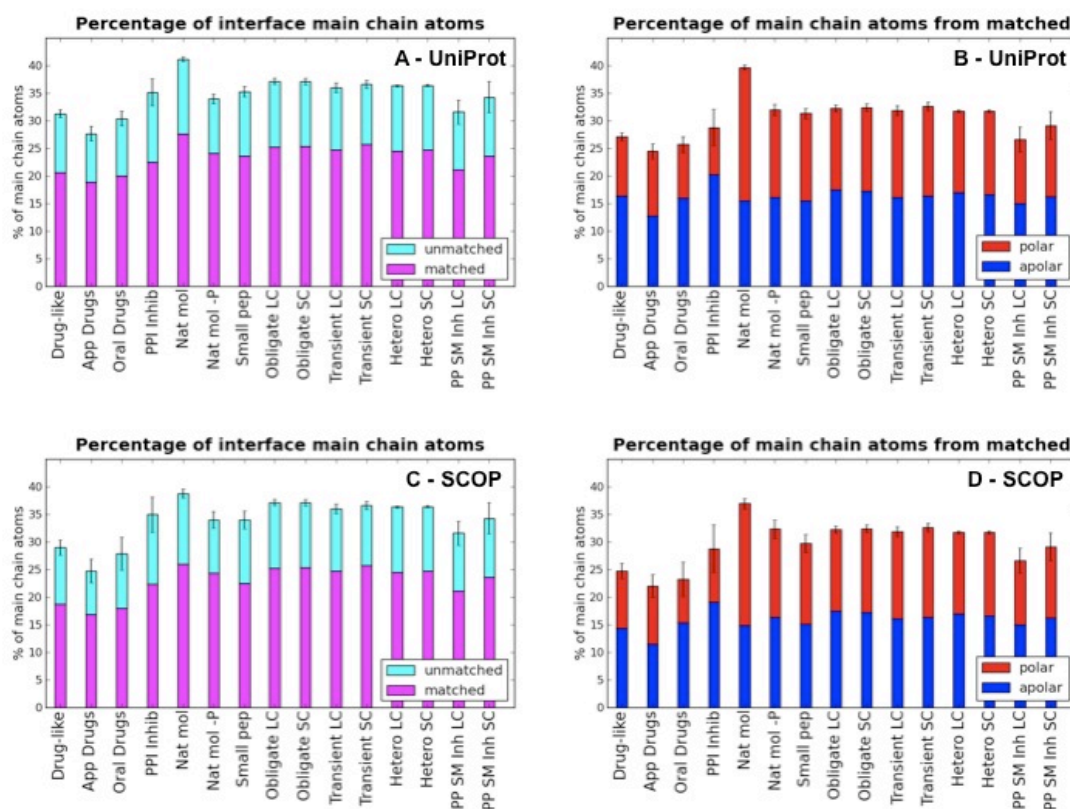


Figure 5.33. Average of the percentage of main chain atoms for each molecular subset at the UniProt level: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. A and C: percentage of main chain atoms at the interface (defined as atoms within 4.5Å of the binding partner) colour coded by the proportion that are matched (magenta) or unmatched (cyan). B and D: percentage of main chain atoms from the matched atoms colour coded by polar (red) and apolar (blue). Both levels of redundancy are plotted, A and B: protein-small molecule complexes with distinct UniProt identifiers. C and D: proteins-small molecule complexes belonging with distinct SCOP families.

Figure 5.33 shows that on average (and for distinct UniProt proteins) drugs and drug-like molecules have a significantly ($P < 0.05$) smaller

proportion of main chain atoms in the active sites (28-32%) than natural molecules that have 44% of main chain atoms at the binding interface and small peptides and proteins that have 35-38%. Amongst the protein-protein complexes, there is no significant difference between sets for any of the ratios considered, with the remarkable exception of the long chain of the protein-protein complexes inhibited by small molecules. For these chains, all the proportions of main chain atoms considered (Figure 5.33 and Figure 5.34) are significantly ($P < 0.05$) lower than the rest of protein interfaces. Although there are only 15 complexes in this set, this is an interesting result implying that a higher content of side chain atoms at an interface makes it more amenable to bind small molecules, arguably to facilitate site adaptability. This result is in consonance with previous findings for these interfaces (Fuller *et al.* 2009; Bourgeas *et al.* 2010) that highlighted the accommodation of the small molecules at the interface by side chain rearrangement. The absolute proportion of main chain atoms of the 20 standard amino acids is 48% and rises to 52% if one takes into account the natural abundance (Voet *et al.* 1992) of each residue. In fact, main chain atoms are more common in the protein core and are involved in secondary structure interactions (Chothia 1976). From this main chain atom composition, Figure 5.33 also shows main chain atoms as a proportion of the total number of atoms matched at the binding site. The same trends are maintained; drugs and drug-like molecules have significantly ($P < 0.05$) fewer main chain atoms (24-27%) engaged in successful interactions (as defined in chapter 4, section 4.2.2) than natural molecules (40%) and small peptides and protein complexes (31-33%).

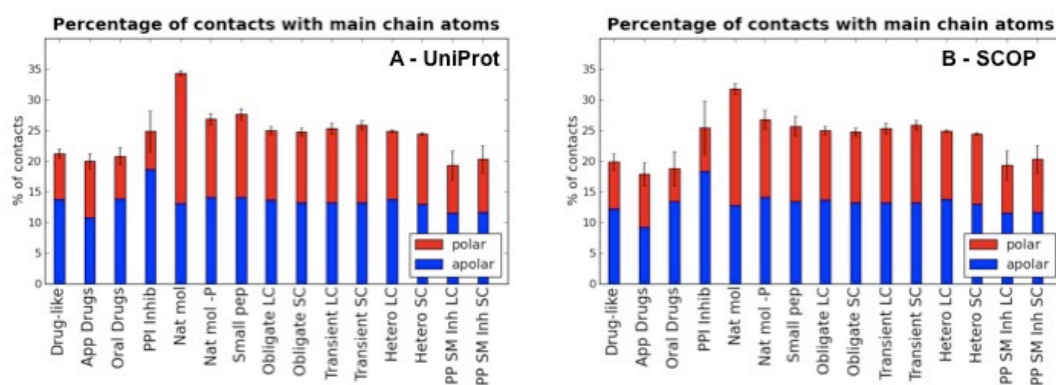


Figure 5.34. Average percentage of contacts involving main chain atoms for each molecular subset for both levels of protein redundancy: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. Colour coded by polar (red) and apolar (blue).

Figure 5.34 shows that natural molecules also have a higher proportion of contacts involving main chain atoms. Indeed, 34% and 32% at UniProt and SCOP family redundancy level respectively, of the contacts made by these molecules interact with protein main chain atoms. Furthermore, more than half of these main chain atoms are polar atoms. This trend may be a consequence of selective pressure in evolution through non-synonymous single nucleotide polymorphisms; a main chain interaction would be more robust to mutation of the amino acid. This may be more crucial in the small binding sites of endogenous ligands than in the large protein complex interfaces. In the latter, compensating mutations can be accepted over time and the proportion of main chain to side chain interactions is much lower. Furthermore, natural molecules tend to be more flexible and able to optimise interactions. However, natural molecules without phosphorus present a lower proportion of contacts involving main chain atoms and are similar to proteins and small peptides in this respect. To investigate this further, Figure 5.35 shows the level of small molecule redundancy of the natural molecule set.

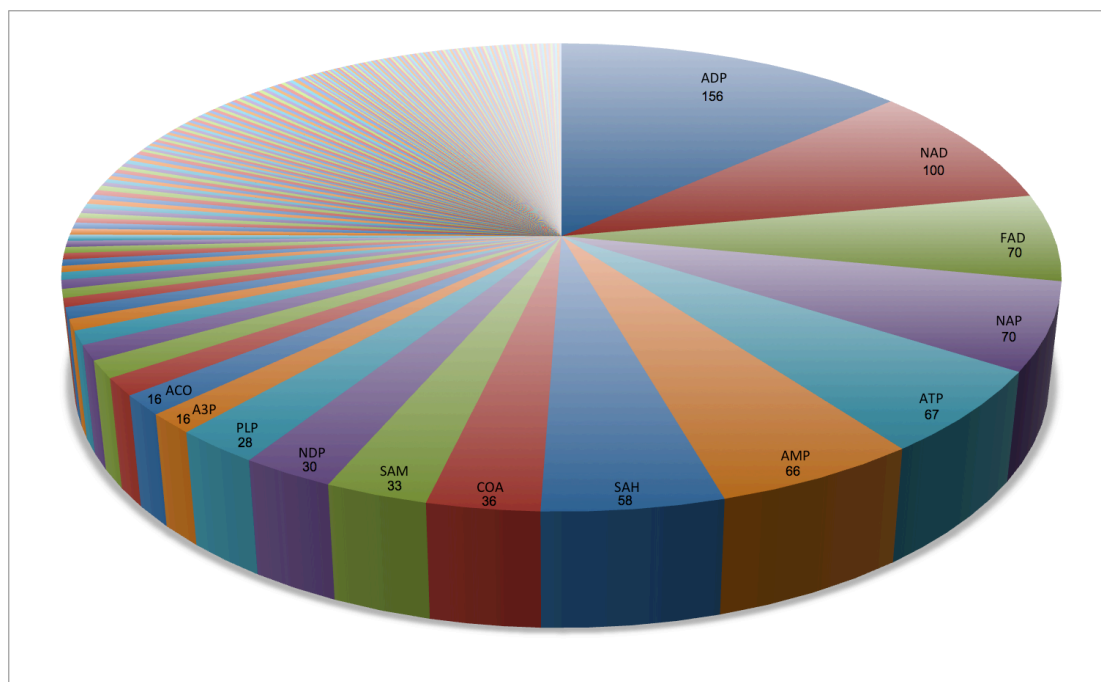


Figure 5.35. Distribution of the natural small molecule subset (filtered for protein redundancy by distinct UniProt) in terms of entries per chemical structure of the small molecule bound to protein. Only higher frequency entries are labelled for clarity. Note that more than half of the subset is composed of the complexes with seven different molecules: ADP, NAD, FAD, NAP, ATP, AMP and SAH.

The natural molecule set analysed here is composed of 1159 different proteins interacting with 216 small natural molecules. As shown in Figure 5.35, more than half of these complexes are formed by nucleotides with saccharides and phosphates, which in turn are the complexes using more main chain atoms.

HetID	Num atoms SM	Num SCOP fa	% mc contacts	% mc atoms	Ratio
ADP	27	41	40%	46%	0.52
SAH	26	23	40%	47%	0.36
ATP	31	18	31%	38%	0.36
SAM	27	18	39%	46%	0.37
AMP	23	17	40%	45%	0.48
NAD	27	15	39%	46%	0.43
FAD	53	11	37%	45%	0.38
COA	48	9	32%	40%	0.40
NAP	48	7	47%	54%	0.49

Table 5.1. The nine most promiscuous small molecules. They all belong to the natural molecule set. Columns in the table are from left to right: HetID is the PDB residue identifier, Num atoms SM is the number of atoms of the small molecule ligand, Num SCOP fa is the number of different SCOP families the small molecule binds to, %mc contacts is the average of the percentage of contacts by protein main chain atoms across all SCOP families bound for a particular small molecule, %mc atoms is the average of the percentage of protein main chain atoms at the binding interface across all SCOP families bound, Ratio is the average of ratio of polar contacts by sum of contacts.

Table 5.1 shows the average proportion of contacts involving main chain atoms, as well as binding site main chain atom content for the more frequent natural molecules binding to different SCOP families. Only distinct SCOP families are considered here in order to avoid bias by protein families, like protein kinases for instance. These molecules bind to a wide range of proteins and SCOP families and they have the highest main chain contact ratio. Although they also present the highest polar contact ratio, these molecules are multipurpose and not selective for a single protein. However, this result can also be interpreted from the protein side. These protein molecules have evolved to bind to the same nucleotide even though they have different folds. In this respect, the proportion of main chain atoms in the active site is one of the factors that can assist in identifying promiscuous

drug-binding sites in therapeutic targets. Figure 5.36 shows that promiscuous binders in the PDB present a high content of main chain atom contacts.

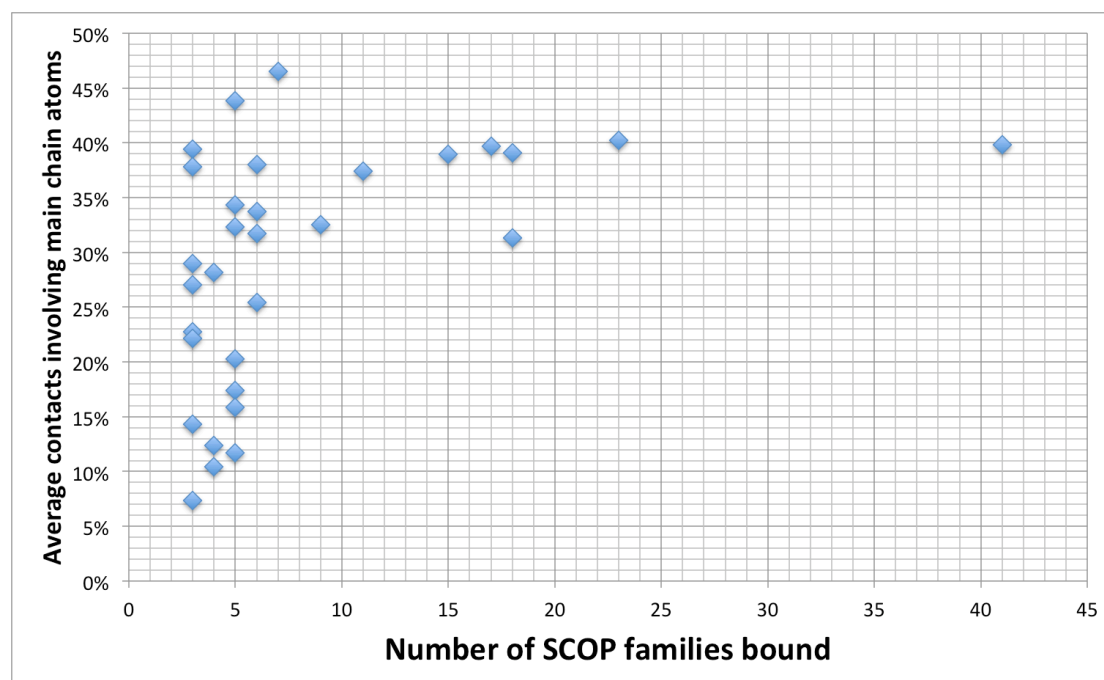


Figure 5.36. Scatter plot of the number of different SCOP families bound to the same small molecule versus the average of contacts involving main chain atoms that these molecules are engaging.

5.3.4 Proportion of polar atoms at the binding interface

This section analyses the polar topology of the binding interfaces for each molecular subset. As before, for protein-small molecule complexes, both levels (UniProt and SCOP) of protein redundancy are assessed (upper versus lower panels Figure 5.37). Although there are changes in the absolute average numbers, the trends for each subset are maintained when bias due to over-representation of certain SCOP families is removed. For the protein-protein complexes, analysis of the polar atom content has been carried out for both long and short chains.

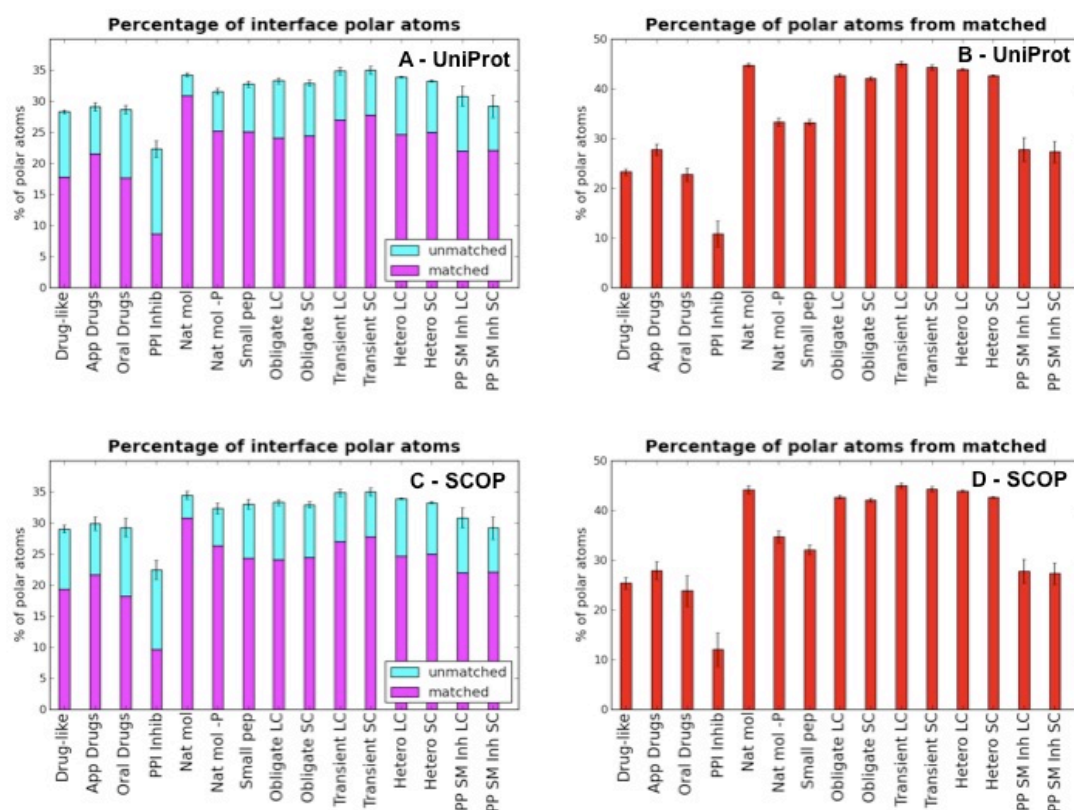


Figure 5.37. Average percentage of protein polar atoms for each molecular subset: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For the PP complexes, both long chain (LC) and short chain (SC) are plotted. Error bars denote the standard error of the mean. A and C: percentage of protein polar atoms at the interface (defined as atoms within 4.5Å of the binding partner) colour coded by the proportion that are matched (magenta) or unmatched (cyan). B and D: percentage of protein polar atoms from the total atoms that are matched. Both levels of redundancy are plotted, A and B: protein-small molecule complexes with distinct UniProt identifiers. C and D: proteins-small molecule complexes belonging with distinct SCOP families.

Figure 5.37 shows that the proportion of polar atoms at the interface of small molecule protein-protein inhibitors is significantly ($P < 0.05$) lower (22% for both UniProt and SCOP families) than all the other sets of molecules. These binding sites represent the highest proportion of unmatched polar atoms, and the lowest proportion of polar atoms from the matched protein atoms at the interface. This result has been discussed in chapter 4 (section 4.3.5) from the ligand viewpoint. Although only protein atoms are considered here, the interfaces are defined by the binding partner, i.e. only protein

atoms proximal to the ligand are taken into account. On the other hand, interfaces of protein-protein complexes inhibited by small molecules are significantly ($P < 0.05$) less polar (30%, Figure 5.37 A the two bars on far right) than the other protein-protein interfaces (33-35%) and similar to the drug interfaces (29%). Thus, this may be the reason for the lower proportion of polar atoms at the binding sites of protein-protein inhibitors. However, this result shows that small molecules binding at the protein-protein interfaces target the most hydrophobic patches on the surface and do not take advantage of the possibility of engaging available specific contacts.

The absolute proportion of polar atoms in the 20 standard amino acids is 35% and rises to 37% if one takes into account the natural abundance (Voet *et al.* 1992) of each residue. Binding sites for natural molecules and transient protein complexes are in this range of polar atoms (34 and 35% respectively). Indeed, binding interfaces for transient complexes are slightly more polar than obligate interfaces, (2% more, $P < 0.05$) as reported by other studies comparing obligate with non-obligate complexes (Nooren *et al.* 2003). With respect to drug-like and drug-binding sites, the proportion of polar atoms is significantly lower (28-29%) than natural molecules. This result corroborates the use of hydrophobicity scores to predict druggability of binding sites, however polar interactions in hydrophobic environments are stronger and cannot be dismissed in the assessment of druggability (Schmidtke *et al.* 2010). Furthermore, the proportion of unmatched polar atoms at the binding sites of drugs suggests that drug-like molecules could, in principle, engage more specific interactions as discussed in chapter 4.

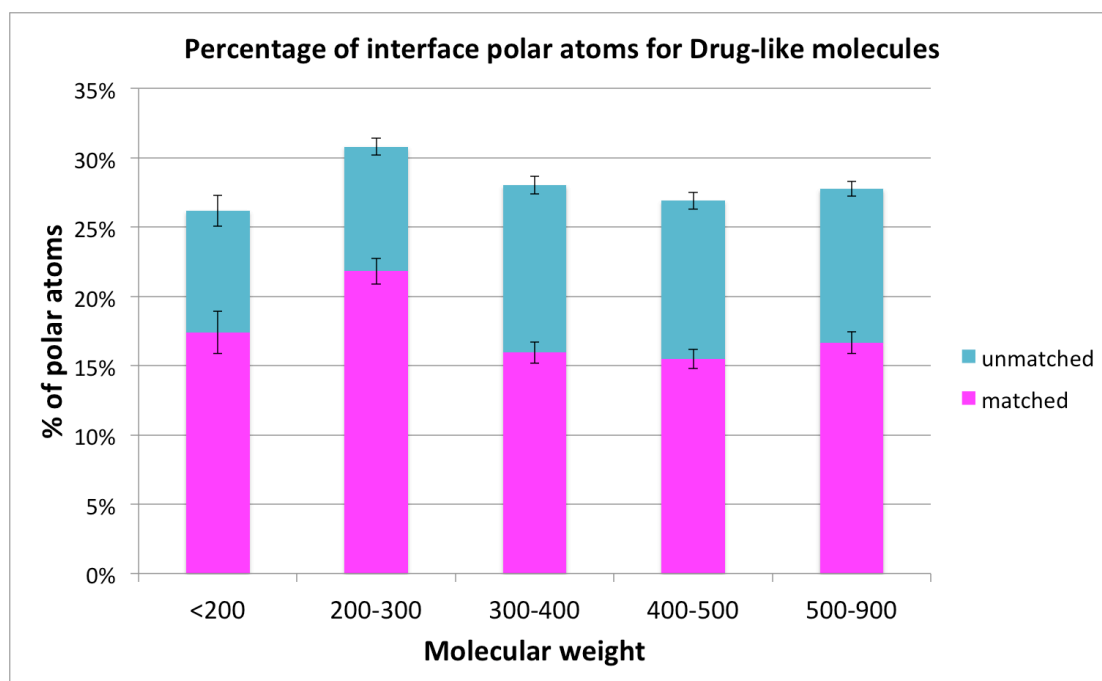


Figure 5.38. Distribution of the average of protein polar atoms at the binding interface for drug-like molecules at the UniProt level by molecular weight of the small molecule. The proportion of polar atoms is colour coded if they are engaged in successful interactions with the ligand (magenta) or are unmatched (cyan). Error bars denote the standard error of the mean.

Figure 5.38 shows the distribution of polar atoms at the binding interfaces of drug-like molecules. As described in 5.2.2, the protein atoms considered in the binding interfaces are defined by a distance cut-off from the ligand. In this respect, small fragments of molecular weight between 200-300Da bind to regions that are significantly more polar than the binding regions of bigger molecules. This result is consistent with the results described in chapter 4, where small fragments engaged more polar contacts than bigger molecules.

5.3.5 Depth of protein atoms at the binding interface

Using the ghecom program (Kawabata 2010), Rinaccess is calculated for all protein atoms at the interfaces, i.e. within 4.5Å of the binding partner. Rinaccess is a measure of the depth of the considered atom with respect to the protein surface. See Methods for details. For protein complexes, only the

longest chain is considered here as discussed before. The special cases where a large protein interacts with a shorter adaptable chain will be discussed separately.

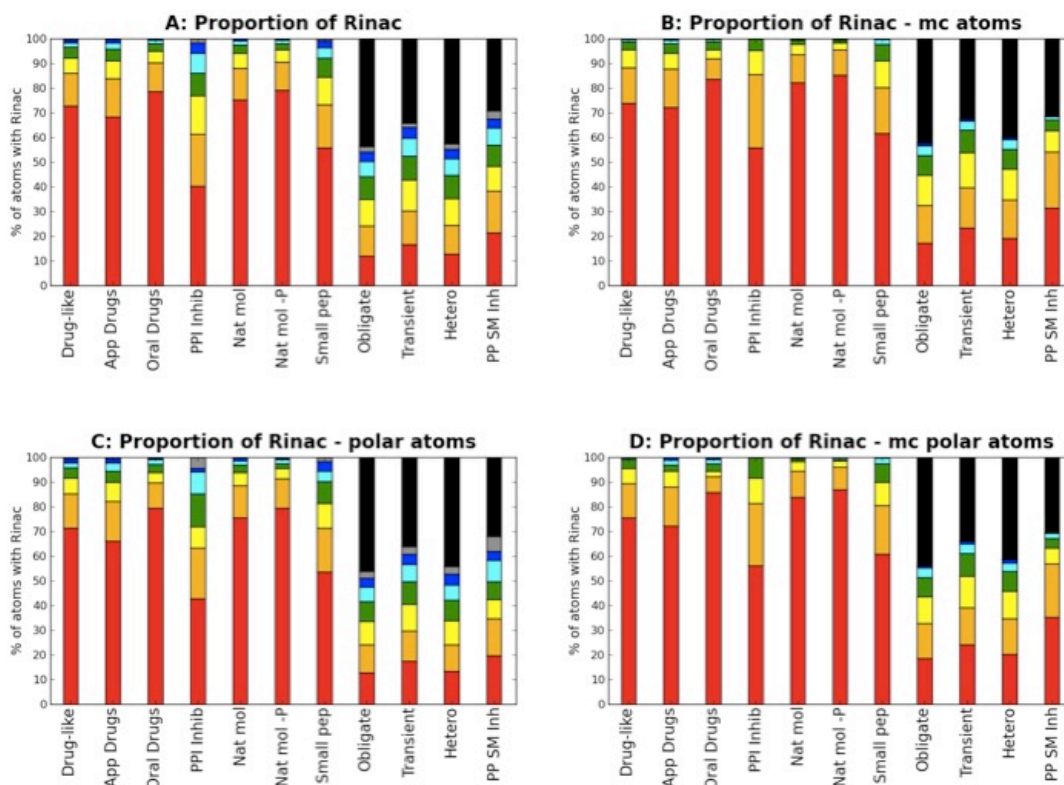


Figure 5.39. Proportion of Rinaccess values for the atoms at the interface for each molecule set at the SCOP family redundancy level. The colour in the bars denotes the Rinaccess values: red ($<2\text{\AA}$), orange ($2-3\text{\AA}$), yellow ($3-4\text{\AA}$), green ($4-5\text{\AA}$), cyan ($5-6\text{\AA}$), blue ($6-7\text{\AA}$), grey ($7-10\text{\AA}$) and black ($> 10\text{\AA}$). A: for all atoms at the interface, B: for main chain atoms at the interface, C: for polar atoms at the interface and D: for polar main chain atoms at the interface. For protein-protein complexes, only the longest chain is considered.

As expected, Figure 5.39 (A) shows that the small molecule sets have defined pockets, as shown by the higher proportion of small values of Rinaccess in comparison with protein-protein complexes. Within the small molecule subsets, there is significant difference in the average of Rinaccess for the small peptides and small molecule protein-protein inhibitors, which present less deep pockets than drugs, drug-like and natural molecules. Regarding protein-protein complexes, transient dimers have on average

deeper cavities than the obligate dimers, as shown by the significant difference in median for Rinaccess values. Interestingly, protein complexes that are inhibited by small molecules also have deeper cavities than the obligate and quaternary interfaces but cannot be distinguished from the transient subset. Figure 5.39 (B) shows that main chain atoms are on average deeper than the side chain atoms as the proportion of smaller Rinaccess values is bigger for these atoms. Polar atoms seem not to have a preferred position within the pockets, as Figure 5.39 (C) shows. The proportion of Rinaccess values for polar atoms in comparison with all atoms at the interface (Figure 5.39 (A)) does not change significantly.

5.3.5.1 Depth of the protein atoms at the interface versus chain length

Results in chapter 2 highlighted the success in finding small molecules binding to the protein-protein interfaces where one of the partners in the complex is a small peptide motif that probably undergoes a disorder-order transition upon binding to a globular domain. Indeed, the majority of protein complexes that have been successfully inhibited by small molecules are in this category. In order to quickly discriminate between the content of protein complexes studied here, in terms of relative protein sizes, the ratio of the length of the short chain to the long chain was calculated. Figure 5.40 and Table 5.2 show the distribution of this ratio.

	N complexes	Ratio (SC/LC)	% of R < 0.5
Obligate	161 (67 homo)	0.76	25%
Transient	154 (all hetero)	0.49	58%
Hetero interfaces	2,271	0.57	45%
Homo interfaces	12,034	0.98	1%
PP inh by SM	15 (1 homo)	0.39	67%

Table 5.2. Size differences between long chain (LC) and short chain (SC) for each subset of protein-protein complexes: Obligate dimers, Transient dimers, Hetero and Homo quaternary interfaces and protein-protein (PP) complexes inhibited by small molecules (SM). 'Ratio (SC/LC)' is the average of the ratio between the lengths of long and short chain. '% of R < 0.5' is the percentage of complexes where the short chain is smaller than half the long chain. See Figure 5.40 for the distribution of these ratios.

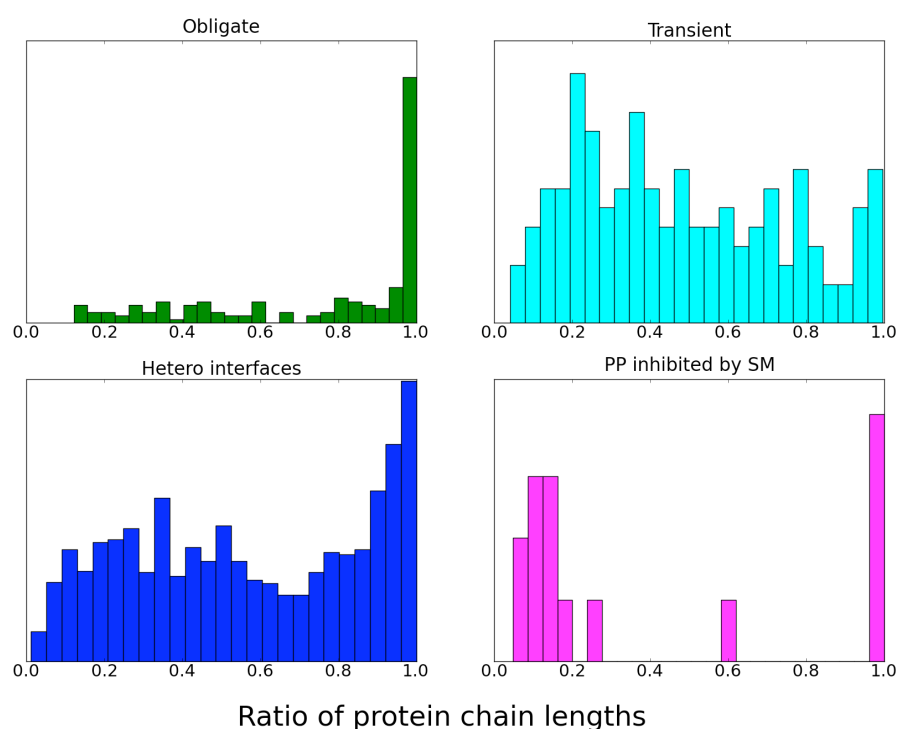


Figure 5.40. Normalised distribution of the ratio between the lengths of short and long chain for the protein-protein complexes subsets: Obligate dimers, Transient dimers, Hetero quaternary interfaces and protein-protein (PP) complexes inhibited by small molecules (SM). Homo quaternary interfaces are not plotted, as they have virtually no difference in chain length, see Table 5.2.

The biggest set of protein-protein interfaces, the homo quaternary interfaces, is composed of same length protein dimers. Only 1% of these complexes have significant difference in chain length, as this implies one of the partners has been truncated. The hetero quaternary interfaces subset presents a spread distribution of relative sizes, with the complexes of similar chain length being more common. The subset of obligate dimers (42% of which are homo dimers) is mainly composed by similar chain length complexes, although there are also cases of small peptides binding to bigger proteins. The Transient dimers (all of them are hetero dimers) have a spread distribution of relative size for the binding partners, skewed towards smaller ratios; almost 60% of the Transient complexes are composed by one partner that is, at least, double the length of the other. Here, the interest is to explore the subset of complexes where a large usually globular domain recognises a short chain. In this respect, the ratio of the chain lengths can be misleading, as interactions with small ratio can be between globular domains. For example, the structure of a RNA polymerase, 1YNN (Campbell *et al.* 2005), where the alpha chain (314 residues) interacts with the beta chain (1119 residues). However, this ratio allows focusing in the hetero quaternary protein interfaces and transient complexes. In the previous section, we have seen that these interfaces have on average a larger proportion of atoms in deeper cavities than the obligate subset.

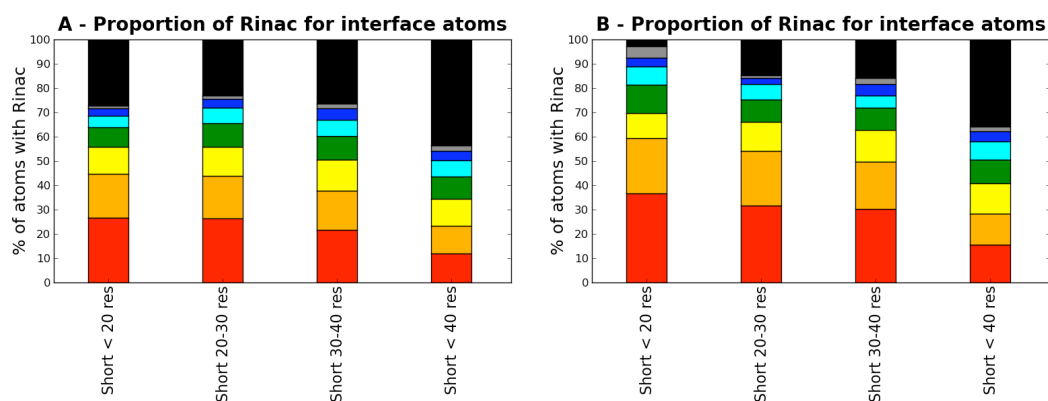


Figure 5.41. Proportion of Rinaccess values for the atoms at the interface of the long chain of the hetero quaternary interfaces (A) and for the long chain (with at least 100 residues in length) of the transient dimers and protein-protein complexes inhibited by small molecules (B). The colour in the bars denotes the Rinaccess values: red ($<2\text{\AA}$), orange (2-3 \AA), yellow (3-4 \AA), green (4-5 \AA), cyan (5-6 \AA), blue (6-7 \AA), grey (7-10 \AA) and black ($> 10\text{\AA}$). Each bar represents different length range for the short length of the complex.

Figure 5.41 shows a tendency for this proportion to increase with decreasing size of the short chain of the complex. Considering that shorter chains will define smaller interface areas. The fact that the proportion of atoms in deeper cavities for smaller chains is larger highlights the preference of these shorter peptides to target cavities at the interface and support the hypothesis that these interfaces are more amenable to be inhibited by a small molecule (Blundell *et al.* 2006).

5.3.6 Density of contacts at the binding interface

For a given cavity on the protein surface where a ligand binds, the density of contacts can be calculated per protein atom or per ligand atom. In the case of protein complexes, densities can be calculated per protein atom of each chain. As previously, I labelled protein chains as long and short depending on the length of the polypeptide. The difference in density of contacts between sides gives a rough measure of the wrapping ability of one side towards the other. In the case of small molecules, usually the ligand is wrapped inside a concave shape in the protein surface. In the case of protein

complexes, protein-protein interfaces alternate pockets in both chains of the interface. Thus, on average the density of contacts for each chain will not differ much.

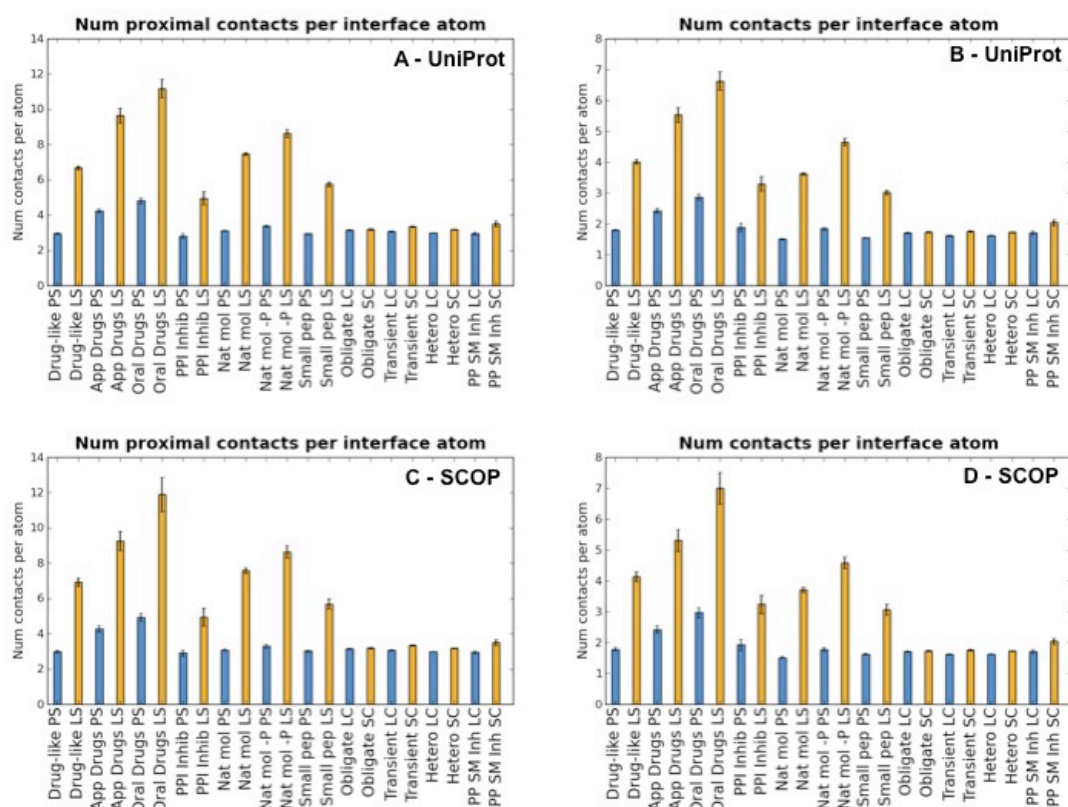


Figure 5.42. Average density of contacts per interface atom for each subset: Drug-like, Approved drugs, Oral drugs, small molecule protein-protein (PP) interaction inhibitors, natural molecules, natural molecules without phosphorous, small peptides, PP obligate dimers, PP transient dimers, PP hetero quaternary interfaces and PP complexes successfully inhibited by small molecules. For small molecule complexes, both protein side (PS, pale blue) and ligand side (LS, orange) are plotted. For the PP complexes, both long chain (LC, pale blue) and short chain (SC, orange) are plotted. Error bars denote the standard error of the mean. A: density of proximal contacts (atom pairs within 4.5\AA) at UniProt level of protein redundancy. B: density of successful contacts at UniProt level. C: density of proximal contacts at SCOP family level of protein redundancy. D: density of successful contacts at SCOP level.

Figure 5.42 (A) shows the average density of proximal contacts (atom pairs within 4.5\AA) for all sets for distinct UniProt families. Similar trends were found for distinct SCOP families Figure 5.42 (C). The results for the differences in density of contacts between long chain (or protein) and short

chain (or ligand) corroborate the distribution of R_{in} discussed in the previous section. Drugs, drug-like and natural molecules have deeper cavities that translate into a greater difference in contact densities between protein and ligand atoms. These are followed by small molecule protein-protein inhibitors and small peptides, which have shallower pockets and therefore smaller differences in contact densities between protein and ligand atoms. Regarding protein complexes, there is no significant difference between contact density for long and short chains for the obligate dimers subset. For quaternary hetero, transient interfaces and protein-protein interfaces inhibited by small molecules there is a significant difference ($P < 0.05$) in the contact density between chains, although it is small. This difference is probably due to the proportion of these complexes where a larger globular domain recognises a shorter peptide as discussed in the previous section.

Regarding differences in contact density across subsets, oral drugs are the most contact efficient in both proximal (Figure 5.42 A and C) and successful contacts (Figure 5.42 B and D). Examples of oral drugs with high and low density contacts are displayed in Figure 5.43. The obvious question that arises is whether this efficiency is because oral drugs are, on average, smaller molecules. Figure 5.44 shows that oral drugs have similar size to other drugs, drug-like and natural molecules. Furthermore, it also shows no correlation between contact density and molecular size. Oral drugs have a greater density of contacts because they are, on average, the small molecules that best fit the deep cavities in proteins, as reported previously (Fuller *et al.* 2009). This is arguably because they have been optimised to achieve tight binding. In addition, oral drugs considered here have been structurally characterised suggesting that maybe there is a bias in the data where optimisation has been achieved with structural information in hand. In contrast, natural molecules that are the product of millennia of evolution have significantly lower contact density. This is because; natural molecules evolve to preserve function, not tight binding. Indeed, Kahraman and co-workers (Kahraman *et al.* 2007) found that pockets were on average three times

bigger than the ligands they bound. In that study, a hundred protein complexes with nine ligands were considered (ATP, AMP, FAD, FMN, Glucose, Heme, NAD, Phosphate and Steroid-like molecules). These ligands belong to the natural molecule subset analysed here. They are substrates, cofactors or products of enzymatic reactions that need to transfer chemical groups to pass a response in the signalling cascade, indeed none of these are likely to be the sole occupants of the binding pocket. This is probably the reason why these molecules bind less tightly to bigger pockets as they need room to manoeuvre as well as leaving the site once the signalling is achieved.

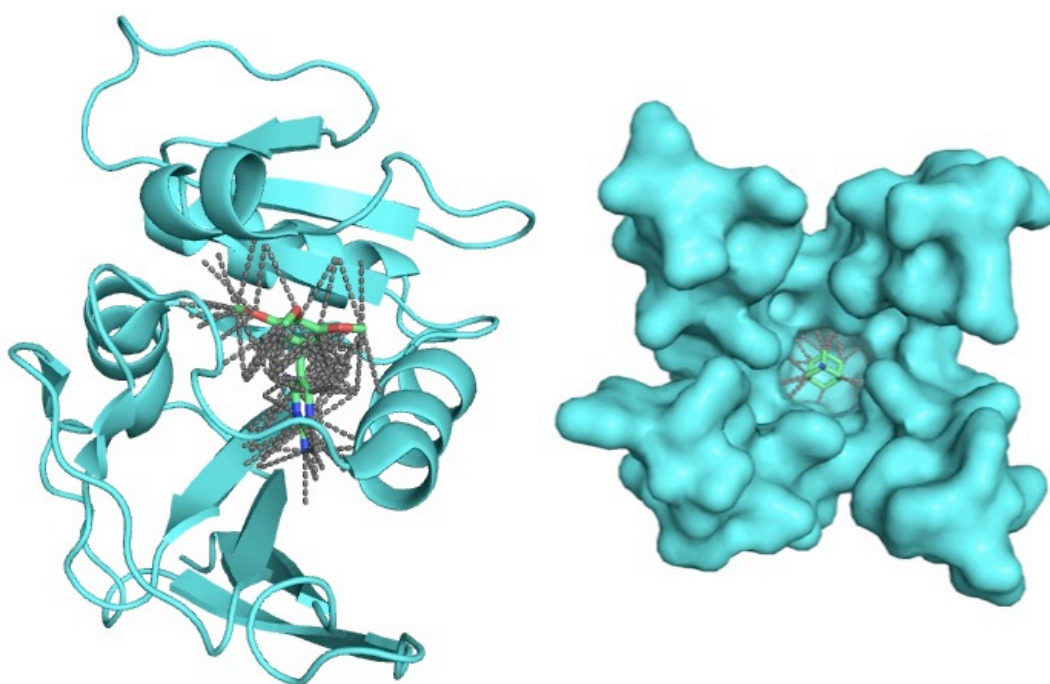


Figure 5.43. Examples of oral drugs binding to proteins, proximal contacts are represented by grey dotted lines. LEFT: 2HM9, dihydrofolate reductase complexed with trimethoprim, 16.3 proximal contacts per buried ligand atom. RIGHT: 3C9J, transmembrane domain of M2 protein complexed with amantadine, 4.6 proximal contacts per buried ligand atom.

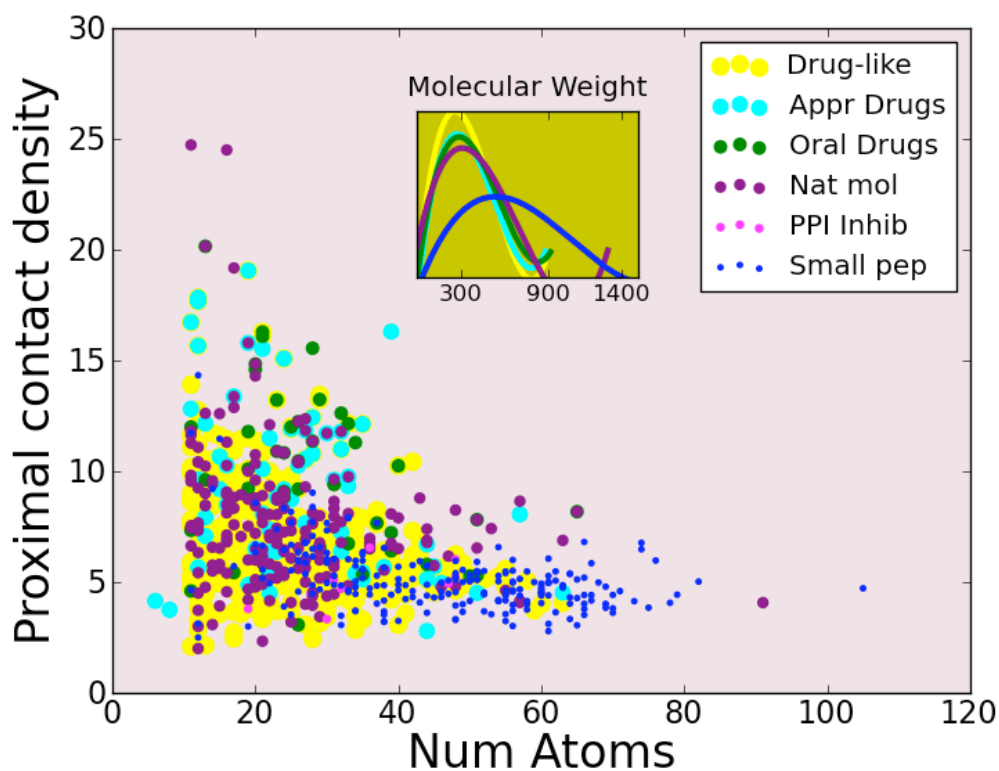


Figure 5.44. Scatter plot of the proximal contact density (the number of contacts per interacting ligand atom) versus the number of ligand atoms for the small molecule subset. The redundancy filter applied here is by distinct UniProt and distinct small molecule. Drug-like (yellow), approved drugs (cyan), oral drugs (green), natural molecules (purple), protein-protein inhibitors (magenta) and small peptides (blue). The histogram in the centre of the figure represents the molecular weight distribution.

5.4 Conclusions

The first conclusion from this analysis of the binding interfaces is that, although *binding interface* is a simple concept, it is difficult to encapsulate in a universal definition that allows unbiased comparison across different binding sites. Regarding pockets, one can distinguish between ligand-defined pocket and protein-geometry defined pocket. Comparison of cavities defined by interacting partners is biased by the interactions these entities prefer to make, whereas comparison of pockets defined by cavity detection programs is biased by the software used rather than by the potential of the sites. Furthermore, the polar ambivalent and flexible nature of the amino acids enables a range of binding profiles at the same interface, especially for protein-protein complexes.

In this chapter, I have used the binding interfaces as defined by the bound molecule.

Amongst the protein-protein interfaces studied here, the transient complexes appear to be, from the structural point of view, the most amenable to be targeted by small molecule drug based therapies. These complexes have on average deeper pockets at the interface than obligate and quaternary interfaces. In addition, these complexes are often formed by a small chain binding to a bigger protein. Indeed, most of the successfully inhibited protein-protein complexes have characteristics similar to those of the transient dimers subset with the exception of the TNF trimer.

Drug-like molecule binding sites are on average more hydrophobic and have higher aromatic content than those binding small natural molecules. However, these sites have a higher proportion of unmatched polar atoms, suggesting that in principle, the polar interaction profile for drug-like molecules could be improved. Indeed, small drug-like fragments (200-300Da) bind on average to more polar sites than larger molecules. Furthermore, for

protein-protein complexes inhibited by small molecules, comparison of the protein-protein with the protein-small molecule binding interfaces reveals that, on average, small molecules target the hydrophobic and aromatic residues instead of the polar residues available at the sites. Moreover, the higher content of flexible side chains in protein-protein interfaces confirms that a degree of adaptability to bind to small molecules and to match polar contacts is possible.

Natural molecules have on average a higher proportion of contacts with main chain atoms and a higher content of Gly at the binding site. This behaviour is mainly due to the small molecules nucleotides, such as ATP, binding to a variety of different folds. Indeed it may be that a high proportion of main chain atoms in a binding site may be characteristic of a promiscuous binding region.

Analysis of the depth of the atoms at the interfaces confirms that drugs, drug-like and natural molecules bind to deeper pockets than small peptides and small molecules inhibiting protein complexes. The density of contacts at the interfaces also corroborates this result. In this respect, oral drugs are the most contact efficient group.

The conclusions reached in chapter 4 are generally supported by the analysis of binding sites. Drug-like molecules in general, but especially those that inhibit protein complexes do not make full use of the polar signature on binding. Although the subset of natural molecules binding to many different folds has more polar contacts, they do so through a higher proportion of main chain atoms, which may well explain the intended promiscuity that has been helpful in evolution.

I start this last chapter with a quote Brian Warrington used in one of his talks:

Life can only be understood backwards; but it must be lived forwards.

Søren Kierkegaard, Danish philosopher (1813 - 1855)

The work reflected in this thesis has been an exercise in trying to understand what we have done so far with the aim of identifying areas where we can move forwards. The drug discovery community agrees that we are at an inflexion point; classical practices are being questioned and new areas are being explored. In chapter 2, we have seen, however, that when new areas are explored with the classical tools the outcome often brings us back where we started. Small molecules disrupting new drug targets, in particular protein-protein interactions, are more lipophilic than the already-too-lipophilic drug-like molecules. This is our starting point for exploring a new, more challenging drug space. In this dissertation I have looked backwards to review the progress made so far, but I have also sought to look forwards to new approaches to targeting protein-protein interactions.

6.1 Protein-protein interactions as drug targets

I have described a new public resource, TIMBAL, a database that holds small molecules inhibiting protein-protein interfaces. Comparison of these molecules with drugs on the market and those in most screening libraries underlined the fact that TIMBAL molecules tend to be bigger, more rigid, more lipophilic and with fewer hydrogen bonded atoms. Comparing the binding interfaces of the protein target with its small molecule inhibitor and with its protein partner highlighted that the small molecules prioritise hydrophobic contacts instead of the available polar patches at the protein surface. Although one of these big lipophilic molecules, ABT-263 at the Bcl-2 interface, has made its successful way into oncology clinical trials, it should not be a general model for future campaigns; rather efforts should be invested into maximising specific contacts that are available at these

interfaces. Furthermore, residue propensity comparisons between protein-protein interfaces and small molecule binding sites have confirmed that multi-protein complexes have a higher proportion of flexible side chains, which are able to match polar contacts. In terms of available cavities, studies of the depth below the surface of protein atoms that constitute a cavity have shown that transient protein complexes (especially those which are composed of a large domain interacting with a short chain) have, on average, deeper pockets. They may therefore offer greater opportunities for binding ligands with high efficiency, so making them more amenable as targets for candidate drug molecules.

6.2 Molecular recognition, synthetic versus natural molecules

Molecular recognition is a concept that describes the outcome of a complexity of both attractive and opposing forces. Atomic interactions between two molecules are not the only factors to consider. However, in drug discovery, they encode the relationship between binding affinity and molecular properties, and in turn define the ADMET space where the small molecules operate. I have demonstrated by comparisons of atomic interaction profiles between different sets of molecules that natural molecules (small molecules but also other proteins) bound to proteins have a larger proportion of polar contacts than protein-synthetic molecule complexes. Exogenous compounds are restricted within a window of “drug-like” properties that facilitate their journey in the body in order to reach their target. Specifically, oral drugs should not be very polar. Furthermore, matching too many hydrogen bonds is not only extremely difficult but also will confer a lipophilicity that would be too low to cross membranes. The results presented in this thesis, however, have shown that drug-like molecules have a higher proportion of buried polar unmatched atoms than the natural sets and probably for this reason, there is no correlation between logP and the

proportion of polar contacts. I conclude that, in principle, it should be possible to increase the specific contacts achieved by synthetic molecules without changing drastically the molecular properties of drug-like compounds. In practical terms, this seems to be feasible through fragment-based approaches. Analysis of the proportion of polar contacts versus size of molecules confirmed that a higher content in specific interactions occurs when the compounds are small. Indeed, small drug-like fragments bind on average to more polar binding patches than larger molecules. It is accepted now that the evolution of these initial hits should be along a path that optimises the affinity and molecular properties in a concerted fashion.

6.3 Concluding remarks

Structurally characterised protein complexes offer a wealth of information about molecular recognition and much insight can be gained by studying atomic interaction profiles of different types of molecules. However, data curation and redundancy assessment are paramount to extract robust conclusions. The results of such studies should reveal trends for each molecular type, arbitrarily defined in the study, but not the particular solution nature has found for that particular molecule with that particular function. Nevertheless, trends can guide us on a journey to understand what we have done so far with the aim of further improving what we should do moving forwards. I hope the work reflected in this thesis can contribute towards that end.

Bibliography

- ABAD-ZAPATERO C & MET JT (2005). Ligand efficiency indices as guideposts for drug discovery. *Drug Discov Today* 10:464–469.
- ABAD-ZAPATERO C, PERISIC O, WASS J, BENTO AP, OVERINGTON J, AL-LAZIKANI B & JOHNSON ME (2010). Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discov Today* 15(19-20):804-811.
- ACHARYA KR & LLOYD MD (2005). The advantages and limitations of protein crystal structures. *Trends Pharmacol Sci* 26(1):10-14.
- ADAIR JR & LAWSON ADG (2005). Therapeutic Antibodies. *Drug Design Reviews - Online* 2(3):209-217.
- ALA PJ, HUSTON EE, KLABE RM, JADHAV PK, LAM PYS & CHANG C-H (1998). Counteracting HIV-1 Protease Drug Resistance: Structural Analysis of Mutant Proteases Complexed with XV638 and SD146, Cyclic Urea Amides with Broad Specificities. *Biochemistry* 37(43):15042-15049.
- AN J, TOTROV M & ABAGYAN R (2005). Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Molecular & Cellular Proteomics* 4(6):752-761.
- ARKIN MR, RANDAL M, DELANO WL, HYDE J, LUONG TN, OSLOB JD, RAPHAEL DR, TAYLOR L, WANG J, MCDOWELL RS, WELLS JA & BRAISTED AC (2003). Binding of small molecules to an adaptive protein-protein interface. *Proc Natl Acad Sci U S A* 100:1603–1608.
- ARUNAN E, DESIRAJU GR, KLEIN RA, SADLEJ J, SCHEINER S, ALKORTA I, CLARY DC, CRABTREE RH, DANNENBERG JJ, HOBZA P, KJAERGAARD HG, LEGON AC, MENNUCCI B & NESBITT DJ (2011). Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure and Applied Chemistry* 83(8):1637-1641.
- BAHADUR RP, CHAKRABARTI P, RODIER F & JANIN J (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* 336(4):943-955.
- BALDWIN R (2007). Energetics of Protein Folding. *J Mol Biol* 371(2):283-301.
- BARLOW DJ & THORNTON JM (1983). Ion-pairs in proteins. *J Mol Biol* 168(4):867-885.

- BAUER RA, WURST JM & TAN DS (2010). Expanding the range of 'druggable' targets with natural product-based libraries: an academic perspective. *Curr Opin Chem Biol* 14(3):308-314.
- BERG T (2003). Modulation of protein-protein interactions with small organic molecules. *Angew Chem Int Ed Engl* 42(22):2462-2481.
- BERMAN HM, WESTBROOK J, FENG Z, GILLILAND G, BHAT TN, WEISSIG H, SHINDYALOV IN & BOURNE PE (2000). The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242.
- BICKERTON GR (2009). Molecular characterization and evolutionary plasticity of protein-protein interfaces. *PhD University of Cambridge, Cambridge, UK*
- BICKERTON GR, HIGUERUELO AP & BLUNDELL TL (2011). Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC Bioinformatics* 12(1):313.
- BIRD GH, MADANI N, PERRY AF, PRINCIOTTO AM, SUPKO JG, HE X, GAVATHIOTIS E, SODROSKI JG & WALENSKY LD (2010). Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proc Natl Acad Sci U S A* 107(32):14093-14098.
- BIRTALAN S, ZHANG Y, FELLOUSE FA, SHAO L, SCHAEFER G & SIDHU SS (2008). The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol* 377(5):1518-1528.
- BISSANTZ C, KUHN B & STAHL M (2010). A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem* 53(14):5061-5084.
- BLUNDELL TL, SIBANDA BL, MONTALVAO RW, BREWERTON S, CHELLIAH V, WORTH CL, HARMER NJ, DAVIES O & BURKE D (2006). Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos Trans R Soc Lond B Biol Sci* 361(1467):413-423.
- BOELSTERLI UA, HO HK, ZHOU S & LEOW KY (2006). Bioactivation and hepatotoxicity of nitroaromatic drugs. *Curr Drug Metab* 7(7):715-727.
- BOGAN AA & THORN KS (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol* 280(1):1-9.
- BÖHM H-J & KLEBE G (1996). What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? *Angew Chem Int Ed Engl* 35(22):2588-2614.
- BONDI A (1964). van der Waals Volumes and Radii. *The Journal of Physical Chemistry* 68(3):441-451.

- BOURGEAS RL, BASSE M-J, MORELLI X & ROCHE P (2010). Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. *PLoS ONE* 5(3):e9598.
- BRADEN BC & POLJAK RJ (2000). Chapter 5: Structure and energetics of anti-lysozyme antibodies. Protein-Protein Recognition. C. Kleanthous. Oxford, *Oxford University Press*: 123-161.
- BRODERSEN DE, CLEMONS WM, JR., CARTER AP, MORGAN-WARREN RJ, WIMBERLY BT & RAMAKRISHNAN V (2000). The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell* 103(7):1143-1154.
- BROWN S & HAJDUK P (2006). Effects of Conformational Dynamics on Predicted Protein Druggability. *ChemMedChem* 1(1):70-72.
- CALABRESE JC, JORDAN DB, BOODHOO A, SARIASLANI S & VANNELLI T (2004). Crystal structure of phenylalanine ammonia lyase: multiple helix dipoles implicated in catalysis. *Biochemistry* 43(36):11403-11416.
- CAMPBELL EA, PAVLOVA O, ZENKIN N, LEON F, IRSCHIK H, JANSEN R, SEVERINOV K & DARST SA (2005). Structural, functional, and genetic analysis of sorangicin inhibition of bacterial RNA polymerase. *EMBO J* 24(4):674-682.
- CAPRA JA, LASKOWSKI RA, THORNTON JM, SINGH M & FUNKHOUSER TA (2009). Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Comput Biol* 5(12):e1000585.
- CCDC. "Elemental Radii." 2011, from <http://www.ccdc.cam.ac.uk/products/csd/radii/>.
- CHAKRABARTI P & BHATTACHARYYA R (2007). Geometry of nonbonded interactions involving planar groups in proteins. *Prog Biophys Mol Biol* 95(1-3):83-137.
- CHAKRABARTI P & JANIN J (2002). Dissecting protein-protein recognition sites. *Proteins* 47(3):334-343.
- CHARPENTIER T, WILDER P, LIRIANO M, VARNEY K, ZHONG S, COOP A, POZHARSKI E, MACKERELL A, TOTH E & WEBER D (2009). Small Molecules Bound to Unique Sites in the Target Protein Binding Cleft of Calcium-Bound S100B As Characterized by Nuclear Magnetic Resonance and X-ray Crystallography. *Biochemistry* 48(26):6202-6212.
- CHEN K & KURGAN L (2009). Investigation of Atomic Level Patterns in Protein-Small Ligand Interactions. *PLoS ONE* 4(2):e4473.
- CHO KI, LEE K, LEE KH, KIM D & LEE D (2006). Specificity of molecular interactions in transient protein-protein interaction interfaces. *Proteins* 65(3):593-606.

- CHOTHIA C (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 105(1):1-12.
- CHRISTOPOULOS A (2002). Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat Rev Drug Discov* 1:198-210.
- CLACKSON T & WELLS JA (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* 267(5196):383-386.
- COCHRAN AG (2000). Antagonists of protein-protein interactions. *Chem Biol* 7(4):R85-R94.
- CONGREVE M, CHESSARI G, TISI D & WOODHEAD AJ (2008). Recent Developments in Fragment-Based Drug Discovery. *J Med Chem* 51(13):3661-3680.
- CONTI E, RIVETTI C, WONACOTT A & BRICK P (1998). X-ray and spectrophotometric studies of the binding of proflavin to the S1 specificity pocket of human alpha-thrombin. *FEBS Lett* 425(2):229-233.
- COOPER TWJ, CAMPBELL IB & MACDONALD SJF (2010). Factors Determining the Selection of Organic Reactions by Medicinal Chemists and the Use of These Reactions in Arrays (Small Focused Libraries). *Angew Chem Int Ed Engl* 49(44):8082-8091.
- COWAN-JACOB SW, FENDRICH G, FLOERSHEIMER A, FURET P, LIEBETANZ J, RUMMEL G, RHEINBERGER P, CENTELEGHE M, FABBRO D & MANLEY PW (2007). Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallogr D Biol Crystallogr* 63(Pt 1):80-93.
- COYNE AG, SCOTT DE & ABELL C (2010). Drugging challenging targets using fragment-based approaches. *Curr Opin Chem Biol* 14(3):299-307.
- CUMMINGS MD, FARNUM MA & NELEN MI (2006). Universal screening methods and applications of ThermoFluor. *J Biomol Screen* 11(7):854-863.
- DAVIS AM, TEAGUE SJ & KLEYWEGT GJ (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* 42(24):2718-2736.
- DAYLIGHT SMARTS query for aminoacids.
[*]([NX3H,NX4H2+]),[*]([NX3](C)(C)(C))1[CX4H]([CH2][CH2][CH2]1)[CX3](=[OX1])[OX2H,OX1-,N]),[*]([NX3H2,NX4H3+]),[*]([NX3H](C)(C))[CX4H2][CX3](=[OX1])[OX2H,OX1-,N]),[*]([NX3H2,NX4H3+]),[*]([NX3H](C)(C))[CX4H]([*])[CX3](=[OX1])[OX2H,OX1-,N])].

- DE S, KRISHNADEV O, SRINIVASAN N & REKHA N (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol* 5:15.
- DI MICCO S, VITALE R, PELLECCIA M, REGA M, RIVA R, BASSO A & BIFULCO G (2009). Identification of Lead Compounds As Antagonists of Protein Bcl-xL with a Diversity-Oriented Multidisciplinary Approach. *J Med Chem* 52(23):7856-7867.
- DINKEL H, MICHAEL S, WEATHERITT RJ, DAVEY NE, VAN ROEY K, ALTENBERG B, TOEDT G, UYAR B, SEILER M, BUDD A, JÖDICKE L, DAMMERT MA, SCHROETER C, HAMMER M, SCHMIDT T, JEHL P, MCGUIGAN C, DYMECKA M, CHICA C, LUCK K, VIA A, CHATRYAMONTRI A, HASLAM N, GREBNEV G, EDWARDS RJ, STEINMETZ MO, MEISELBACH H, DIELLA F & GIBSON TJ (2011). ELM - the database of eukaryotic linear motifs. *Nucleic Acids Res* 2011 Nov 21 [Epub ahead of print].
- DOBSON CM (2004). Chemical space and biology. *Nature* 432:824-828.
- DOBSON PD & KELL DB (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Discov* 7(3):205-220.
- ECK MJ & SPRANG SR (1989). The structure of tumor necrosis factor-alpha at 2.6 Å resolution. Implications for receptor binding. *J Biol Chem* 264(29):17595-17605.
- EDFELDT FNB, FOLMER RHA & BREEZE AL (2011). Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov Today* 16(7-8):284-287.
- ERLANSON D (2011). Sixth Annual Fragment-Based Drug Discovery. Blog: Practical fragments. D. Erlanson and T. Zartler. 2011.
- ERLANSON D, WELLS J & BRAISTED A (2004). TETHERING: Fragment-Based Drug Discovery. *Annu Rev Biophys Biomol Struct* 33(1):199-223.
- FEHER M & SCHMIDT JM (2002). Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J Chem Inf Comput Sci* 43(1):218-227.
- FERENCZY GG & KESERÜ GM (2010). Enthalpic efficiency of ligand binding. *J Chem Inf Model* 50(9):1536-1541.
- FERENCZY GG & KESERÜ GM (2010). Thermodynamics guided lead discovery and optimization. *Drug Discov Today* 15(21-22):919-932.
- FISCHER PM (2005). Protein-Protein Interactions in Drug Discovery. *Drug Design Reviews - Online* 2(3):179-207.

- FOLOPPE N, FISHER LM, HOWES R, POTTER A, ROBERTSON AG & SURGENOR AE (2006). Identification of chemically diverse Chk1 inhibitors by receptor-based virtual screening. *Bioorg Med Chem* 14(14):4792-4802.
- FREIRE E (2008). Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov Today* 13(19-20):869-874.
- FREIRE E (2009). A Thermodynamic Approach to the Affinity Optimization of Drug Candidates. *Chemical Biology & Drug Design* 74(5):468-472.
- FRY DC (2008). Drug-like inhibitors of protein-protein interactions: a structural examination of effective protein mimicry. *Curr Protein Pept Sci* 9(3):240-247.
- FRY DC & VASSILEV L (2005). Targeting protein-protein interactions for cancer therapy. *J Mol Med (Berl)* 83(12):955-963.
- FULLER JC, BURGOYNE NJ & JACKSON RM (2009). Predicting druggable binding sites at the protein-protein interface. *Drug Discov Today* 14(3-4):155-161.
- GALLIVAN JP & DOUGHERTY DA (1999). Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A* 96(17):9459-9464.
- GALLOWAY WR, ISIDRO-LLOBET A & SPRING DR (2010). Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat Commun* 1:80.
- GARMAN SC, WURZBURG BA, TARCHEVSKAYA SS, KINET JP & JARDETZKY TS (2000). Structure of the Fc fragment of human IgE bound to its high-affinity receptor Fc epsilonRI alpha. *Nature* 406(6793):259-266.
- GARNIER JP (2008). Rebuilding the R&D engine in big pharma. *Harv Bus Rev* 86(5):68-70.
- GAULTON A, BELLIS LJ, BENTO AP, CHAMBERS J, DAVIES M, HERSEY A, LIGHT Y, MCGLINCHEY S, MICHALOVICH D, AL-LAZIKANI B & OVERINGTON JP (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40(D1):D1100-D1107.
- GAVATHIOTIS E, SUZUKI M, DAVIS ML, PITZER K, BIRD GH, KATZ SG, TU H-C, KIM H, CHENG EHY, TJANDRA N & WALENSKY LD (2008). BAX activation is initiated at a novel interaction site. *Nature* 455(7216):1076-1081.
- GHERARDINI PF, AUSIELLO G, RUSSELL RB & HELMER-CITTERICH M (2010). Modular architecture of nucleotide-binding pockets. *Nucleic acids res* 38(11):3809-3816.

- GHUMAN J, ZUNSZAIN PA, PETITPAS I, BHATTACHARYA AA, OTAGIRI M & CURRY S (2005). Structural basis of the drug-binding specificity of human serum albumin. *J Mol Biol* 353(1):38-52.
- GLEESON MP, HERSEY A, MONTANARI D & OVERINGTON J (2011). Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nature Reviews Drug Discovery* 10(3):197-208.
- GONZALEZ-RUIZ D & GOHLKE H (2006). Targeting Protein-Protein Interactions with Small Molecules: Challenges and Perspectives for computational Binding Epitope Detection and Ligand Finding. *Curr Med Chem* 13(22):2607-2625.
- GRASBERGER BL, LU T, SCHUBERT C, PARKS DJ, CARVER TE, KOBLISH HK, CUMMINGS MD, LAFRANCE LV, MILKIEWICZ KL, CALVO RR, MAGUIRE D, LATTANZE J, FRANKS CF, ZHAO S, RAMACHANDREN K, BYLEBYL GR, ZHANG M, MANTHEY CL, PETRELLA EC, PANTOLIANO MW, DECKMAN IC, SPURLINO JC, MARONEY AC, TOMCZUK BE, MOLLOY CJ & BONE RF (2005). Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *J Med Chem* 48(4):909-912.
- GUNEY E, TUNCBAG N, KESKIN O & GURSOY A (2008). HotSprint: database of computational hot spots in protein interfaces. *Nucleic acids res* 36(suppl 1):D662-D666.
- HAJDUK P (2006). SAR by NMR: Putting the Pieces Together. *Mol. Interv.* 6(5):266-272.
- HAJDUK PJ, HUTH JR & FESIK SW (2005). Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* 48(7):2518-2525.
- HAMELRYCK T & MANDERICK B (2003). PDB file parser and structure class implemented in Python. *Bioinformatics* 19(17):2308-2310.
- HANN MM (2011). Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* 2(5):349-355.
- HANN MM, LEACH AR & HARPER G (2001). Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J Chem Inf Comput Sci* 41(3):856-864.
- HARTSHORN MJ, VERDONK ML, CHESSARI G, BREWERTON SC, MOOIJ WT, MORTENSON PN & MURRAY CW (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* 50(4):726-741.
- HE M, SMITH A, OSLOB J, FLANAGAN W, BRAISTED A, WHITTY A, CANCELLA M, WANG J, LUGOVSKOY A, YOBURN J, FUNG A, FARRINGTON G, ELDREDGE J, DAY E, CRUZ L, CACHERO T, MILLER S, FRIEDMAN J, CHOONG I & CUNNINGHAM B (2005). Small-Molecule Inhibition of TNF-alpha. *Science* 310(5750):1022-1025.

- HEADD JJ, BAN YEA, BROWN P, EDELSBRUNNER H, VAIDYA M & RUDOLPH J (2007). Protein-Protein Interfaces: Properties, Preferences, and Projections. *J. Proteome Res.* 6(7):2576-2586.
- HIGUERUELO AP, SCHREYER A, BICKERTON GRJ, PITT WR, GROOM CR & BLUNDELL TL (2009). Atomic Interactions and Profile of Small Molecules Disrupting Protein-Protein Interfaces: the TIMBAL Database. *Chem Biol Drug Des* 74(5):457-467.
- HIRSCHLER B (2009). Glaxo CEO admits R&D overhaul has been traumatic. *Reuters.com* 19 June:<http://www.reuters.com/article/2009/2006/2019/glaxo-ceo-idUSLJ9191220090619>.
- HOPKINS A, GROOM C & ALEX A (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* 9(10):430-431.
- HOPKINS AL & GROOM CR (2002). The druggable genome. *Nat Rev Drug Discov* 1(9):727-730.
- HORVÁTH I, HARMAT V, PERCZEL A, PÁLFI V, NYITRAY L, NAGY A, HLAVANDA E, NÁRAY-SZABÓ G & OVÁDI J (2005). The Structure of the Complex of Calmodulin with KAR-2. *J Biol Chem* 280(9):8266-8274.
- HYVÖNEN M, BEGUN J & BLUNDELL TL (2000). Chapter 7: Protein-protein interactions in eukaryotic signal transduction. Protein-Protein Recognition. C. Kleanthous. Oxford, *Oxford University Press*: 189-227.
- JANIN J (2000). Chapter 1: Kinetics and Thermodynamics of protein-protein interactions. Protein-Protein Recognition. C. Kleanthous. Oxford, *Oxford University Press*: 1-32.
- JANIN J, RODIER F, CHAKRABARTI P & BAHADUR RP (2007). Macromolecular recognition in the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 63(Pt 1):1-8.
- JI H-F, KONG D-X, SHEN L, CHEN L-L, MA B-G & ZHANG H-Y (2007). Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biology* 8(8):R176.
- JONES E, OLIPHANT T, PETERSON P & OTHERS. (2001 -). "SciPy: Open Source Scientific Tools for Python ", from <http://www.scipy.org>.
- JONES S & THORNTON JM (2000). Chapter 2: Analysis and classification of protein-protein interactions from a structural perspective. Protein-Protein Recognition. C. Kleanthous. Oxford, *Oxford University Press*: 33-59.
- KAHRAMAN A, MORRIS RJ, LASKOWSKI RA, FAVIA AD & THORNTON JM (2010). On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins* 78(5):1120-1136.

- KAHRAMAN A, MORRIS RJ, LASKOWSKI RA & THORNTON JM (2007). Shape variation in protein binding pockets and their ligands. *J Mol Biol* 368(1):283-301.
- KANEHISA M, ARAKI M, GOTO S, HATTORI M, HIRAKAWA M, ITOH M, KATAYAMA T, KAWASHIMA S, OKUDA S, TOKIMATSU T & YAMANISHI Y (2008). KEGG for linking genomes to life and the environment. *Nucleic acids res* 36(Database):D480-D484.
- KANEHISA M, GOTO S, FURUMICHI M, TANABE M & HIRAKAWA M (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38(Database issue):D355-360.
- KANG L (2008). Microfluidics for drug discovery and development: From target selection to product lifecycle management. *Drug Discov Today* 13(1-2):1-13.
- KAWABATA T (2010). Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* 78(5):1195-1211.
- KAWABATA T & GO N (2007). Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* 68(2):516-529.
- KELLENBERGER E, MULLER P, SCHALON C, BRET G, FOATA N & ROGNAN D (2006). sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J Chem Inf Model* 46(2):717-727.
- KESERÜ GM & MAKARA GM (2009). The influence of lead discovery strategies on the properties of drug candidates. *Nat Rev Drug Discov* 8(3):203-212.
- KESKIN O, GURSOY A, MA B & NUSSINOV R (2008). Principles of Protein-Protein Interactions: What are the Preferred Ways For Proteins To Interact? *Chem. Rev.*
- KESKIN O, MA B & NUSSINOV R (2005). Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* 345(5):1281-1294.
- KIM MK, LEE JH, KIM H, PARK SJ, KIM SH, KANG GB, LEE YS, KIM JB, KIM KK, SUH SW & EOM SH (2006). Crystal structure of visfatin/pre-B cell colony-enhancing factor 1/nicotinamide phosphoribosyltransferase, free and in complex with the anti-cancer agent FK-866. *J Mol Biol* 362(1):66-77.
- KLEBE G (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 11(13-14):580-594.
- KNOX C, LAW V, JEWISON T, LIU P, LY S, FROLKIS A, PON A, BANCO K, MAK C, NEVEU V, DJOUMBOU Y, EISNER R, GUO AC & WISHART DS (2011). DrugBank 3.0: a

- comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 39(Database issue):D1035-1041.
- KOCH M, SCHUFFENHAUER A, SCHECK M, WETZEL S, CASALTA M, ODERMATT A, ERTL P & WALDMANN H (2005). Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci U S A* 102(48):17272-17277.
- KOZAKOV D, HALL DR, CHUANG G-Y, CENCIC R, BRENKE R, GROVE LE, BEGLOV D, PELLETIER J, WHITTY A & VAJDA S (2011). Structural conservation of druggable hot spots in protein-protein interfaces. *Proc Natl Acad Sci U S A* 108(33):13528-13533.
- KRISSINEL E & HENRICK K (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372(3):774-797.
- KUMAR S & NUSSINOV R (2002). Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophysical journal* 83:1595-1612.
- KUSSIE PH, GORINA S, MARECHAL V, ELENBAAS B, MOREAU J, LEVINE AJ & PAVLETICH NP (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 274(5289):948-953.
- LADBURY JOHN E (2010). Calorimetry as a tool for understanding biomolecular interactions and an aid to drug design. *Biochem Soc Trans* 38(4):888.
- LADBURY JE, KLEBE G & FREIRE E (2010). Adding calorimetric data to decision making in lead discovery: a hot tip. *Nat Rev Drug Discov* 9(1):23-27.
- LASKOWSKI RA (1995). SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13(5):323-330.
- LAURIE AT & JACKSON RM (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21(9):1908-1916.
- LAWRENCE MC & COLMAN PM (1993). Shape complementarity at protein/protein interfaces. *J Mol Biol* 234(4):946-950.
- LE GUILLOUX V, SCHMIDTKE P & TUFFERY P (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10:168.
- LEE EF, CZABOTAR PE, SMITH BJ, DESHAYES K, ZOBEL K, COLMAN PM & FAIRLIE WD (2007). Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. *Cell Death Differ* 14(9):1711-1713.
- LEE S & BLUNDELL TL (2009). BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics* 25(12):1559-1560.

- LEESON P & SPRINGTHORPE B (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* 6(11):881-890.
- LEESON PD & EMPFIELD JR (2010). Chapter 24 - Reducing the Risk of Drug Attrition Associated with Physicochemical Properties. Annual Reports in Medicinal Chemistry. E. M. John, *Academic Press*. Volume 45: 393-407.
- LEVINSON NM, KUCHMENT O, SHEN K, YOUNG MA, KOLDOBSKIY M, KARPLUS M, COLE PA & KURIYAN J (2006). A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol* 4(5):e144.
- LI JW & VEDERAS JC (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science* 325(5937):161-165.
- LIPINSKI C, LOMBARDO F, DOMINY B & FEENEY P (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 23(1-3):3-25.
- LIU Z, SUN C, OLEJNICZAK ET, MEADOWS RP, BETZ SF, OOST T, HERRMANN J, WU JC & FESIK SW (2000). Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. *Nature* 408(6815):1004-1008.
- LIVESAY DR & SUBRAMANIAM S (2004). Conserved sequence and structure association motifs in antibody-protein and antibody-hapten complexes. *Protein Eng Des Sel* 17(5):463-472.
- LO CONTE L, CHOTHIA C & JANIN J (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol* 285(5):2177-2198.
- LOMBARDINO JG & LOWE JA (2004). The role of the medicinal chemist in drug discovery - then and now. *Nat Rev Drug Discov* 3(10):853-862.
- LOVERING F, BIKKER J & HUMBLET C (2009). Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J Med Chem* 52(21):6752-6756.
- MACARRON R & LUENGO JI (2011). Yin and Yang in medicinal chemistry: what does drug-likeness mean? *Future Medicinal Chemistry* 3(5):505-507.
- MAITA N, OKADA K, HATAKEYAMA K & HAKOSHIMA T (2002). Crystal structure of the stimulatory complex of GTP cyclohydrolase I and its feedback regulatory protein GFRP. *Proc Natl Acad Sci U S A* 99(3):1212-1217.
- MARCOU G & ROGNAN D (2007). Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model* 47(1):195-207.

- MATHEWS FS, MAUK AG & MORAS D (2000). Chapter 3: Protein-protein complexes formed by electron transfer proteins. *Protein-Protein Recognition*. C. Kleanthous. Oxford, *Oxford University Press*: 60-101.
- MCDONALD IK & THORNTON JM (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 238:777-793.
- MINTSERIS J & WENG Z (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 102(31):10930-10935.
- MORELLI X, BOURGEAS R & ROCHE P (2011). Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). *Curr Opin Chem Biol* 15(4):475-481.
- MOSYAK L, ZHANG Y, GLASFELD E, HANEY S, STAHL M, SEEHRA J & SOMERS WS (2000). The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *EMBO J* 19(13):3179-3191.
- MURZIN A, BRENNER S, HUBBARD T & CHOTHIA C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-540.
- NEDUVA V & RUSSELL RB (2006). Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol* 17(5):465-471.
- NEUGEBAUER A, HARTMANN RW & KLEIN CD (2007). Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. *J Med Chem* 50(19):4665-4668.
- NICOLA G & VAKSER IA (2007). A simple shape characteristic of protein-protein recognition. *Bioinformatics* 23(7):789-792.
- NOOREN IM & THORNTON JM (2003). Diversity of protein-protein interactions. *EMBO J* 22(14):3486-3492.
- ODA A, YAMAOTSU N & HIRONO S (2009). Evaluation of the searching abilities of HBOP and HBSITE for binding pocket detection. *J Comp Chem* 30(16):2728-2737.
- OFRAN Y & ROST B (2003). Analysing six types of protein-protein interfaces. *J Mol Biol* 325(2):377-387.
- OLSSON TSG, WILLIAMS MA, PITT WR & LADBURY JE (2008). The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design. *J Mol Biol* 384(4):1002-1017.

- OOST TK, SUN C, ARMSTRONG RC, AL-ASSAAD AS, BETZ SF, DECKWERTH TL, DING H, ELMORE SW, MEADOWS RP, OLEJNICZAK ET, OLEKSIJEW A, OLTERSDORF T, ROSENBERG SH, SHOEMAKER AR, TOMASELLI KJ, ZOU H & FESIK SW (2004). Discovery of potent antagonists of the antiapoptotic protein XIAP for the treatment of cancer. *J Med Chem* 47(18):4417-4426.
- OSBORNE J & WATERS E (2002). Four assumptions of multiple regression that researchers should always test *Practical Assessment, Research & Evaluation* 8:2.
- PAGLIARO L, FELDING J, AUDOUZE K, NIELSEN SJ, TERRY RB, KROG-JENSEN C & BUTCHER S (2004). Emerging classes of protein-protein interaction inhibitors and new tools for their development. *Curr Opin Chem Biol* 8(4):442-449.
- PAL D & CHAKRABARTI P (2001). Non-hydrogen bond interactions involving the methionine sulfur atom. *J Biomol Struct Dyn* 19(1):115-128.
- PATERA A, BLASZCZAK LC & SHOICHET BK (2000). Crystal Structures of Substrate and Inhibitor Complexes with AmpC-Lactamase: Possible Implications for Substrate-Assisted Catalysis. *J Am Chem Soc* 122:10504-10512.
- PAUL SM, MYTELKA DS, DUNWIDDIE CT, PERSINGER CC, MUNOS BH, LINDBORG SR & SCHACHT AL (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 9(3):203-214.
- PDB TEAM (2011). PDB Statistics <http://www.rcsb.org/pdb/statistics/holdings.do>.
- PEROT S, SPERANDIO O, MITEVA MA, CAMPROUX AC & VILLOUTREIX BO (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today* 15(15-16):656-667.
- PILLAI B, CHERNEY MM, HIRAGA K, TAKADA K, ODA K & JAMES MN (2007). Crystal structure of scytilidoglutamic peptidase with its first potent inhibitor provides insights into substrate specificity and catalysis. *J Mol Biol* 365(2):343-361.
- PITT WR, PARRY DM, PERRY BG & GROOM CR (2009). Heteroaromatic Rings of the Future. *J Med Chem* 52(9):2952-2963.
- REICHMANN D, RAHAT O, COHEN M, NEUVIRTH H & SCHREIBER G (2007). The molecular architecture of protein-protein binding sites. *Curr Opin Struct Biol* 17(1):67-76.
- REYNES C, HOST H, CAMPROUX A-C, LACONDE G, LEROUX F, MAZARS A, DEPRez B, FAHRAEUS R, VILLOUTREIX BO & SPERANDIO O (2010). Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors using Machine-Learning Methods. *PLoS Comput Biol* 6(3):e1000695.

- RICKERT M, WANG X, BOULANGER MJ, GORIATCHEVA N & GARCIA KC (2005). The structure of interleukin-2 complexed with its alpha receptor. *Science* 308(5727):1477-1480.
- RODIER F, BAHADUR RP, CHAKRABARTI P & JANIN J (2005). Hydration of protein-protein interfaces. *Proteins* 60(1):36-45.
- RUSH TS, GRANT JA, MOSYAK L & NICHOLLS A (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 48(5):1489-1495.
- RUSTANDI RR, BALDISSERI DM & WEBER DJ (2000). Structure of the negative regulatory domain of p53 bound to S100B(beta-beta). *Nat Struct Biol* 7(7):570-574.
- RUZHEINIKOV S. (2007). "Protein Crystallography." from <http://proteincrystallography.org/>.
- SCHMIDTKE P & BARRIL X (2010). Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J Med Chem* 53(15):5858-5867.
- SCHMIDTKE P, SOUAILLE C, ESTIENNE F, BAURIN N & KROEMER RT (2010). Large-scale comparison of four binding site detection algorithms. *J Chem Inf Model* 50(12):2191-2200.
- SCHREYER A (2010). Characterisation of protein-ligand interactions and their application to drug discovery. *PhD University of Cambridge, Cambridge, UK*
- SCHREYER A & BLUNDELL T (2009). CREDO: A Protein-Ligand Interaction Database for Drug Discovery. *Chem Biol Drug Des* 73(2):157-167.
- SEGURA J & FERNANDEZ-FUENTES N (2011). PCRPI-DB: a database of computationally annotated hot spots in protein interfaces. *Nucleic acids res* 39(suppl 1):D755-D760.
- SINGH J & THORNTON J (1992). Atlas of protein side-chain interactions, *Oxford University Press (IRL Press)*.
- SINGH N, GUHA R, GIULIANOTTI MA, PINILLA C, HOUGHTEN RA & MEDINA-FRANCO JL (2009). Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J Chem Inf Model* 49(4):1010-1024.
- SINGH N, JABEEN T, SHARMA S, SOMVANSHI RK, DEY S, SRINIVASAN A & SINGH TP (2006). Specific binding of non-steroidal anti-inflammatory drugs (NSAIDs) to phospholipase A2: structure of the complex formed between phospholipase

- A2 and diclofenac at 2.7 Å resolution. *Acta Crystallogr D Biol Crystallogr* 62(Pt 4):410-416.
- SOGA S, SHIRAI H, KOBORI M & HIRAYAMA N (2007). Use of amino acid composition to predict ligand-binding sites. *J Chem Inf Model* 47(2):400-406.
- SPERANDIO O, REYNES CH, CAMPROUX A-C & VILLOUTREIX BO (2010). Rationalizing the chemical space of protein-protein interaction inhibitors. *Drug Discov Today* 15(5-6):220-229.
- STAMS T, CHEN Y, CHRISTIANSON DW, BORIACK-SJODIN PA, HURT JD, LAIPIS P, SILVERMAN DN, LIAO J, MAY JA & DEAN T (1998). Structures of murine carbonic anhydrase IV and human carbonic anhydrase II complexed with brinzolamide: Molecular basis of isozyme-drug discrimination. *Protein Sci* 7(3):556-563.
- STEIN A & ALOY P (2008). Contextual Specificity in Peptide-Mediated Protein Interactions. *PLoS ONE* 3(7): e2524.
- SUGANO K, KANSY M, ARTURSSON P, AVDEEF A, BENDELS S, DI L, ECKER GF, FALLER B, FISCHER H, GEREBTZOFF G, LENNERNAES H & SENNER F (2010). Coexistence of passive and carrier-mediated processes in drug transport. *Nat Rev Drug Discov* 9(8):597-614.
- SVENSSON S, OSTBERG T, JACOBSSON M, NORSTROM C, STEFANSSON K, HALLEN D, JOHANSSON IC, ZACHRISSON K, OGG D & JENDEBERG L (2003). Crystal structure of the heterodimeric complex of LXRalpha and RXRbeta ligand-binding domains in a fully agonistic conformation. *EMBO J* 22(18):4625-4633.
- SWINNEY DC & ANTHONY J (2011). How were new medicines discovered? *Nat Rev Drug Discov* 10(7):507-519.
- THAISRIVONGS S, SKULNICK HI, TURNER SR, STROHBACH JW, TOMMASI RA, JOHNSON PD, ARISTOFF PA, JUDGE TM, GAMMILL RB, MORRIS JK, ROMINES KR, CHRUSCIEL RA, HINSHAW RR, CHONG KT, TARPLEY WG, POPPE SM, SLADE DE, LYNN JC, HORNG MM, TOMICH PK, SEEST EP, DOLAK LA, HOWE WJ, HOWARD GM, WATENPAUGH KD & ET AL. (1996). Structure-based design of HIV protease inhibitors: sulfonamide-containing 5,6-dihydro-4-hydroxy-2-pyrones as non-peptidic inhibitors. *J Med Chem* 39(22):4349-4353.
- THANOS CD, DELANO WL & WELLS JA (2006). Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A* 103:15422-15422.
- THANOS CD, RANDAL M & WELLS JA (2003). Potent Small-Molecule Binding to a Dynamic Hot Spot on IL-2. *J Am Chem Soc* 125(50):15280-15281.
- THE UNIPROT C (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* 39(suppl 1):D214-D219.

- THORN KS & BOGAN AA (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17(3):284-285.
- TINA KG, BHADRA R & SRINIVASAN N (2007). PIC: Protein Interactions Calculator. *Nucleic Acids Res* 35(Web Server issue):W473-476.
- TOOGOOD PL (2002). Inhibition of protein-protein association by small molecules: approaches and progress. *J Med Chem* 45(8):1543-1558.
- TOWNEND J (2002). Correlation and regression. Practical Statistics for Environmental and Biological Scientists, *John Wiley & Sons Ltd*: 129-152.
- TSAI J, TAYLOR R, CHOTHIA C & GERSTEIN M (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol* 290:253-266-253-266.
- UNDERWOOD KW, PARRIS KD, FEDERICO E, MOSYAK L, CZERWINSKI RM, SHANE T, TAYLOR M, SVENSON K, LIU Y, HSIAO CL, WOLFROM S, MAGUIRE M, MALAKIAN K, TELLIEZ JB, LIN LL, KRIZ RW, SEEHRA J, SOMERS WS & STAHL ML (2003). Catalytically active MAP KAP kinase 2 structures in complex with staurosporine and ADP reveal differences with the autoinhibited enzyme. *Structure* 11(6):627-636.
- VAJDOS FF, ULTSCH M, SCHAFFER ML, DESHAYES KD, LIU J, SKELTON NJ & DE VOS AM (2001). Crystal Structure of Human Insulin-like Growth Factor-1: Detergent Binding Inhibits Binding Protein Interactions. *Biochemistry* 40(37):11022-11029.
- VAN DE WATERBEEMD H, SMITH DA & JONES BC (2001). Lipophilicity in PK design: methyl, ethyl, futile. *J Comput Aided Mol Des* 15(3):273-286.
- VAN REGENMORTEL M (1999). Molecular recognition in the post-reductionist era. *J Mol Recognit* 12:1-2.
- VEBER DF, JOHNSON SR, CHENG HY, SMITH BR, WARD KW & KOPPLE KD (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 45(12):2615-2623.
- VOET D & VOET JG (1992). Bioquímica. Barcelona, *Ediciones Omega, S.A.*
- WALENSKY LD (2004). Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science* 305(5689):1466-1470.
- WALTERS WP, GREEN J, WEISS JR & MURCKO MA (2011). What Do Medicinal Chemists Actually Make? A 50-Year Retrospective. *J Med Chem* 54(19):6405-6416.

- WANG R, FANG X, LU Y & WANG S (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* 47:2977-2980.
- WELLS J & MCCLENDON C (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450(7172):1001-1009.
- WHITTY A & KUMARAVEL G (2006). Between a rock and a hard place? *Nat Chem Biol* 2(3):112-118.
- WIESMANN C, BARR KJ, KUNG J, ZHU J, ERLANSON DA, SHEN W, FAHR BJ, ZHONG M, TAYLOR L, RANDAL M, MCDOWELL RS & HANSEN SK (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat Struct Mol Biol* 11(8):730-737.
- WILLIAMS DH, STEPHENS E, O'BRIEN DP & ZHOU M (2004). Understanding Noncovalent Interactions: Ligand Binding Energy and Catalytic Efficiency from Ligand-Induced Reductions in Motion within Receptors and Enzymes. *Angew Chem Int Ed Engl* 43(48):6596-6616.
- WISHART DS, KNOX C, GUO AC, EISNER R, YOUNG N, GAUTAM B, HAU DD, PSYCHOGIOS N, DONG E, BOUATRA S, MANDAL R, SINELNIKOV I, XIA J, JIA L, CRUZ JA, LIM E, SOBSEY CA, SHRIVASTAVA S, HUANG P, LIU P, FANG L, PENG J, FRADETTE R, CHENG D, TZUR D, CLEMENTS M, LEWIS A, DE SOUZA A, ZUNIGA A, DAWE M, XIONG Y, CLIVE D, GREINER R, NAZYROVA A, SHAYKHUTDINOV R, LI L, VOGEL HJ & FORSYTHE I (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 37(Database issue):D603-610.
- WLODAWER A & VONDRASEK J (1998). Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* 27:249-284.
- WONG-HAWKES SYF, MATTEO JC, WARRINGTON BH & WHITE JD (2007). Microreactors as new tools for drug discovery and development. *Ernst Schering Found Symp Proc* 3:39-55.
- WU J, ADOMAT JM, RIDKY TW, LOUIS JM, LEIS J, HARRISON RW & WEBER IT (1998). Structural basis for specificity of retroviral proteases. *Biochemistry* 37(13):4518-4526.
- WU JP, WANG J, ABEYWARDANE A, ANDERSEN D, EMMANUEL M, GAUTSCHI E, GOLDBERG DR, KASHEM MA, LUKAS S, MAO W, MARTIN L, MORWICK T, MOSS N, PARGELLIS C, PATEL UR, PATNAUDE L, PEET GW, SKOW D, SNOW RJ, WARD Y, WERNEBURG B & WHITE A (2007). The discovery of carboline analogs as potent MAPKAP-K2 inhibitors. *Bioorg Med Chem Lett* 17(16):4664-4669.
- YAN C, WU F, JERNIGAN RL, DOBBS D & HONAVAR V (2008). Characterization of protein-protein interfaces. *Protein J* 27(1):59-70.

- YU EW, AIRES JR, MCDERMOTT G & NIKAIIDO H (2005). A Periplasmic Drug-Binding Site of the AcrB Multidrug Efflux Pump: a Crystallographic and Site-Directed Mutagenesis Study. *J Bacteriol* 187(19):6804-6815.
- ZHANG Z, LI Y, LIN B, SCHROEDER M & HUANG B (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* 27(15):2083-2088.
- ZHOU P, TIAN F, LV F & SHANG Z (2009). Geometric characteristics of hydrogen bonds involving sulfur atoms in proteins. *Proteins* 76(1):151-163.
- ZHU H, DOMINGUES FS, SOMMER I & LENGAUER T (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 7:27.