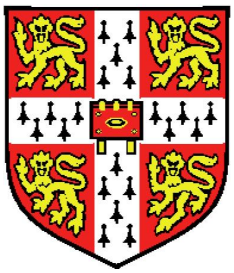
Methods for the Investigation of Protein-Ligand Complexes



Benjamin H Stauch

Robinson College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

June 2013

Meinen Eltern.

"What we see before us is just one tiny part of the world. We get in the habit of thinking, this is the world, but that's not true at all. The real world is a much darker and deeper place than this, and much of it is occupied by jellyfish and things. We just happen to forget all that. Don't you agree? Two-thirds of earth's surface is ocean, and all we can see with the naked eye is the surface: the skin."

Haruki Murakami
The Wind-Up Bird Chronicle.

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."

Douglas Adams
The Long Dark Tea-Time of the Soul.

Declaration

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

This thesis does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using LaTeX according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Acknowledgements

This Ph.D. has been a long and strange journey. I owe my gratitude to many people I met on the way, and to those who were with me from the start. Although this section is therefore rather long, it is also necessarily incomplete.

First and foremost, I would like to thank my supervisor John Overington for his leap of faith of accepting me into his group at the EBI, providing me with ideas, inspiration and criticism. Your constant support and advice has made this thesis possible.

Daniel Nietlispach is acknowledged for his willingness to provide supervision from the side of Cambridge University, and support in and understanding of complicated issues.

I owe my deepest gratidude to Helke Hillebrand, Iain Mattaj, and the EMBL International Ph.D. Programme for funding and solving minor and major problems, making me feel like nothing can go wrong.

Michele Cianci is thanked for his exceptional commitment, patience and support during most of the experimental work presented in this thesis. Thanks to your hospitality, my numerous trips to Hamburg have never felt like work, even though we have been clocking long hours.

I must thank Bernd Simon for his support and expertise during my time in Heidelberg. It was an absolute pleasure working with you, and I like to think that you are the prime example of how EMBL really is held together.

The members of my thesis advisory committee, Maja Köhn and Gerard Kleywegt, are gratefully acknowledged for their duties during my Ph.D., providing constant input and hands-on support.

Teresa Carlomagno is acknowledged for accepting me into the EMBL and her supervision during my time at the EMBL Heidelberg.

Importantly, my parents and my family, who were essential to me during my entire education (and before), trusting my judgment and supporting me in every which way. This is for you.

Sophie Shephard has been a great support through many of the challenges on my way, and I am sure at this stage, together with my parents and John, she will be among the five most relieved people this is coming to a conclusion.

Daniel Mende and Christina Mertens have been great friends for many years. It means a lot to me to know I can always rely on you.

Felix Krüger and Nenad Bartoniček have been essential in making me feel at home in Cambridge, and I never had to ask them twice to help me out one of the many times I did. You guys are awesome! Also thanks to you, John and Sophie for proof-reading of this thesis.

My housemates Benedetta Baldi and Tamara Steijger are credited for enduring me especially during the last phase of this thesis.

Frank Thommen is acknowledged for facilitating many last-minute, non-standard, and semi-official solutions, and work-arounds.

Adriana Rullmann has been a most exceptional teacher and inspiration early during my education. Without you I would probably not be here. Thanks for encouraging me to take on slightly thicker 'boards'.

Julien Orts, Irene Amata, John Kirpatrick, Domenico Sanfelice, Thierry Rohmer, Audronė Lapinaitė, Carmen Fernandez, Teodora Basile, Magdalena Rakwalska, and Stephen Fullerton are acknowledged for being great colleagues and friends during my time in Heidelberg. You made it worth it.

Patrícia Bento, Sam Croset, Gerard van Westen, Louisa Bellis, Shaun McGlinchey, George Papadatos, Rita Santos, and the entire ChEMBL team for their welcoming me and help to settle in to Cambridge and the Genome Campus. It has been a pleasure working with you.

Henning Hofmann, Jörg Zielonka, and Carten Münk are thanked for giving me confidence and encouraging me to pursue my ambitions, and providing me with a great environment during my master's thesis.

I furthermore thank Teresa Tiffert and Robinson College, Cambridge for their support during my graduate studies here.

Sander Timmer, Christine Seeliger and Steven Wilder are gratefully ackonwledged for the banter, the long hours on the ergs, the rare hours in the sun, and contributing their part in keeping me going.

I must thank Ângela Gonçalves and Catalina Schwalie for their willingness to casually overlook initial dissents.

Tracey Andrew and Milanka Stoijkovic are acknowledged as representatives of the Ph.D. Programme. Their commitment and interest in their 'subjects' has been exceptional.

I am grateful to Pedro Ballester and Michael Menden for their willingness to discuss machine learning related matters.

The Cambridge University Cycling Club, and especially Dan Ahearn, Sarah Gallagher, Jenny Haskell, and Ming-Chee Chung, are acknowledged for getting me away from the screen and bench, or trying to.

Konrad Rudolph is thanked for last-minute support with LaTeX typesetting.

I want to explicitly thank my friends Karina Mayer, Jonas Brand, Isa Serma, Patrick Schubert, Swetlana Derksen, and Sina Muhler for keeping in touch, visiting, or providing accommodation during my own visits back home.

Thomas Weisswange, Martin Weisel, Tim Geppert, Tim Werner, Gosia Drwal, Daniel Gottstein, Felix Reisen, and Frederik Hefke deserve a mention, too.

Last but not least, I owe my gratitude to countless people at the EMBL Heidelberg, the EBI, and Cambridge University for making my time in Cambridge and Heidelberg into an experience I will keep dearly.

Abstract

The interaction of small molecules with biological receptors is fundamental for cellular processes and plays a central role in disease and homeostasis. In this thesis, two multi-disciplinary methods aiding the investigation of protein-ligand interactions are presented that make use of data obtained by diverse experimental techniques, Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography.

In the first part of the thesis, an existing NMR-based method for the determination of ligand binding modes, termed INPHARMA, is improved by incorporating a rigorous model of protein internal motion using the NMR order parameter formalism. In this approach, indirect magnetisation transfer between two competitive ligands relayed by the receptor is simulated computationally and compared to experimental NMR data. This yields a ranking of binding poses, allowing for the true ligand orientations to be selected from a pool of possible complex structures. Using Protein Kinase A (PKA) as a test system, and order parameters extracted from Molecular Dynamics simulations of two PKA-ligand complexes, it is demonstrated that the overestimation of the magnetisation transfer that had been observed when *not* using a motional model, can be corrected for. Additionally, a 'generic' statistical characterisation of protein motional behaviour is obtained by investigating other, unrelated proteins, and applied to increase the ability of the method to discriminate true from decoy ligand orientations in the PKA system.

In the second part, the empirical notion that the noble gas xenon tends to bind to ligand binding site of proteins is investigated systematically. It is found that xenon is over-represented in protein ligand binding sites of known structures within the Protein Data Bank, showing dispersive interactions with aliphatic and aromatic protein moieties. A knowledge-based interaction potential for xenon is developed and used to construct a 'xenon likeness score' that can discriminate xenon from water molecules and ions. The performance of a classifier based on this score is examined by cross-validation using common performance measures from the field of machine learning, indicating high discrimination power. In a prospective application, xenon binding to the N-terminal domain (NTD) of the molecular chaperone Hsp90 is predicted, and for validation, the X-ray structure of a xenon derivative of Hsp90-NTD is solved, demonstrating that xenon binding sites can be predicted with good accuracy.

These findings have implications for fragment-based drug discovery, increasing both the throughput and accuracy of the INPHARMA method that is especially suited for fragment-like, weakly binding ligands, and suggesting the usefulness of adding xenon to standard fragment libraries.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Protein-ligand interactions are central to biology. Proteins have evolved to catalyse chemical reactions many orders of magnitude faster and more efficiently than those achievable by typical artificial chemical catalysts, and small molecule ligands are furthermore involved in processes as diverse as the regulation of gene transcription, the communication between synapses, and metabolic processes in the mitochondrial respiratory chain. Engineered small molecules provide a therapeutically tractable way of tweaking and manipulating the interplay and assembly of proteins that form cells, and cells that form organs and organisms, in order to address and restore normative phenotypes. In this therapeutic setting, orphan targets, pathogen adaption and evolution of resistence requires a constant supply of novel therapeutic agents. The advent of routine biophysical techniques for structure guided drug design and high throughput ligand centred methods, aided by chemoinformatics approaches, have resulted in a rational approach to drug design that attempts to supply novel therapeutics and to rival historical, serendipitous discoveries such as those of opiates, penicillin and aspirin.

In recent decades, important progress has been made in the field of translational molecular biomedicine, e.g. effectively rendering HIV infections a chronic disease efficiently controlled by cocktails of antiretroviral drugs, and being able to target more and more different types of cancer specifically by chemotherapeutic agents. In a scenario where a fundamental limitation of the number of tractable drug targets exists, novel drug design approaches are constantly sought after, addressing

the high attrition rates in drug design campaigns, resulting in the high costs of drug development that translate into a financial burden, or the sub-optimal therapeutic treatment of patients. Improving the throughput of existing methods and establishing novel methods providing complementary information about protein-ligand interactions continues to be a key challenge in modern multi-disciplinary life science that will allow crucial global health challenges to be addressed in the future, such as neuro-degenerative diseases, or emerging multi-drug-resistant bacterial strains.

In the following, an introduction to the concept of *fragment based drug design* (FBDD) is given (Section 1.1). This drug design approach is of crucial interest for both of the computational tools aiding drug discovery that are developed throughout this thesis. After that, two protein systems will be reviewed, both of which are implicated in cancer and molecular targeted cancer chemotherapy and have played a central role in the *proof-of-principle* of FBDD. Firstly, the cAMP-dependent protein kinase (PKA) is described as a prototypical protein kinase (Section 1.2). In Chapter 2 of this thesis it was used to study protein-ligand interactions using Nuclear Magnetic Resonance (NMR) spectroscopy. Section 1.3 describes the molecular chaperone Hsp90 which was used to predict xenon binding sites using a technique that was devised as part of this thesis and is described in Chapter 3.

1.1 Fragment based drug design

In recent years, a paradigm shift has started to take place in drug design campaigns, away from brute force, high throughput screening (HTS) of vast numbers of diverse drug-like small molecules to detect biological activity, towards design intensive approaches that start with lead-like molecular fragments and iteratively expand them into potent compounds (Jencks, 1981; Shuker et al., 1996; Erlanson, 2012). This *fragment based drug design* (FBDD) approach has been facilitated by advances in structural biology and biophysics, particularly by improvements in throughput and sensitivity of biophysical screening methods (Davies and Tickle, 2012). As opposed to HTS approaches that aim at identifying potent inhibitors as early as possible during a drug design campaign, in FBDD, smaller, weakly

binding fragments are identified. This has multiple advantages over traditional approaches, as will be outlined in this section. FBDD is now widely applied to a plethora of molecular targets and indications (Hajduk and Greer, 2007; de Kloe et al., 2009; Murray and Blundell, 2010; Baker, 2013), including traditionally challenging systems such as G-protein coupled receptors (GPCRs) and protein-protein interactions (Valkov et al., 2012). FBDD has very recently played a central role in the development of vemurafenib (PLX4032) (Tsai et al., 2008; Bollag et al., 2010; Flaherty et al., 2010; Bollag et al., 2012), a kinase inhibitor (see section 1.2), in 2011.

The space of drug-like molecules with a molecular weight of 300 to 500 Da has been estimated to exceed 10^{60} molecules (Bohacek et al., 1996). It becomes immediately apparent that this chemical space cannot be sampled exhaustively, with the size of the worldwide collection of isolated small molecules estimated to be 10^8 (Hann and Oprea, 2004), and the size of a typical HTS compound library around 10^5 to 10^6 molecules. On the other hand, for smaller molecules, the search space can be reduced dramatically (Fink et al., 2005): there are just over 10^8 molecules with up to 11 non-hydrogen atoms (Fink and Reymond, 2007), and less than 10^9 with up to 13 non-hydrogen atoms (Blum and Reymond, 2009). Among these chemically tractable molecules there is a smaller fraction (an estimated 10^7 and 10^8, respectively) with favourable, 'lead-like' molecular properties as formalised by the 'Rule of Three' (Congreve et al., 2003): molecular weight (MW) less than 300 Da, not more than 3 hydrogen bond donors, not more than 3 hydrogen bond acceptors, ClogP not more than 3, not more than 3 rotatable bonds, and a polar surface area of 60 Å^2 or less.[1] This rule has been defined in analogy to the familiar 'Lipinski's Rule of Five' for orally bioavailable drug-like molecules (Lipinski et al., 2001) which requires a MW of less than 500 Da, not more than 5 hydrogen bond donors, not more than 10 hydrogen bond acceptors, and a logP of 5 or less. Therefore, 'fragment space' can be screened much more efficiently by small compound collections of typically around 10^3 molecules, allowing starting points for lead optimisation to be determined very rapidly, at lower cost and largely indepen-

[1] it is useful to further establish a lower bound on the MW, around 150 Da, to limit multiple or ambiguous binding modes of very small fragment ligands.

dent of prior knowledge about ligand requirements of the molecular target. This makes FBDD applicable to novel and challenging targets and especially appealing in academic settings (Scott et al., 2012). Additionally, in general higher hit rates can be expected from fragment screening than from traditional approaches such as HTS, as for molecules being extended by medicinal chemistry, each newly introduced moiety has an increasing probability of interfering with binding (Hann et al., 2001). Fragments therefore probe both ligand binding sites and chemical space more efficiently.

FBDD campaigns comprise three stages: library design, fragment screening, and fragment elaboration (Scott et al., 2012). Fragment libraries can be assembled either for diversity, or tailored to specific targets, incorporating prior knowledge. For screening, typically high compound concentrations (0.1 to 10 mM) are used, and an array of complementary and orthogonal biophysical screening techniques exists. Some of these techniques provide detailed atomic models of the protein-ligand interaction, facilitating rational drug design and modification by medicinal chemistry. After a brief characterisation of library design and fragment elaboration approaches the remainder of this section will focus on the most important biophysical screening and hit validation techniques in FBDD, providing a setting for the methodologies developed throughout this thesis.

1.1.1 Fragment library design and fragment elaboration

Screening library design allows to incorporate different levels of prior knowledge into the drug design campaign. In particular, general purpose libraries containing a diverse set of molecules can be used to detect new chemotypes. These general purpose libraries are designed to cover chemical space efficiently and are suitable for screening against a diverse range of targets. One example of such a library is the Core Fragment Set used by Astex, making use of scaffolds and functional groups commonly found in drug molecules (Bemis and Murcko, 1996, 1999; Hartshorn et al., 2005). It contains about 1,000 fragments that have been selected to cover chemical space uniformly, out of a larger compound collection, using clustering by topological fingerprints (Davies and Tickle, 2012). Another noteworthy example

is the 'Fragments of Life' collection that contains fragments of natural metabolites and peptide motif mimetics (Davies et al., 2009). If knowledge about the target is available, such as existing drugs or previous successful lead compounds, a focused library can be built from fragmenting larger compound databases. These can be pre-filtered, e.g. by molecular docking (Kitchen et al., 2004). Often, focused libraries are built around initial hits from primary biophysical screens.

Ideally, fragment libraries already exhibit favourable molecular properties for further optimisation, i.e. they are Rule of Three compliant. Fragment hits out of the biophysical screens can then be ranked according to their affinity for the target, if the respective biophysical technique allows for the determination of this property (see below). An additional criterion for the selection of promising lead compounds is represented by the *ligand efficiency* (LE) concept (Hopkins et al., 2004), which suggests that fragments should be ranked by the ratio of their binding free energy and non-hydrogen atom count.[2] The metric is based on the observation that the incorporation of larger lipophilic groups into molecules is the easiest way of improving their potency. This has lead to 'molecular obesity' (Hann, 2011), the downside of potency driven optimisation in medicinal chemistry. It is tempting to achieve high potency leads early on in the optimisation process, but often this leads to dead ends later in the optimisation process, as it can result in insoluble, aggregating compounds which are more likely to non-specifically inhibit their targets, and later on display poor bio-availability parameters in cell based or *in vivo* assays. It is therefore important to establish how *efficiently* a fragment makes use of added molecular weight and lipophilic character (Hopkins et al., 2004). In general, the most promising fragment elaboration strategy is to select for specific binding early on, which is best achieved through hydrogen bonds to the protein by polar groups of fragments, and later on add hydrophobic groups, if necessary, and tolerated.

Retrospective analysis has revealed that typically, ligand efficiency is not improved during optimisation. Therefore, in order to obtain a drug with K_d of 10 nM and a MW of 500 Da, a LE index of 0.29 kcal mol^{-1}/non-hydrogen atom would

[2] an alternative, related metric, the *binding efficiency* index (BEI), makes use of the $\mathrm{pK_i}$ or $\mathrm{pIC_{50}}$ instead of the binding free energy (Abad-Zapatero and Metz, 2005), among many others, such as the ligand-lipophilicity efficiency index (Ryckmans et al., 2009).

be necessary. Thus, a commonly set target for the initial screening is to obtain a ligand efficiency score of at least 0.3 for the fragments to be followed up on as a promising compound. It is found that typically, binding efficiency is higher for fragments than for HTS hits, but they are less specific for their target and more promiscuous binders. However, promiscuity of fragments usually does not pose a severe problem to the design process as selectivity is improved during lead optimisation (Hennig et al., 2012).

After obtaining initial hits in fragment screening, computational hit expansion can be carried out, employing concepts such as molecular similarity or pharmacophore[3] searches (Hennig et al., 2012). In order to establish pharmacophores with high confidence, it is beneficial to obtain the binding modes of the fragments within the protein ligand binding pocket, allowing fragment elaboration to conclude in a rational, structure-guided way (Murray and Blundell, 2010). This requires detailed visualisation of the binding mode by molecular models with atomic resolution. The reference technique to obtain high confidence atomic models is X-ray crystallography, but also NMR spectroscopy can provide useful structural information (see below). Importantly, the NMR based methodology described in Chapter 2 of this thesis, termed INPHARMA, provides relative or absolute ligand binding modes equivalent to atomic resolution.

It has been found that fragments most commonly bind to 'hot spot' regions of proteins that serve as anchoring points from which selective ligands can be constructed (see also section 1.1.3) (Hann et al., 2001; Hajduk et al., 2005; Ciulli et al., 2006). For fragment elaboration, three basic concepts have been described: fragment *linking*, -*growing*, and -*merging*. Fragment *linking* (Shuker et al., 1996) describes the concept of introducing a linker region between two fragments that in the screen were found to bind independently and simultaneously at two different sites within the protein ligand binding site. It is the most appealing concept from a theoretical standpoint as in joining two fragments, there is a gain in *Gibbs connection energy* because the rigid body entropy penalty that has to be overcome is largely independent of the molecular weight of the compounds involved, and so the

[3]a *pharmacophore* is, according to the International Union of Pure and Applied Chemistry (IUPAC), defined as the 'ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response' (http://dx.doi.org/10.1351/pac199870051129).

linked compound only experiences a single rigid body penalty term, which is less than the sum for two fragments (Jencks, 1981; Rognan, 2012). Therefore, joining two non-competitive fragments with independent micromolar affinity can in principle yield a nanomolar lead compound. However, fragment linking is challenging in practice due to strict geometric requirements imposed by allowed bond lengths and angles. Strain on the linker region joining the fragments can lead to loss of potency or imperfect positioning of the fused fragments (Chung et al., 2009). In practice, a *growing* strategy appears to prevail more often, in which the initial fragment is progressively expanded within the pocket to entertain additional interactions with the protein. Typically, the initial fragment maintains its position and orientation while moieties are added (Erlanson et al., 2000). Alternatively, a *merging* strategy can be pursued, where a functional moiety is borrowed from one chemical series and transplanted onto another, in order to generate a novel molecule (Huth et al., 2007).

For the first drug approved by the FDA that was discovered and optimised using a FBDD strategy, the kinase inhibitor vemurafenib (see section 1.2), a fragment growing strategy was employed (Bollag et al., 2010; Flaherty et al., 2010).

1.1.2 Biophysical techniques for fragment identification

In FBDD, issues of solubility for screening compounds are aggravated by the high compound concentrations necessary (around 1 mM). Non-specific inhibition or aggregation can occur and lead to false positives, depending on the screening method used. Routinely, small quantities of organic solvents such as DMSO can be used to improve solubility, and light scattering techniques can be employed to check solubility (Erlanson, 2012). Different biophysical techniques are prone to different artifacts, and at the same time provide complementary information about the binding event. For example, X-ray crystallography is the only technique that provides detailed atomic resolution binding mode data in standard applications, while quantitative affinity data can be provided by isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), and NMR spectroscopy (see below). At the same time, throughput and material requirements imposed by the target systems can differ dramatically between different techniques, influencing the stage

of the FBDD campaign they are best suited for.

It is useful to discriminate primary from secondary screening techniques, the former ideally having high throughput and not requiring large amounts of target protein, and the latter having lower throughput, but providing more detailed information such as the atomic binding mode to be used for hit conformation. Traditionally, X-ray crystallography and ITC are low throughput techniques, while SPR and thermal shift are higher throughput. Depending on whether *ligand-* or *target-detected* techniques are used, NMR spectroscopy can be relatively high or low in throughput, and potentially also provide an atomic model of binding mode. It has been suggested that NMR spectroscopy is probably the experimental technique least prone to unspecific binding (Hennig et al., 2012). The role of X-ray crystallography in the screening process, on the other hand, is changing due to recent developments (see below) - its throughput is steadily increasing, making it fit for application as primary screening method if robust protein crystals can be obtained reliably (Davies and Tickle, 2012).

Given their different individual strengths and limitations, it is beneficial to use several orthogonal screening methods in conjunction with each other. NMR spectroscopy and X-ray crystallography play a central role as structural biology techniques in FBDD, with the former standing out as an extremely versatile tool to obtain different levels of information on a ligand binding event under solution conditions, and the latter the reference technique for obtaining detailed structural information at atomic level, including robust inter-atomic distances. Chapter 2 of this thesis describes the improvement of an NMR-based technique to obtain detailed ligand binding modes, addressing a classic shortcoming of NMR-based techniques for the investigation of ligand binding events by closing a gap between target- and ligand-detected techniques (see below), while Chapter 3 and 4 present the development of a computational technique that relies heavily on information obtained by X-ray crystallography. Therefore, this introduction discusses both the application of NMR spectroscopy and X-ray crystallography to FBDD in greater detail. In addition, a short overview of other biophysical techniques is given.

1.1.2.1 NMR spectroscopy

NMR spectroscopy was the first experimental technique used for FBDD screening (Wyss et al., 2012). In the *SAR by NMR* methodology (Shuker et al., 1996), changes in protein chemical shift in the presence of a fragment indicate binding, and the ligand binding epitope of the protein can be mapped if protein resonances assignments are available. Chemical shift changes are a result of the strong dependency of chemical shifts on the electronic environment of the nucleus under observation. Therefore, if a ligand associates with a protein moiety, it will influence the chemical shifts of nearby protein moieties in a way that can be measured by NMR spectroscopy. In NMR based FBDD screening experiments, ligand cocktails consisting of multiple ligands can be used to reduce the number of experiments and duration of the screening campaign. In this case, ligands should be *chemical shift encoded*, i.e. possess largely orthogonal sets of non-degenerate chemical shifts so that ligands that bind to the target can be uniquely identified. SAR by NMR is an example of a *target-detected* method (see below), and as such limited to relatively small protein systems with a maximum MW of about 30 to 40 kDa, and large amounts of isotope-labelled protein are required. Alternative to target-detected methods, *ligand-detected* techniques can be used such as saturation transfer difference (STD) (Mayer and Meyer, 2001), WaterLOGSY (Dalvit et al., 2001; Dalvit, 2009), or target-immobilised NMR screening (TINS) (Vanwetswinkel et al., 2005).

In general, *target-detected techniques* provide information about the ligand binding site, and in favourable cases, structural information about the ligand binding mode (Carlomagno, 2005). Typically, *heteronuclear single quantum coherence* (HSQC) spectroscopy experiments are performed in the presence and absence of the ligand, correlating ^{1}H and ^{15}N resonances in a two dimensional NMR spectrum, and the chemical shift perturbations that are observed can be mapped to the sequential protein resonance assignment, if it is available. This spatially highlights the protein ligand binding site. Affinity information in the form of a NMR-K_{d} can be obtained from ligand titration. However, two dimensional experiments are time consuming, and the isotope labelling required (^{15}N, and possibly ^{13}C) is expensive and prohibitive for some protein expression systems.

Ligand-detected techniques, on the other hand, require less protein, no isotope

labelling, and are applicable to larger protein systems. In fact systems with larger proteins yield stronger signal. However, ligand-detected techniques do not provide information about the ligand binding epitope of the protein, as the protein takes a passive role in the NMR experiment and is therefore not visible. Thus, traditional ligand-detected experiments do not provide information about ligand binding modes. This shortcoming is addressed by the INPHARMA methodology that is presented in Chapter 2 of this thesis. Ligand-detected techniques require the ligands to be in *fast exchange* with the protein. This means that on the NMR timescale that is defined by chemical shift differences that can be expressed in Hertz (Hz), binding events need to be *fast*, i.e. larger than about 100 Hz. In terms of affinity, this usually translates to a K_d value of above 10 μM, a requirement that is typically fulfilled in FBDD settings. Shift differences measured in Hz, or *parts per million* (ppm) difference to a reference compound, naturally relate to chemical rates measured in s^{-1}.

STD and WaterLOGSY are examples of ligand-detected NMR techniques. In these methods, an NMR irradiation pulse is applied either to the bulk water resonance (in case of WaterLOGSY), or some protein resonances (for STD), and results in a measurable transfer of magnetisation to the fragment in case of a binding event. If this binding event is fast enough and the ligand is present in large excess to the protein, target saturation is achieved and signal can be observed on ligand resonances that correspond to those of the free state, as free and bound ligand states are constantly and quickly exchanging. Thus, even in the presence of large proteins, narrow ligand peaks are observed, and resonance frequencies lie close to those of the free ligand. This is especially beneficial as it implies that ligand resonance assignment can be performed in the absence of the protein receptor. Yet other ligand-detected approaches make use of the change of translational and rotational mobility of the ligand that influence its relaxation properties upon binding ('relaxation dependent experiments'). A ligand in its protein bound state displays longer average rotational correlation time, and therefore has a shortened transverse relaxation time, due to its more efficient relaxation. This results in measurable line broadening in the spectrum of the bound ligand, in a protein size dependent manner (Carlomagno, 2005).

All ligand-detected experiments mentioned so far can be performed as one-

dimensional proton detected experiments and are therefore quick and sensitive compared to the two-dimensional target-detected experiments mentioned above. However, there is an added benefit in recording two dimensional spectra of relaxation dependent experiments. If ligand-detected NOESY experiments are recorded, intramolecular magnetisation transfer through space between ligand protons is observed. The efficiency of this transfer depends most strongly on the relaxation properties of the spin system. In its bound state, the ligand adopts some relaxation properties of the complex; most importantly, its effective rotational correlation time is the average of its own rotational correlation time and that of the protein, weighted by their relative concentrations. As the rotational correlation time of a molecule largely depends on its molecular weight, it is much larger for a protein than for a small molecule, and upon binding the relaxation properties of the ligand are strongly affected. This difference can be measured and quantified by NMR spectroscopy. In the absence of binding, the ligand has positive or zero cross peak intensities in a two dimensional proton-proton NOESY spectrum, indicating inefficient magnetisation transfer between ligand protons. Upon binding, however, the signs of the cross peaks are inverted with respect to the diagonal peaks, and their intensity gets much stronger. Cross peak intensities are strongly distance dependent;[4] therefore, from a two dimensional NOESY experiment of a ligand in the presence of a receptor it binds to in fast exchange, the bioactive, bound ligand conformation can be deduced. However, its absolute binding mode cannot be deduced by standard methods, as the protein is not observed in ligand-detected experiments (see discussion of the INPHARMA methodology in Chapter 2).

1.1.2.2 X-ray crystallography

X-ray crystallography is in many respects the most natural approach because it yields direct information and interpretable binding modes at atomic resolution (Jhoti et al., 2007; Davies and Tickle, 2012). Most frequently, protein crystals are soaked with mother liquor containing high concentrations (commonly around 50 mM) of the test fragment, and hits are deduced from observing fragments binding to the protein in the electron density. Due to its direct nature, no false positives

[4] they scale with the internuclear distance raised to the power of -6.

occur, but false negatives can appear due to protein ligand binding sites getting occluded by crystal contacts, or ligand binding requiring a protein conformational change that is not allowed within the crystal framework. The latter problem can be circumvented by attempting co-crystallisation of the protein in the presence of the ligand, at the expense of not being able to pre-produce batches of protein crystals prior to screening. The technique is also applicable to large proteins and affinity regimes of up to $K_\mathrm{d} \gtrsim 5$ mM and yields high resolution data and the most detailed information possible for well-behaved protein systems. While it can be regarded the gold standard for final hit validation, its perceived low throughput prevents it from being widely used a a primary screening tool.

However, in recent years, throughput of the method has increased markedly (Davies and Tickle, 2012), making use of automated, rapid data collection at powerful synchrotron beam sources that allow the collection of high-resolution data in minutes, employing sample changing robots, and semi-automatic processing and structure solution.[5] *Molecular replacement strategies* (discussed in more detail in the introduction of Chapter 3, sections 3.1.1 and 3.1.2) can essentially reduce structure solution to manual inspection of ligand placements in difference electron maps. The ligand orientations are obtained by ligand fitting routines, employing similar strategies to those used in *molecular docking* (Kitchen et al., 2004), but maximising the fit of the ligand with the experimental electron density. If multiple ligands are used during the same soaking experiment, it is important to use a set of *diverse* ligands within the same cocktail to make them uniquely identifiable from the shape of their electron density, and minimising the chances of two or more ligands within the same cocktail competing for the same binding site. X-ray crystallography based screening or hit validation requires the protein of interest to crystallise reproducibly, and ligands binding the crystals under isomorphic conditions. Solubility issues affecting the ligands can be addressed by using low concentrations of organic co-solvents, such as 1 to 10 % DMSO. Unlike other methods (NMR spectrocopy, SPR, ITC), X-ray crystallography does not yield information about binding affinity, so it requires other experimental methods to determine this important quantity. Ideally, X-ray crystallography is used

[5] routine acquisition of 100 datasets within 24 hours has been reported (Davies and Tickle, 2012).

in conjunction with a method with higher throughput like SPR that also provides affinity/rate information, or verified independently after crystallisation using ITC.

1.1.2.3 Surface plasmon resonance

In typical surface plasmon resonance (SPR) studies (Johnsson et al., 1991), proteins are immobilised on a metal coated surface and ligands flow past. When a ligand binds to the protein, it causes changes in the refractivity/reflectivity properties of the metal that are related to the mass of the ligand and the mass of the protein that can be detected by an optical device. SPR is well suited as a primary screening tool, and screening campagins are rapid and straightforward to set up. It has the additional advantage that very small amounts of protein are needed, unlike for ITC (see below), and distinctively, affinity data can be obtained: as the recorded sensogramm is time dependent and the approach represents a continuous flow system, ligands first saturate the protein and then wash off, and hence, k_{on} and k_{off} can be determined. SPR allows for the screening of several thousand compounds in a few days and is therefore ideally suited for prioritising subsequent X-ray experiments (Hennig et al., 2012).

1.1.2.4 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) measures heat release upon ligand binding (Ward and Holdgate, 2001; Holdgate et al., 2013). It is the technique of choice for precise determination of binding constants, and the only widely used biophysical technique that is able to deconvolute the contributions of *enthalpy* and *entropy* to ligand binding. The relative contributions of enthalpy and entropy can provide with insights into the relative importance of polar and hydrophobic interactions, respectively, and information about their relative magnitude can be very valuable for the fragment expansion process, as discussed in the context of ligand efficiency (LE) above. One drawback of the method is that it requires relatively large amounts of protein and has rather low throughput, so it is not suited as a primary screening method.

1.1.2.5 Thermal shift assay

Small molecule ligands typically bind to proteins in a well-defined, folded state, thereby stabilising this folded state by increasing its heat capacity. In a thermal shift assay (Niesen et al., 2007; Schulz and Hubbard, 2009; Kranz and Schalk-Hihi, 2011), the protein unfolding temperature is determined in the presence and absence of a ligand by optical means, as folded and unfolded proteins have different fluorescent properties. In the case of ligand binding, an increase in unfolding temperature can be measured. Thermal shift assays are well suited to determine the presence or absence of binding with high throughput, and are mainly utilised as an enrichment process before a secondary screening method is applied.

1.1.3 Computational techniques for fragment scanning

Computational alternatives to experimental methods attract interest for their ability to save resources, and their potential to provide complimentary information at relatively small expense. With the notion that fragments tend to bind hot spot regions of proteins, serving as anchoring points from which larger selective ligands can be constructed (Hann et al., 2001; Hajduk et al., 2005; Ciulli et al., 2006) (see above), the identification of these hot spot regions becomes a prerequisite for any FBDD campaign. In the early days of FBDD, it was noted that hot spot regions have the propensity to bind multiple types of solvent molecules. In the MSCS method (multiple solvent crystal structures), X-ray crystallography of proteins in a variety of organic solvents has been used to probe protein surfaces for complimentary binding sites, which are potential hot spot regions and ligand binding sites (Mattos and Ringe, 1996). Along these lines, NMR spectroscopy has been used to detect binding of solvent molecules with long residence times by observing NOE signals between solvent resonances and protons of the protein (Liepinsh and Otting, 1997).[6]

This solvent based approach can be considered to be parallel or complimentary to FBDD, as solvent molecules are considerably smaller than Rule of Three compliant fragments. Their much lower affinity can be expected to be balanced to some extent by their extremely high concentration. For instance, water is present

[6] the conceptual similarity to the WaterLOGSY approach discussed above should be noted.

at 55 M in solution, and a large number of ordered water molecules are found at discrete positions in every crystal structure. Using multiple solvents with diverse molecular properties and consensus site detection, ligand binding sites and their ability to interact specifically with different functional groups of organic ligand molecules can be detected. Due to their small size and extremely high abundance, solvent molecules are very efficient in sampling the protein crystals by diffusion.

In practice, with the improvements in highly sensitive biophysical techniques for FBDD, there is arguably little added benefit in solvent-based characterisation of proteins if they are already established as druggable targets, except for cases where individual solvent molecules are of functional or structural importance. However, due to their more universal properties, solvent molecules do have justified applications to establish druggablity, especially by computational means. Potential novel targets can be investigated for their likely ability to bind small molecule drugs by establishing the presence of hot spot regions computationally from the three dimensional coordinates of the protein (Rognan, 2012). Several approaches have been described in the literature to address this task, and could be applied to select targets with good success, in favourable cases being in good agreement with experimental findings.

The computational solvent mapping (CS-Map) technique, for instance, uses 14 different organic solvent molecules as computational probes to detect hot spots, and addresses the problem of nonspecific binding by identifying clusters of overlapping binding sites (Dennis et al., 2002). This is achieved by using an empirical energy function and a continuum electrostatic solvation model. Starting from probes positioned at points on the first water layer of the protein, an energy minimisation is performed using the program CHARMM (MacKerell et al., 1998), moving the probe molecules to favourable positions on the protein surface. Along similar lines, the FTMap approach uses a more efficient technique to sample the possible orientations of 16 types of solvent molecules (Brenke et al., 2009).

A different approach is implemented in the SILCS (site identification by ligand competitive saturation) method, where the protein is computationally immersed in high concentrations of an æqueous solution of three different probe molecules, benzene (1 M), propane (1 M), and water (at its experimental density) (Guvench and MacKerell, 2009). Using the CHARMM force field (MacKerell et al., 1998),

multiple runs of short molecular dynamics (MD) simulations are performed, generating a probability map of long residence water, benzene, and propane molecules. The unique feature of this approach is that solvation and protein plasticity are explicitly and rigorously represented, as side chain and loop motions are sampled on the nanosecond timescale. The number of different solvent molecules is restricted by the computational cost of the approach, but assumed to sufficiently represent the limited chemical diversity of amino acid side chains, as all important moieties of drug-like molecules are covered. Specifically, benzene is an important fragment as it is the most common aromatic group, and represents over 40 % of cyclic substructures in drug-like compounds (Kolb and Caflisch, 2006). Due to their low affinity (even when assuming a high ligand efficiency, as they have a very small number of atoms), solvent molecules exchange at the MD timescale, effectively rendering the approach a computational competition experiment.

A number of different biophysical and computational techniques to probe protein-ligand interactions has been discussed in the previous section. Traditionally, a number of protein systems exist to benchmark these in the context of FBDD and have provided proof of principle of the applicability of this drug design approach in industrial and academic settings. Two of these systems, both of which share their implication in cancer and cancer chemotherapy, will be presented in the following, having been used as test systems for the methodologies developed throughout this thesis.

1.2 Protein Kinase A

Protein kinases regulate their substrate proteins by attaching phosphate groups, which in turn act as recognition signals for other proteins. This post-translational modification alters the activation state of the target protein and can modify its shape and dynamics. The kinase catalyses the transfer of the γ-phosphate of adenosine triphosphate (ATP) to the hydroxyl group of serine, threonine or tyrosine of its protein substrate.

Kinases are in every aspect central to signal transduction cascades as they cou-

ple receptors of external stimuli[7] such as G-protein coupled receptors (GPCRs), to cellular effector proteins, thereby amplifying the signal and integrating signal networks (Hunter, 2000). The human genome encodes around 500 kinases (Manning et al., 2002). This renders them an attractive and important therapeutic target family (Cohen, 2002; Melnikova and Golden, 2004; Overington et al., 2006). Protein kinases themselves are stringently regulated and their deregulation is a key feature in several pathophysiological processes (Cohen, 2002; van Linden, 2013) such as oncology (Zhang et al., 2009), infectious disease (Doerig et al., 2005), as well as in processes in neurology (Chico et al., 2009), immunology (Cohen, 2009), and cardiology (Kumar et al., 2007).

The catalytic subunit of cyclic adenosine monophosphate (cAMP) dependent protein kinase, also known as Protein Kinase A (PKA),[8,9] is in many respects a prototype kinase (Taylor et al., 2012). It was the second regulatory protein kinase to be discovered (Walsh et al., 1968), and the first to be sequenced (Shoji et al., 1981) and crystallised (Knighton et al., 1991a).

Protein Kinase A integrates external signals by means of the *second messenger* cAMP (Tao et al., 1970; Brostrom et al., 1971) formed from an ATP precursor by the enzyme adenylyl cyclase, which itself is activated by signalling molecules through GPCRs. Depending on the cell type and tissue PKA is expressed, it fulfills different cellular functions, such as in glycogen and lipid metabolism (Walsh and Van Patten, 1994; Shabb, 2001). It achieves its functional versatility by modular design and combinatorial diversity (Taylor et al., 2012) of an *assembly* of proteins consisting of the enzyme itself, auxiliary proteins, interaction partners, and scaffolding proteins. In this assembly, the conserved catalytic core is linked to its regulators and substrates at a specific subcellular location (Taylor et al., 2004; Taylor and Kornev, 2011). Thus, a multi-protein functional unit, or 'signalling system' that senses intra-cellular levels of cAMP, is formed. PKA itself possesses

[7] in the form of *primary* messenger molecules such as hormones, peptides, or cytokines.

[8] Enzyme Commission number (EC number) (Bairoch, 2000) 2.7.11.11, Pfam identifier (Punta et al., 2012) PF00069, UniProt (UniProt Consortium, 2012) identifier P17612, ChEMBL (Gaulton et al., 2012; Bento et al., 2013) identifier CHEMBL4101, the latter two identifiers referring to the human protein.

[9] EC number key: 2.x, transferase; 2.7.x, transferring phosphorus containing groups; 2.7.11.x protein-serine/threonine kinases.

two stable phosphorylation sites (Thr^{197} and Ser^{338}) that are typically phosphorylated[10] due to the autocatalytic activity of the enzyme even when expressing it in *Escherichia coli* (Taylor et al., 2012). Furthermore PKA is resistant to its deactivation by phosphatases, rendering it responsive exclusively to cAMP (Kornev et al., 2008; Taylor and Kornev, 2011).

The PKA holo enzyme consists of two copies of the catalytic (C) and two copies of the regulatory subunit (R), forming a C_2/R_2 tetramer (Johnson et al., 2001; Taylor et al., 2012) (Figure 1.1). In absence of cAMP, the enzyme is sequestered as an inactive holo enzyme, whereas upon binding of cAMP to the regulatory subunit, the enzyme is activated by dissociation of the regulatory subunits (Figure 1.1) (Taylor et al., 2005).[11] Four different isoforms of the regulatory subunit have been described that control the activity of PKA depending on intracellular levels of cAMP (termed RIα, RIβ, RIIα, and RIIβ), differing mainly by their N-terminal linker amino acids containing the inhibitor site (see below) (Taylor et al., 2012).

1.2.1 Structure and interaction of the catalytic subunit

The catalytic subunit of PKA assumes a bilobal bean or kidney like shape and consists of an N-terminal (N-lobe, comprising amino acid residues 40 to 120) and a C-terminal lobe (C-lobe, residues 128 to 300). The smaller, more flexible N-lobe is dominated by a core formed of five anti parallel β-strands (Knighton et al., 1991a) but contains a single, functionally important α-helix termed 'αC-helix'. The N-lobe predominantly mediates the binding and positioning of the ATP nucleotide ligand. The C-lobe contains seven α-helices and serves as a docking surface for the substrate (Knighton et al., 1991b). The substrate cleft is positioned between the lobes, close to the active centre of the enzyme, facilitating phosphate group transfer (Figure 1.1).

The earliest structures of PKA were determiend in complex with the protein kinase inhibitor (PKI) peptide, $PKA/PKI^{5\text{-}24}$ (PDB identifier 1apm) (Knighton

[10] thereby activating the enzyme.

[11] figure modified after http://dx.doi.org/10.2210/rcsb_pdb/mom_2012_8, PDB Molecule of the Month, by David Goodsell. The model of the full complex was derived from structures with PDB identifiers 3tnp (Zhang et al., 2012), 1j3h (Akamine et al., 2003) (both *Mus musculus*, and 2h9r (Newlon et al., 2001) *Rattus norvegicus*. The regulatory subunits shown are of the RIIβ type (see below).

et al., 1993), and the ternary PKA/PKI$^{5\text{-}24}$/Mg^{2+}:ATP complex (PDB identifier 1atp) (Zheng et al., 1993b). In these structures, the inhibitor peptide both stabilises the catalytic subunit of PKA for crystallisation and provides a substitute for the part of the regulatory subunit that complexes the catalytic subunit, with structures of the catalytic/regulatory complexes within the same crystal not obtained until much later (see below). These first experimental kinase structures have provided a mechanistic understanding of the entire enzyme family as well as a structural basis for drug discovery. Since then, many more structures of PKA in complex with different ligands have been solved.[12]

PKI binds to the free catalytic subunit of PKA and acts as a potent competitive inhibitor (Demaille et al., 1977; Whitehouse and Walsh, 1983). The peptide constitutes the major inhibitory segment of PKI, a naturally occurring thermostable protein kinase inhibitor protein (Cheng et al., 1985, 1986), and has a K_i of around 2 nM (Dalton and Dewey, 2006). In complex with the catalytic subunit, its N-terminus adopts an amphipathic, α-helical conformation, with an extended C-terminus containing the inhibitor site (Knighton et al., 1991b). Both PKI and the regulatory subunit of cAMP-dependent protein kinase contain amino acid sequences that mimic the substrate, allowing them to bind the catalytic subunit, inhibiting its activity by blocking substrate access.[13] Recently, a complex of the catalytic subunit of cAMP-dependent protein kinase (C) and its regulatory subunit (RIIβ), forming a tetrameric holoenzyme C_2/RIIβ_2 comprising a dimer of hetero dimers, has been described (Zhang et al., 2012), and additional structures have been reported for dimers of C_1/RIα_1 (Kim et al., 2005, 2007).[14]

[12] as of May 2013, there are 142 X-ray structures in the PDB, including 74 bovine, 44 murine, and 19 human protein structures.

[13] PKI is a *pseudo* substrate as instead of the phosphorylation site (Ser or Thr) of a canonical substrate it possesses an alanine (Whitehouse and Walsh, 1982).

[14] PDB identifiers 3tnp, 3fhi and 2qcs, respectively.

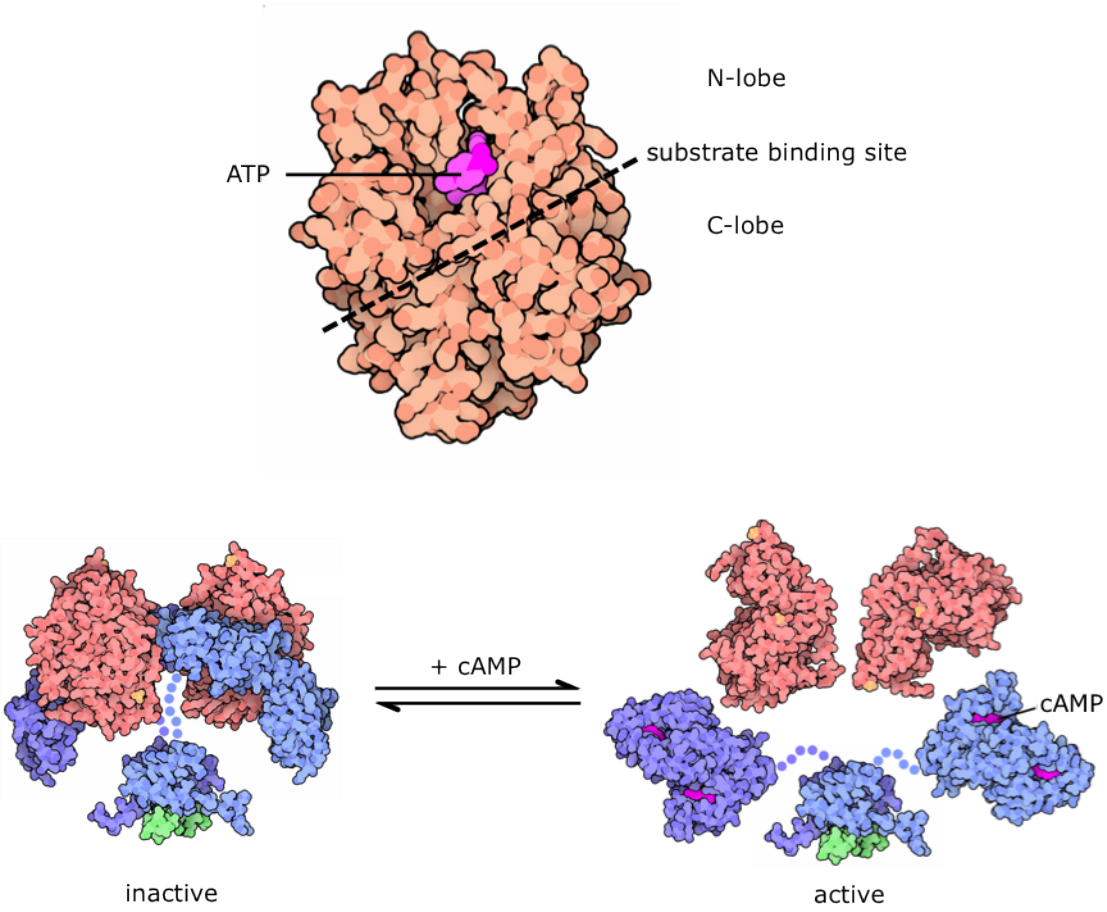


Figure 1.1 – Structure of the catalytic subunit and the holo enzyme of cAMP-dependent protein kinase (PKA). Top, monomer of the bilobal catalytic subunit (orange) of PKA, with ATP bound to the active centre contained in the N-lobe of the enzyme. The substrate recognition site located within the C-lobe is indicated by a dotted line. Bottom, the holo enzyme of PKA is formed by two copies each of the catalytic (C, red) and the regulatory (R, blue) subunit, forming a C_2/R_2 tetramer. Two phosphorylation sites within each catalytic subunit (Thr^{197} and Ser^{338}) are shown in orange. Upon binding of cAMP, the regulatory dissociate from the catalytic subunits, thereby releasing the activated enzymes. The flexible linker regions of the regulatory subunits are depicted as schematic blue chains, linking the N-terminal dimerisation/docking (D/D) domain of each protein to the C-terminal cyclic nucleotide binding (CNB) domains CNBA and CNBB. The inhibitor site competing with substrate of the catalytic subunit and thereby inhibiting its function is structured, and located in the C-terminal region of the linker. The D/D domain both interacts with the complex and mediates interaction with regulatory proteins such as *A kinase anchoring proteins* (AKAPs). AKAPs localise PKA to specific

sites in the cell, thereby creating micro-environments for PKA signalling. A peptide fragment of an AKAP interacting with the D/D domain (blue) is depicted in green.

PKA structure determination endeavours have shed light on the conformational changes the enzyme undergoes at different points in its catalytic cycle (Taylor et al., 2012). The inherent flexibility of the catalytic core allows the substrate cleft to open and close upon binding of ATP and substrate, and the regulatory subunit undergoes major conformational changes upon cAMP binding, causing it to disassociate from the core (Kim et al., 2007). Depending on its conformational state, the regulatory subunit either displays a high affinity towards cAMP or towards the catalytic subunit, rendering the PKA system a cAMP-dependent, dynamic molecular switch. The following section will focus on the conformational changes associated with activation and catalysis of the catalytic subunit of PKA.

1.2.2 Conformational changes during activation and catalysis

There are two kinds of large scale conformational rearrangement that the catalytic subunits of kinases undergo, the first one concerning the initial activation of the enzyme, and the second one the opening and closing of the enzyme during the catalytic cycle (Figure 1.2).

Kinases are tightly regulated. The most common type of activation is the phosphorylation of the kinase by another kinase, or autophosphorylation, but also pH-dependent activation mechanisms have been described (Shan et al., 2009). PKA is an example of a constitutively active enzyme, due to the autophosphorylation of Thr^{197} within its activation loop (Kornev et al., 2006) (amino acid residues 184 to 204), and resistance to dephosphorylation by phosphatases. Its activity is therefore exclusively controlled by the cellular levels of cAMP and the regulatory subunits. Other kinases, however, are dynamically activated and inactivated, and there are certain conformational changes associated with their activation state. In general, inactive kinase states are conformationally more diverse than active conformational states, so they pose an attractive target for the design of specific inhibitors (see below) (Taylor et al., 2004; Liu and Gray, 2006; Taylor and Kornev, 2011). Important drug targets such as the ABL and Src kinase assume an inac-

tive conformation where the substrate cleft is blocked by a segment of the kinase termed the activation loop (Huse and Kuriyan, 2002; Levinson et al., 2006). After phosphorylation, the activation loop changes its conformation and makes accessible the substrate cleft and active centre. This conformational change sometimes is inaptly described as an opening of the enzyme, and can therefore potentially be confused with the opening and closing of the enzyme during the catalytic cycle (see below).

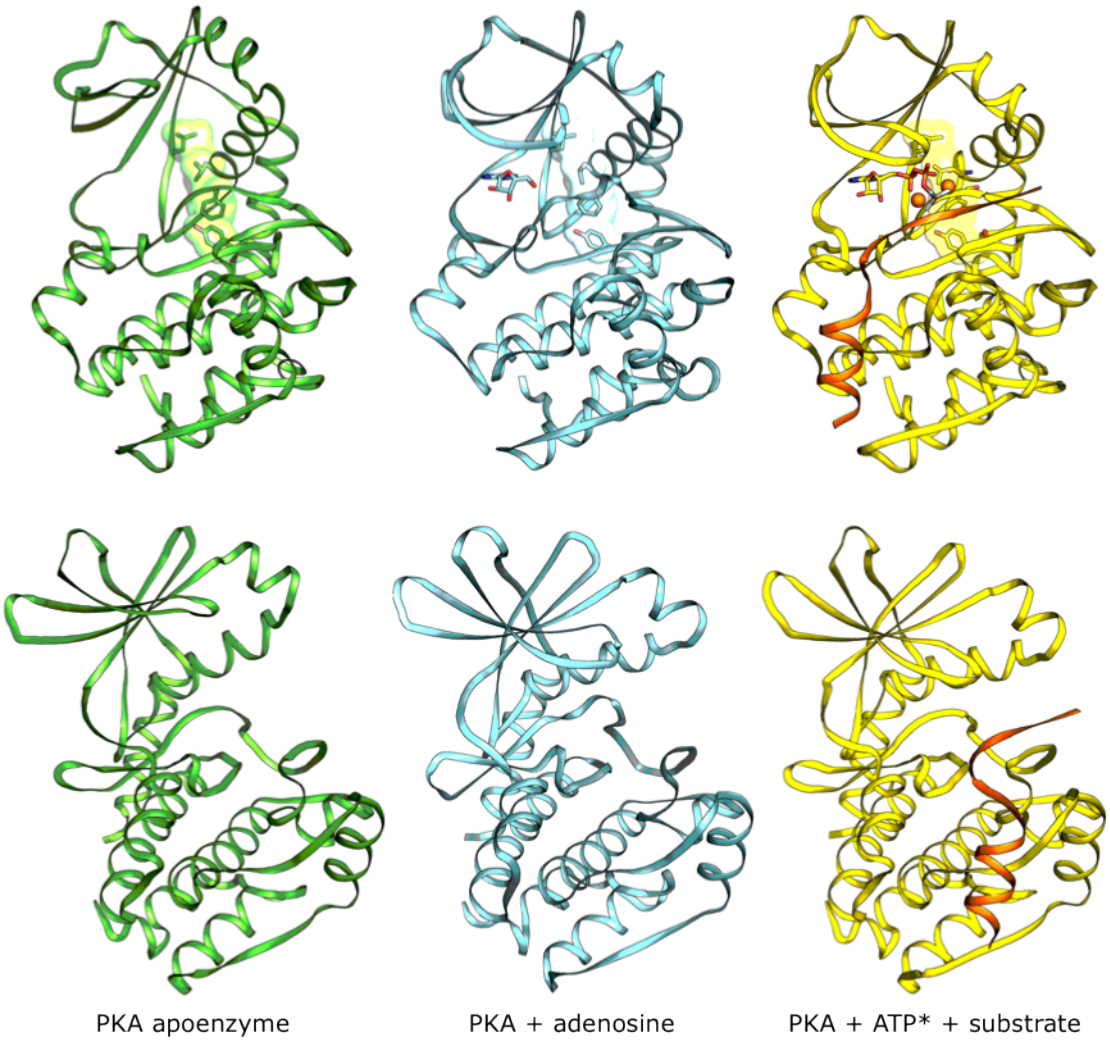


Figure 1.2 – Opening and closing of cAMP-dependent protein kinase during the catalytic cycle. Left to right: frontal (top row) and lateral (bottom row) view of the apoenzyme (green, PDB 1j3h) (Akamine et al., 2003) resembling the inactive

open conformation; the binary adenosine complex (light blue, PDB 1bkx) (Narayana et al., 1997); and the ternary complex of enzyme (yellow), aluminium fluoride-ADP and substrate peptide (orange) (PDB 1l3r) (Madhusudan et al., 2002), resembling the active, closed enzyme (all murine proteins). All systems are in the DFG-in conformation. As the enzyme traverses the catalytic cycle, a closing motion can be perceived, comprising a concerted downward movement of the glycin-rich loop (residues 50 to 55) and positioning of the αC-helix that contributes to the catalytic residues. Proteins and peptides are shown in cartoon representation with smooth loop regions, ligands as sticks (middle and right panel in top row), magnesium ions as orange spheres (not to scale). Side chains of residues comprising the regulatory spine (top to bottom, Leu^{106}, Leu^{95}, Phe^{185}, Tyr^{164}) are shown in stick representation with semi-transparent surface (top row only). Structures were superimposed minimising the C^{α}-RMSD of the C-lobe (residues 120 to 300). Pictures were generated using PyMOL (DeL).

Different (in)activation mechanisms exist in different kinases, but active conformations have common prerequisites, i.e. the accessibility of the substrate cleft, arrangement and positioning of amino acid residues participating in catalysis, and formation of the nucleotide binding pocket. Historically, the conformational state of the conserved DFG motif that is situated at the beginning of the activation segment has been used as a criterion to decide whether a kinase was observed in an active or inactive conformation, the 'DFG-in' conformation being the active and 'DFG-out' conformation the inactive conformation. This is evident as in the DFG-out conformation, the nucleotide binding pocket is indeed blocked, thereby inactivating the enzyme. However, not all kinases are able to adopt DFG-out conformations[15] (Dar et al., 2008), and DFG-in conformations have also been observed for *inactive* kinases.

Therefore, while the DFG-in conformation is not alone *sufficient* to determine the activation state of the kinase, it is *necessary*, as the DFG-out state guarantees an inactive kinase conformation. A more pragmatic conformational description of active and inactive conformational states is given by the model of the regulatory spine (Kornev et al., 2006). This spine consists of four non-consecutive hydrophobic residues which in the active conformation of the kinase are arranged in such a manner that they position different functional residues distributed throughout the primary sequence of the enzyme for catalysis. When the enzyme is activated, the regulatory spine assembles, and stays assembled throughout the entire catalytic

[15] importantly, PKA has never been observed in a DFG-out state.

cycle (Kornev et al., 2006). The assembly of the regulatory spine can therefore be regarded the major signature of active kinases. There are different mechanisms for breaking the regulatory spine, the most common one being the DFG-flip, the conformational transition from DFG-in to DFG-out (see above). While DFG-out conformations are necessarily inactive (Hubbard et al., 1994; Yang et al., 2002; Levinson et al., 2008; Shan et al., 2009), each DFG-in conformation needs closer examination of the intactness of the rest of the regulatory spine to judge its activation state (Taylor et al., 2004; Taylor and Kornev, 2011). In some kinases such as ABL (Shan et al., 2009) or Src (Xu et al., 1999), the outward movement of the αC-helix is sufficient to break the regulatory spine, induced by allosteric regulation through the SH2 and SH3 domains of the enzyme (Huse and Kuriyan, 2002; Panjarian et al., 2013). In this inactive 'αC-out' state, the DFG motif can still be found in an inward conformation.

Spine formation and opening and closing of the enzyme during the catalytic cycle will now be illustrated using PKA as an example. The regulatory spine comprises four residues (Leu^{95}, Leu^{106}, Tyr^{164}, Phe^{185}) each of which is part of different protein secondary structural elements and distinct functional protein segments (Supplementary Figure S1). When the spine is assembled, the four residues adopt a linear arrangement, with Leu^{106} anchored in the β-sheet core of the N-lobe, Leu^{95} fixing the αC-helix (amino acid residues 84 to 96) in proximity to the catalytic residues, Phe^{185} in an inward position,[16] thereby positioning Asp^{184} for catalysis, and Tyr^{164} adjacent to the catalytic loop (residues 166 to 171) spatially arranging further catalytic residues (Supplementary Figure S1). Once the spine is assembled, the enzyme is in its active conformation and remains so until it is specifically inactivated.

The catalytic cycle itself consists of an opening and closing of the enzyme (Figure 1.2). The catalysis then comprises three steps: ligand binding (ATP and substrate), chemical step (phosphate transfer), and product release. The chemical step is fast, while product release is slow (Adams, 2001).[17] Given the high ATP

[16] DFG-in.

[17] rates for the phosphoryl transfer $> 500\ s^{-1}$ and the rate-limiting release of the ADP product $20\ s^{-1}$ have been determined.

concentration under physiological conditions it can be assumed that ATP binds before the substrate (Zhou and Adams, 1997).

In the absence of substrate and the nucleotide, the apoenzyme adopts an open conformation and is *catalytically inactive*,[18] with the small lobe displaced relative to the large lobe and the glycine-rich loop (residues 50 to 55) positioned away from the active site in order to maximise accessibility to the nucleotide binding site (Figure 1.2). When ATP binds and a binary complex is formed, the enzyme gradually closes and compacts. This closing movement is predominantly mediated by a twist of the β-sheet of the N-lobe, and downward movement of the glycine rich loop, closer to the catalytic loop but not yet anchoring it firmly (Zheng et al., 1993a; Karlsson et al., 1993; Akamine et al., 2003). When substrate binds, the glycine rich loop fully closes, with the backbone nitrogen of Ser^{53} partaking in nucleotide coordination. This engaging of the top of the loop to the γ-phosphate of ATP is the critical step for catalysis, and the conserved salt bridge between Lys^{72} and Glu^{91} is formed (Supplementary Figure S1). The distance between the N-terminus of αC-helix and the activation loop can be regarded as a gauge that defines the open and closed state of the activated enzyme (Taylor et al., 2004; Taylor and Kornev, 2011). In the closed conformation, nucleotide and substrate hydroxyl are in close proximity, and all catalytic residues are arranged in such a way as to facilitate the phosphate transfer (Huse and Kuriyan, 2002). After phosphate transfer, substrate and adenosine diphosphate (ADP) product are released, and the enzyme opens up again, thereby providing better accessibility of the nucleotide pocket and a larger surface area for substrate binding.

The open (apo) enzyme and binary complexes were found to be quite dynamic and the ternary complex more rigid (Johnson et al., 2001) with indications of conformational selection being present (Masterson et al., 2010; Jarymowycz and Stone, 2006), i.e. open and closed enzyme conformations coexisting in solution, and their equilibrium being influenced by the presence of ligand and substrate. The flexibility of the enzyme is mainly due to changes in its N-lobe as overall, the large lobe is quite stable, as can be seen from crystallographic B-factors and investigations by NMR spectroscopy (Masterson et al., 2008, 2010). The flexibility of kinases is likely to be of consequence to investigations of the dynamics of protein-

[18] although it is in its *activated* state.

substrate interactions, especially by MD simulations and NMR spectroscopy (see Chapter 2). Representative structures for different states of the catalytic cycle have been captured by crystal structures (Figure 1.2), and NMR spectroscopy and molecular mechanics simulation approaches (Hyeon et al., 2009) are beginning to probe the solution dynamics and ligand-induced conformational transitions of the enzyme.

1.2.3 Kinase inhibitors

Because of their implication in cell proliferation and -survival, kinases are the primary target for therapeutic intervention in cancer chemotherapy. Kinase deregulation and autonomous activation is a central aspect of many cancer phenotypes (Blume-Jensen and Hunter, 2001). The ligand-dependent activation of the kinase is bypassed by kinase overexpression (Wang et al., 1996), gain of function mutations activating the kinase by stabilising its active conformation (Weiner et al., 1989), or autocrine stimulation (Sierra et al., 2010). However, cancer cells often display a characteristic known as *oncogene addiction*, rendering them overly susceptible to therapeutic interference with a single target they are highly reliant on, compared to cells in non-pathogenic states (Weinstein and Joe, 2008). A high number of such targets are kinases, providing a clear rationale for molecular targeted cancer chemotherapy by kinase inhibitors.

Kinase inhibitors face the challenging task of achieving specificity in spite of the highly conserved kinase fold. To date, most kinase inhibitors are direct ATP competitors, a molecule which, together with its precursors, is arguably the most frequent metabolite within a cell (Edwards and Palsson, 2000; Forster et al., 2003; Borodina et al., 2005),[19] and the endogenous ligand or cofactor of many protein families. In recent years, there has been an increase in FDA approved kinase inhibitors, with a higher number of kinase inhibitors approved after 2010 than before, as of 2013 (see Supplementary Table S1 in the Appendix). Although to date all of them are ATP competitive, more recently, allosteric and covalent ki-

[19] indeed, it can be estimated that on average about 1.7×10^7 ATP molecules are consumed per second and cell in the human body to maintain the basal metabolic rate (assuming a daily energy requirement of 2,000 kcal and using an energy density of 8 kcal mol^{-1} for ATP, and 10^{14} as the number of cells in the human body).

nase inhibitors begin to emerge (Melnikova and Golden, 2004; Hantschel et al., 2012), some of which have entered clinical development (Zhang et al., 2009; Dar and Shokat, 2011). There are two types of ATP-competitive kinase inhibitors: type I inhibitors target the active kinase conformation, while type II inhibitors target the inactive kinase conformation (Zhang et al., 2009) which in many relevant cases is the DFG-out conformation (see above).[20] Imatinib / Gleevec has been the first type II inhibitor to be described, targeting the cABL kinase implied in myelogenous leukemia (Dar et al., 2008). Importantly, vemurafenib, a type I inhibitor (Supplementary Table S1 in the Appendix), was the first ever inhibitor to be granted approval by the FDA for any target that has been developed by *fragment based drug design*, a drug design paradigm discussed in greater detail earlier in this introductory chapter (Section 1.1).

For functional reasons, in activated kinases the residues surrounding the active centre assume a more conserved conformation than in inactive kinases where a wide range of additional conformational states is observed, as different kinases are inactivated in different ways (Taylor and Kornev, 2011). Furthermore, only the closed state of the activated kinase is able to phosphorylate a target, making stabilising the inactive kinase conformation, or the open conformation of the active kinase by rational inhibitor design (Liu and Gray, 2006) a promising strategy.

Most inhibitors aim to mimick the nucleotide by targeting the hinge region of the kinase that links N- and C-lobes of the enzyme (Akritopoulou-Zanze and Hajduk, 2009). The active centre of a kinase can be divided in different subpockets, one of which overlaps with the adenine region of the endogenous nucleotide ligand, plus additional hydrophobic subpockets (Zhang et al., 2009) which are less conserved and can therefore be exploited for the design of specific inhibitors. In the DFG-out conformation of the enzyme, an additional selectivity pocket in the catalytic cleft immediately adjacent to the region occupied by the nucleotide becomes available. Therefore, in general, type II inhibitors are considered more

[20] remarkably, some type I inhibitors were found to bind the DFG-out conformation of a kinase (see Sunitinib, Bosutinib and Axitinib in Supplementary Table S1 in the Appendix), and for Lapatinib, based on its selectivity for the active versus the inactive form of human ABL1 kinase, not clear cut inhibitor type could be assigned (Zuccotto et al., 2010; Davis et al., 2011), although the molecule possesses a type II inhibitor like pharmacophore (Liu and Gray, 2006; Zhang et al., 2009).

selective than type I inhibitors. However, the inhibitor type does not necessarily dictate selectivity, as can be seen from comparing the selectivity profiles of lapatinib (highly selective DFG-in binder) and sorafenib (less selective, type II) in an extensive panel screen of kinase inhibitors (Davis et al., 2011). The DFG-out conformation however requires a significant conformational change of the enzyme which has been observed in a limited number of kinases so far (Dar et al., 2008), so there is still a considerable interest in developing type I inhibitors that achieve selectivity e.g. by targeting the variable gate-keeper residue.[21]

1.3 Heat shock protein of 90 kDa

Heat shock protein of 90 kDa (Hsp90)[22] is a molecular chaperone, assisting folding and assembly of many regulatory and signalling proteins involved in cell growth and survival (Picard, 2012; Whitesell and Lin, 2012).[23] Among them are several oncogenic protein kinases such as ErbB2 and Cdk4 (Pearl, 2005), and other proteins involved in hallmark processes of cancer (Hanahan and Weinberg, 2011): induction of angiogenesis, tumor metastasis and anti-apoptosis (Neckers and Workman, 2012; Hong et al., 2013). To exert its function, Hsp90 needs to bind and hydrolise ATP (Panaretou et al., 1998; Obermann et al., 1998). Consequently, ATP competitive Hsp90 inhibition offers a pharmacological mode of indirect interference with downstream client proteins (Mimnaugh et al., 1996; Schneider et al., 1996), promoting their degradation, or aggregation of non-functional proteins that can lead to apoptosis (Taipale et al., 2010). This, additional or alternative to targeting protein kinases directly, renders Hsp90 an attractive target for cancer chemotherapy (Workman, 2004).

Hsp90 is highly abundant in eukaryotic cells, accounting for 1 to 2 % of the total cellular protein amount, a value that can increase to 4 to 6 % from baseline levels under stress conditions (Borkovich et al., 1989; Taipale et al., 2010). This corresponds to the excess of 10^7 individual Hsp90 protein copies under equilibrium

[21] Met^{120} in PKA.

[22] Pfam identifier PF00183, Uniprot P07900, CHEMBL3880, the latter two identifiers referring to the human protein.

[23] both articles are part of a special issue solely dedicated to Hsp90.

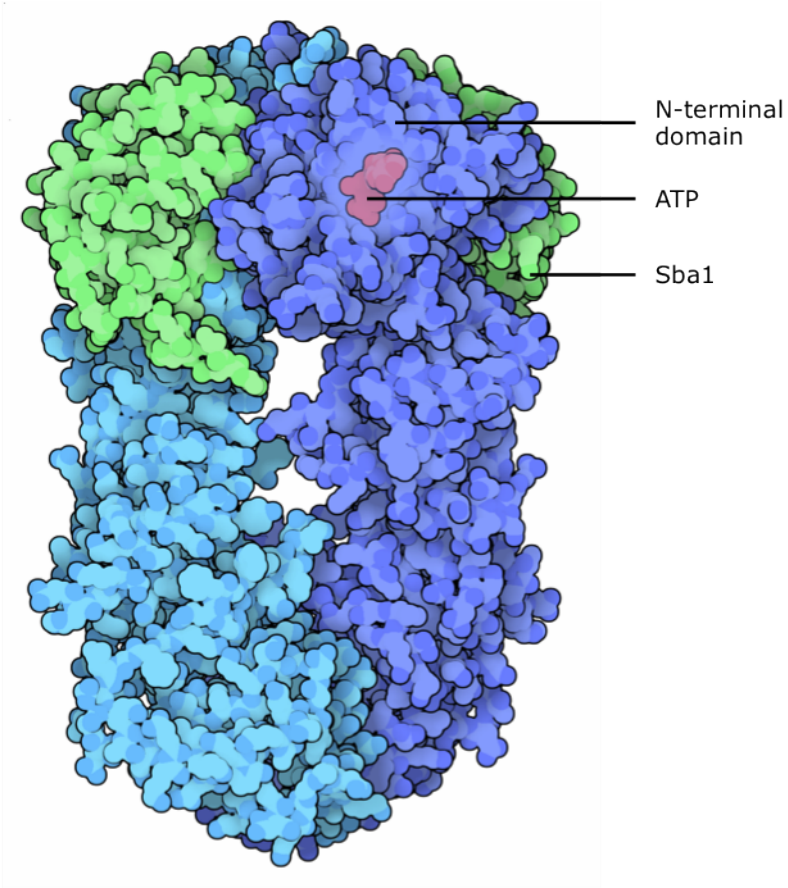


Figure 1.3 – Structure of the Hsp90 holo enzyme. Hsp90 (blue) from *Saccharomyces cerevisiae* (UniProt P02829) in complex with ATP (red) bound to its N-terminal ATPase domain, and the co-chaperone Sba1 (green). It assumes a dimeric clamp-like quarternary structure that opens and closes during its catalytic cycle while consuming ATP, interacting with substrate proteins, amongst which are many oncogenic kinases. The closed conformation of the complex is depicted. It should be noted that the ATP binding site is not accessible in the complex conformation shown here, and that the position of the nucleotide is only visible due to the semi-transparent graphical representation of the protein.

conditions, or a protein concentration of about 10 μM.[24] Chaperone proteins such as Hsp90 promote protein folding in a crowded celluar environment (Ellis, 2007) where the exposed hydrophobic amino acid residues of nascent polypeptide chains face the problem of encountering many different possible non-native intermolecular interactions, diverting them from productive folding and possibly leading to disfunctional aggregation of immature proteins (Zou et al., 2008). Chaperones help to prevent these off-pathway reactions (Taipale et al., 2010) by transiently binding to vulnerable folding intermediates (Hartl and Hayer-Hartl, 2009). Furthermore, chaperones are known to target corrupt, terminally misfolded proteins for degradation by the ubiquitin-proteasome system (McClellan et al., 2005).

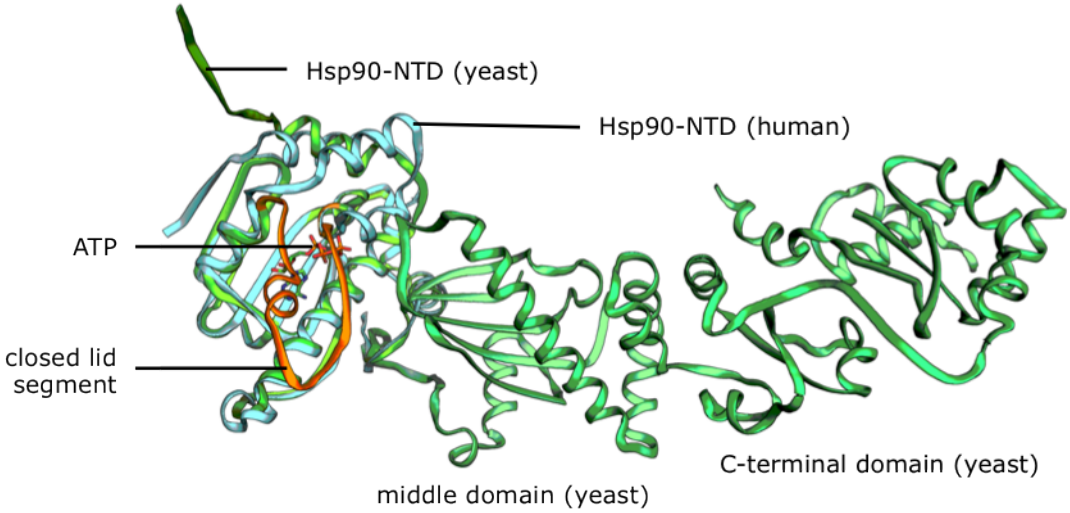


Figure 1.4 – Ligand-induced closing of a lid segment (orange) in the N-terminal ATPase domain (green) of full-length Hsp90 from *S. cerevisiae* (PDB 2cg9) (Ali et al., 2006). One monomer of the protein is shown in cartoon representation (compare to the dimeric structure depicted in Figure 1.3). For comparison, the isolated N-terminal domain of human Hsp90 (blue) is shown, assuming an open conformation (PDB 1yet) (Stebbins et al., 1997). The nucleotide ligand of the yeast protein is shown in stick representation with blue nitrogen, red oxygen, and orange phosphorus atoms. The picture was generated using PyMOL (DeL).

[24] protein count estimations taken from quantitative mass spectrometry studies in human bone osteosarcoma cells where about 1.05×10^7 instances of Hsp90 were observed ($> 7.3 \times 10^8$ copies of more than 7,300 unique proteins overall) (Beck et al., 2011), and murine embryonic fibroblasts where about 1.6×10^7 instances of Hsp90 were observed ($> 3 \times 10^9$ copies of more than 5,000 unique proteins overall) (Schwanhausser et al., 2011). Concentration estimates depending on the numbers of protein molecules found in a single cell were derived from (Moran et al., 2010).

Cancer cells are faced with an exceptional amount of environmental stress such as malnutrition and reduced oxygen levels (Neckers and Workman, 2012), and their high growth rate requires efficient and accurate protein folding. Under these conditions, Hsp90 acts as an important molecular stress buffer. Furthermore, cancer chemotherapy itself applies stress to tumor cells which is potentially alleviated by chaperones such as Hsp90 (Nahleh et al., 2012). Therefore, targeting Hsp90 facilitates a 'two-pronged' interference mechanism with cancer (Travers et al., 2012).

The Hsp90 protein assumes a homodimeric, clamp-like quarternay structure, interacting with different co-chaperones (Prodromou et al., 1999; Panaretou et al., 2002; Siligardi et al., 2002) (Figure 1.3).[25] A monomer of Hsp90 consists of three domains (Pearl and Prodromou, 2000, 2001; Prodromou and Pearl, 2003): a highly conserved N-terminal ATPase domain (NTD) of about 25 kDa, a charged linker region connecting the NTD with the middle domain of about 40 kDa involved in client protein binding, and a C-terminal dimerisation domain of about 12 kDa. During its ATP dependent catalytic cycle, the enzyme is believed to open and close (Chadli et al., 2000; Pearl and Prodromou, 2006), interacting with its protein substrates (Pearl et al., 2008). The ATP binding cleft is unusually shaped and is characterised by a left handed $\beta - \alpha - \beta$ (Bergerat) fold (Bergerat et al., 1997; Prodromou et al., 1997; Roughley et al., 2012).In the ATP bound state, the enzyme undergoes both local and global conformational changes. A lid segment within the NTD closes over the substrate site (Figure 1.4), enhancing its ATPase activity (Ali et al., 2006). Additionally, the NTD transiently partakes in dimerisation (Chadli et al., 2000; Prodromou et al., 2000), altering the quarternary structure of the holoenzyme. It thereby contributes to its opening and closing motion, and ultimately, release of the matured client protein.

Hsp90-NTD is the primary target for therapeutic intervention (Pearl et al., 2008; Solit and Chiosis, 2008) and focus of structure based drug design approaches in recent years. Similar to kinases where the effect of oncogene addiction (Weinstein and Joe, 2008) is observed (see above), cancer cells appear to be hypersensitive to Hsp90 inhibitors (Kamal et al., 2003; Barrott and Haystead, 2013). Among

[25] figure adapted from `http://dx.doi.org/10.2210/rcsb_pdb/mom_2008_12`, PDB Molecule of the Month, by David Goodsell, based on PDB structure 2cg9 (Ali et al., 2006) (*Saccharomyces cerevisiae*).

the first experimentally determined structures of Hsp90-NTD was the complex with the antibiotic geldanamycin[26] (PDB identifier 1yet) (Stebbins et al., 1997), a macrocyclic polyketide from *Streptomyces hygroscopicus* structurally unrelated to ATP, that was already known to target Hsp90 (Whitesell et al., 1994). Since then, structures have been obtained routinely in complex with many different ligands and lead compounds.[27]

Hsp90 inhibitors intended for therapeutic use can be sub-classified into first and second generation inhibitors, with the former derived from geldanamycin, and the latter synthetic small molecule-like compounds many of which were found by *fragment based drug discovery* endeavours (Roughley and Hubbard, 2011) (see section 1.1). Clinical data for several compounds under active development has been reviewed recently (Neckers and Workman, 2012; Hong et al., 2013). While no clinical candidate has been approved by the FDA yet, or successfully progressed past clinical phase 3, currently, there are several clinical trials under way for Hsp90 inhibitors.[28] Despite unfavourable phase 3 clinical trials with first generation compounds, Hsp90 inhibitor design is still an active area of pharmaceutical research. While the development of first generation inhibitors can be considered as suspended, with the possible exception of retaspimycin[29] (IPI-504, Infinity Pharma) where a phase 1 clinical trial is running for a combination therapy, there are three second generation compounds currently being investigated. AT13387 (Astex) is in phase 1 trials, while AUY922[30] (Vernalis and Novartis) and ganetespib (STA9090, Synta Pharma) (Choi and Lee, 2012) have progressed to phase 2. In general, second generation compounds seem more tolerable than first generation compounds which have shown significant liver toxicity. Adverse effects of second generation compounds include reversible visual impairment and possible cardiac effects (Hong et al., 2013). HSP90 has been used extensively as a model system for FBDD (see Introduction section 1.1) by Vernalis and other companies (for a review, see (Roughley and Hubbard, 2011; Roughley et al., 2012)).

[26] CHEMBL 278315.

[27] as of May 2013, there are 242 X-ray structures of eukaryotic Hsp90 in the PDB, 200 of which from human or yeast.

[28] data received from `http://www.clinicaltrials.gov` (May 2013), and Dr. Bissan al-Lazikani (ICR London), personal communication.

[29] CHEMBL1184904.

[30] CHEMBL252164.

Chapter 2

Protein Internal Motions Influence Observables in a Ligand-Detected NMR Experiment

The work presented in this chapter has been conducted at EMBL Heidelberg (Germany) under the supervision of Dr. Teresa Carlomagno, and a part of it has been published previously (Stauch et al., 2012).

2.1 Introduction

Protein surfaces are not static but plastic boundaries, interacting with and adapting to ligands. Besides steric and electrostatic interactions, dynamic features of proteins and protein-ligand interactions have been shown to be functionally relevant (Karplus and Kuriyan, 2005; Kay et al., 1998). Protein dynamics can be probed both experimentally and computationally (Hub and de Groot, 2009; Kohn et al., 2010; van Westen et al., 2010), with nuclear magnetic resonance (NMR) spectroscopy standing out as an especially well-suited experimental tool to study the dynamics of complexes in a close-to-native liquid environment (Mittermaier and Kay, 2006, 2009). NMR spectroscopy is particularly powerful in investigat-

ing transient complexes formed by proteins and synthetic organic ligands (Carlomagno, 2005), or natural products (Carlomagno, 2012). Ligand binding epitopes can be mapped using STD (saturation transfer difference) techniques (Jayalakshmi and Krishna, 2002; Mayer and Meyer, 2001), protein residues in contact with the ligand can be identified using chemical shift perturbation experiments (McCoy and Wyss, 2002), and transferred-NOEs or transferred-CCR (cross-correlated relaxation) rates allow for the determination of the bioactive conformation of the ligand (Blommers et al., 1999; Ni, 1994; Carlomagno et al., 1999).

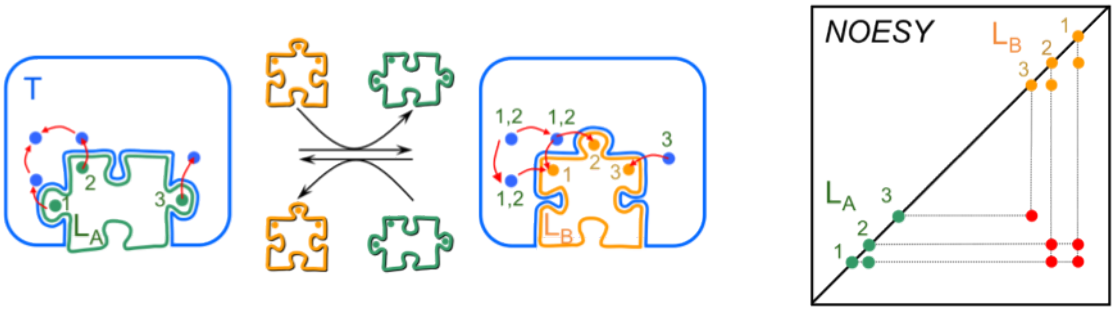


Figure 2.1 – Origin of inter-ligand signals in an INPHARMA experiment. During the mixing time of a NOESY experiment, L_A binds to the receptor T (blue) and the magnetisation is transferred from the ligand protons 1, 2 and 3 to protons of the binding pocket (red arrows). Subsequently, L_A (green) dissociates and L_B (orange) binds. Magnetisation originating from L_A and stored at the protein binding site is transferred to protons of L_B 1, 2 and 3 as this magnetisation originates from L_A, this transfer gives rise to cross-signals between the resonances 1, 2 and 3 of L_A and 1, 2 and 3 of L_B (red circles in NOESY schematic on the right). INPHARMA NOEs thus stem from an indirect, spin diffusion mediated transfer, relayed by receptor protons (Orts et al., 2008b), that can be detected in a two dimensional NOESY experiment.

Recently, the INPHARMA (Interligand Noes for PHARmacophore MApping) method has been described, which allows the determination of the relative binding mode[1] of two competitive, transiently bound ligands (Orts et al., 2008b; Sanchez-Pedregal et al., 2005). INPHARMA relies on *indirect*, inter-ligand, protein mediated, transferred-NOE signals in a NOESY-like experiment between two ligands, L_A and L_B, binding competitively and weakly to a common receptor T (Figure 2.1). The efficiency of the INPHARMA transfer at each ligand site depends on

[1] that is, the superimposition of the ligands.

the relative binding mode of the ligands to the protein. In general, the magnetisation transfer is more efficient between protons of the two ligands that are close to the same *protein* protons in the binding pocket. Because of this dependency, a quantitative analysis of the INPHARMA NOEs allows for the determination of the relative binding mode of L_A and L_B to the protein target. In favourable cases, the *absolute* orientation of the ligands within the protein binding pocket can be determined from INPHARMA data as well (Orts et al., 2008b). Both relative and absolute binding mode allow for establishing *pharmacophores* of the ligands, as chemical groups of equivalent function and interaction are likely to spatially overlap between different ligands. In agreement with standard structure based drug design work flows, the INPHARMA data are used to select the correct binding modes from a pool of pairs of complex structures generated, for example, by molecular docking. The conformation of the bound ligand is easily obtainable from transferred NOE data (Clore and Gronenborn, 1983; Ni and Scheraga, 1994) for ligands that bind their macromolecular target with low affinity ($K_d > 10^{-6}$ M) (Carlomagno, 2012). The agreement between experimental and back-calculated INPHARMA data for each pair (i, j) of complexes $(T \cdot L_A)_i$ and $(T \cdot L_B)_j$ is then used as a selection criterion (Reese et al., 2007) in order to choose the pair of true ligand orientations. This key concept of the INPHARMA methodology is of central importance for the understanding of the remainder of this chapter.

The INPHARMA approach is particularly powerful in a scenario where high-resolution structures need to be obtained from series of low-affinity compounds interacting with a common target, such as fragment-based lead discovery in pharmaceutical industry (Rees et al., 2004; Carr et al., 2005) (see also Introduction section 1.1). Specifically, in cases where the absolute binding mode of one compound to the target of interest is already known, the determination of the orientation of each additional compound becomes very easy compared to the simultaneous *de novo* determination of two ligand orientations, as the pool of possible solutions is not N^2, but only N due to 'fixing' the known orientation for the first ligand. INPHARMA therefore closes a gap in structure-based drug discovery, where the frequent task of binding-mode determination of low-affinity ligands is especially tedious. The INPHARMA methodology relies on the exchange rate of

both ligands to be on the same time-scale as the corresponding NMR experiments, namely the mixing time of a NOESY experiment, requiring $k_{off} > 100$ to 1,000 Hz, and $K_d > 1$ μM, assuming a diffusion limited k_{on}, so that L_A can dissociate from the receptor during the mixing time if the NOESY experiment and leave space for L_B to bind (Carlomagno, 2012). The rest of this chapter will discuss current limitations of the back-calculation procedure and devise means to improve the underlying physical model assumptions, thereby making the model more realistic, and ultimately more useful.

In previous work, Protein Kinase A (PKA) has been used as a model system to test the ability of INPHARMA to determine binding poses. A relatively thorough review of PKA structure and function, and its relevance as a prototype kinase, can be found in the general Introduction (section 1.2). It has been found that for bovine PKA and two competitive ligands L_A and L_B (Figure 2.2), the INPHARMA NOEs allowed for selection of the correct ligand binding poses from a pool of pairs of structures of PKA/L_A and PKA/L_B, representing combinations of very different orientations of the ligands (Orts et al., 2008b). A high correlation coefficient was found between experimental and back-calculated INPHARMA NOEs for the complex pairs representing the correct, reference ligand binding poses as determined by X-ray crystallography. In this favourable case, the INPHARMA data allowed for clear selection not only of the relative, but also of the *absolute* binding mode of both L_A and L_B. Despite obtaining a high Pearson correlation coefficient (R) of 0.82 between the experimental INPHARMA NOEs and the data back-calculated from the two crystal structures, the magnitude of the magnetisation transfer of the experimental data were consistently lower than the theoretical ones (slope a of a linear fit line through the origin, $y = ax$, of 0.33), indicating over-estimation of the magnetisation transfer efficiency (Orts et al., 2008b, 2009). To explain this effect, the influence of protein internal motion on the INPHARMA NOEs was suggested, as order parameters $S^2 < 1$ would reduce the efficiency of the magnetisation transfer (see below). Since the inter-ligand NOEs observed in an INPHARMA experiment are mediated by protein protons through spin diffusion, their value depends on protein internal motion. This is in contrast to intra-ligand transferred-NOEs, which are dominated by direct dipolar-dipolar interaction be-

tween protons of the ligand and are only mildly affected by the protein protons.

In this chapter, NMR order parameters S^2 are included into the INPHARMA calculations. In order to demonstrate the usefulness of order parameters in improving the quality of the fit between experimental and theoretical data and in increasing the discrimination power of the approach, firstly, a set of order parameters S^2 for the PKA/L_A and PKA/L_B complexes is extracted from Molecular Dynamics (MD) simulations. Secondly, the implementation of the *full relaxation matrix approach* (Nilges et al., 1991), used to back-calculate theoretical INPHARMA data from given complex structures, is modified to allow for incorporation of order parameters $S^2 < 1$ in the spectral density function (see below). Thirdly, a set of *generic order parameters* is introduced that can be used in INPHARMA calculations of arbitrary protein-ligand systems, thus by-passing the need for performing computationally expensive MD simulations of each protein-ligand complex of interest. Finally, it is demonstrated that even the use of generic, non-model-specific order parameters can increase the discrimination power of the INPHARMA method.

2.2 Theory

In this section, the theory underlying the influence of protein internal motion on the NOE dependent magnetisation transfer will be outlined concisely. Dipolar cross relaxation rates $\sigma_{kl}^{\mathrm{NOE}}$ determine magnetisation transfer between spins k and l through space and depend on the spectral density functions (Ernst et al., 1987):

$$J_{kl}(\omega) = \int_{-\infty}^{\infty} C_{kl}(t)\exp(-i\omega t)dt \tag{2.1}$$

which are the Fourier transforms of the dipolar correlation functions,

$$C_{kl}(t) = 4\pi \left\langle \frac{Y_{20}(\theta_{kl}^{\mathrm{lab}}(t_0+t))Y_{20}^{*}(\theta_{kl}^{\mathrm{lab}}(t_0))}{r_{kl}^3(t_0+t)r_{kl}^3(t_0)} \right\rangle \tag{2.2}$$

with $\theta_{kl}^{\mathrm{lab}}$ the angle between the inter-nuclear vector r_{kl} and the external magnetic field, Y_{2m} the rank 2 spherical harmonics of order m and the angled brackets denoting a Boltzmann ensemble average. Assuming that the overall tumbling

motion of the molecule is much slower than the fast internal motion, the two kinds of motion are separable and can be treated independently of each other (Wallach, 1967). For isotropic diffusional tumbling, the correlation function of the overall motion is an exponential $C^{\text{tumbling}}(t) = e^{-|t|/\tau_c}$ with τ_c the correlation time of the molecule. The contribution of the internal motions to the dipolar correlation function has the form

$$C_{kl}^{\text{internal}}(t) = \frac{4\pi}{5} \sum_{m=-2}^{2} \left\langle \frac{Y_{2m}(\theta_{kl}^{\text{mol}}(t_0+t), \phi_{kl}^{\text{mol}}(t_0+t)) Y_{2m}^{*}(\theta_{kl}^{\text{mol}}(t_0), \phi_{kl}^{\text{mol}}(t_0))}{r_{kl}^{3}(t_0+t) r_{kl}^{3}(t_0)} \right\rangle \tag{2.3}$$

with r_{kl}^3, θ_{kl}^{mol}, and ϕ_{kl}^{mol} the spherical coordinates in a molecular fixed frame. For $t \to \infty$ the internal correlation function $C^{\text{internal}}(t)$ assumes a plateau value S^2, which is called the NMR order parameter (Lipari and Szabo, 1982a,b).

The dipolar spectral density function can then be rewritten (Brüschweiler et al., 1992) as

$$J_{kl}(\omega) = \left\langle r_{kl}^{-6} \right\rangle S_{kl}^2 \frac{2\tau_c}{1+\omega^2\tau_c^2} + \left\langle r_{kl}^{-6} \right\rangle (1 - S_{kl}^2) \frac{2\tau_{\text{tot}}}{1+\omega^2\tau_{\text{tot}}^2} \tag{2.4}$$

where

$$S_{kl}^2 = \frac{4\pi}{5} \left\langle r_{kl}^{-6} \right\rangle^{-1} \sum_{m=-2}^{2} \left| \left\langle \frac{Y_{2m}(\theta_{kl}^{\text{mol}}, \phi_{kl}^{\text{mol}})}{r_{kl}^3} \right\rangle \right|^2 \tag{2.5}$$

and

$$\frac{1}{\tau_{\text{tot}}} = \frac{1}{\tau_c} + \frac{1}{\tau_{kl}} \tag{2.6}$$

with τ_{kl} the internal correlation time. The second term of Equation 2.4 can be omitted if $\tau_{kl} \ll \tau_c$. This is the case for very fast internal motion in the *extreme narrowing limit.*

Assuming that the angular and radial fluctuations are independent, the order parameter of Equation 2.5 can be factorised as

$$S_{kl}^2 \approx S_{r,kl}^2 \cdot S_{\Omega,kl}^2 \tag{2.7}$$

where

$$S_{r,kl}^2 = \left\langle r_{kl}^{-3} \right\rangle^2 / \left\langle r_{kl}^{-6} \right\rangle \tag{2.8}$$

and

$$S^2_{\Omega,kl} = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \left\langle Y_{2m}(\theta^{\mathrm{mol}}_{kl}, \phi^{\mathrm{mol}}_{kl}) \right\rangle \right| \tag{2.9}$$

are the radial and angular contributions, respectively.

2.3 Results

In previous work (Orts et al., 2008b) conducted on Protein Kinase A (PKA) in complex with two ligands L_A and L_B (Figure 2.2), it was found that a high Pearson correlation coefficient could be achieved for a linear fit between experimental INPHARMA NOEs and INPHARMA NOEs calculated from crystal structures in the complexes PKA/L_A (PDB identifier 3dne) and PKA/L_B (PDB 3dnd) (Pearson correlation coefficient R = 0.82) (Figure 2.3). In this work, internal motions of both the protein and the ligands had been neglected, i.e. $S^2_{kl} = 1$ for all pairs of protons (k, l). The two ligands represent core hinge binding fragments of two known ATP competitive kinase inhibitors with affinities of 6 and 16 μM, respectively, for the test system (Stocks et al., 2005; Orts et al., 2008b).[2] Despite the high correlation coefficient of the linear fit $y = ax$,[3] the slope of only $a = 0.33$ indicated the systematic over-estimation of the magnetisation transfer by a factor of ~3 (Orts et al., 2008b).

In this chapter, these shortcomings are addressed by incorporating a rigorous model of protein internal motion into back-calculation of INPHARMA NOEs. For reference, it should be stated that throughout this chapter, due to practical considerations concerning NMR and X-ray experiments, PKA proteins from different species have been used interchangeably. Specifically, NMR data were generated using PKA from Chinese hamster (UniProt identifier P25321) (UniProt Consortium, 2012), while for the X-ray structures used for the computational back-calculation of theoretical INPHARMA spectra, the bovine protein (UniProt P00517) was used. Accordingly, bovine protein structures have been used for MD simulations conducted in this chapter (see below) and were used in the context of NMR data. Importantly, however, compared to the human protein discussed in Introduction

[2] as determined from competition experiments against ROCK inhibitor Y-27632.
[3] with x the vector of back-calculated and y the vector of experimental data.

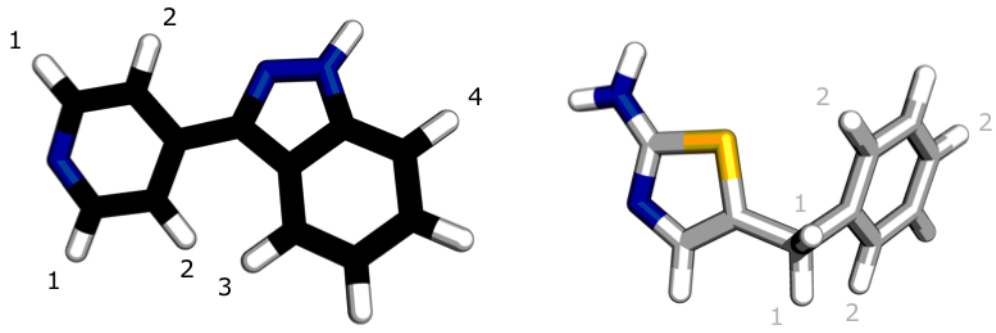


Figure 2.2 – Protein Kinase A ligands L_A (left, black, PDB identifier 3dne for the PKA/ligand complex) and L_B (right, grey, PDB 3dnd) (Orts et al., 2008b). Atoms are represented as sticks (white: hydrogen, blue: nitrogen, yellow: sulphur, black/grey: carbon). Numbers next to hydrogen atoms indicate distinct groups of overlapped chemical shifts of non-exchanging protons between which INPHARMA NOEs were observed and quantified. See Figure 2.3 for structures of the protein-ligand complexes.

section 1.2 (UniProt P17612), the amino acid sequences of the critecine[4] and bovine proteins are highly identical ($\geq$ 98 %), differing only in two positions within the kinase core domain (amino acid residues 44 to 298),[5] and the C^{α} atoms of both residues are located more than 15 Å from the closest ligand non-hydrogen atom in PDB structures 3dnd and 3dne. Therefore, it can be safely assumed that the protein characteristics relevant for this chapter do not differ significantly across species.

2.3.1 INPHARMA calculations using uniform order parameters

In order to explain the deviation of the slope of the linear fit from 1, the impact of including internal motions into the INPHARMA calculations was explored. Protein internal motions are expected to have an impact on the values of the inter-ligand INPHARMA NOEs, as these NOEs are mediated by the protons of the

[4] hamster.

[5] amino acid residue 45 (human reference sequence numbering) is glutamate in human and cow, but aspartate in hamster; residue 64 is lysine in human and hamster, but methionine in cow.

protein via spin diffusion. Unlike transferred-NOEs, which have been shown to depend mostly on direct interactions between protons of the ligand, INPHARMA NOEs strictly depend on the interaction of the ligand(s) with the protein.

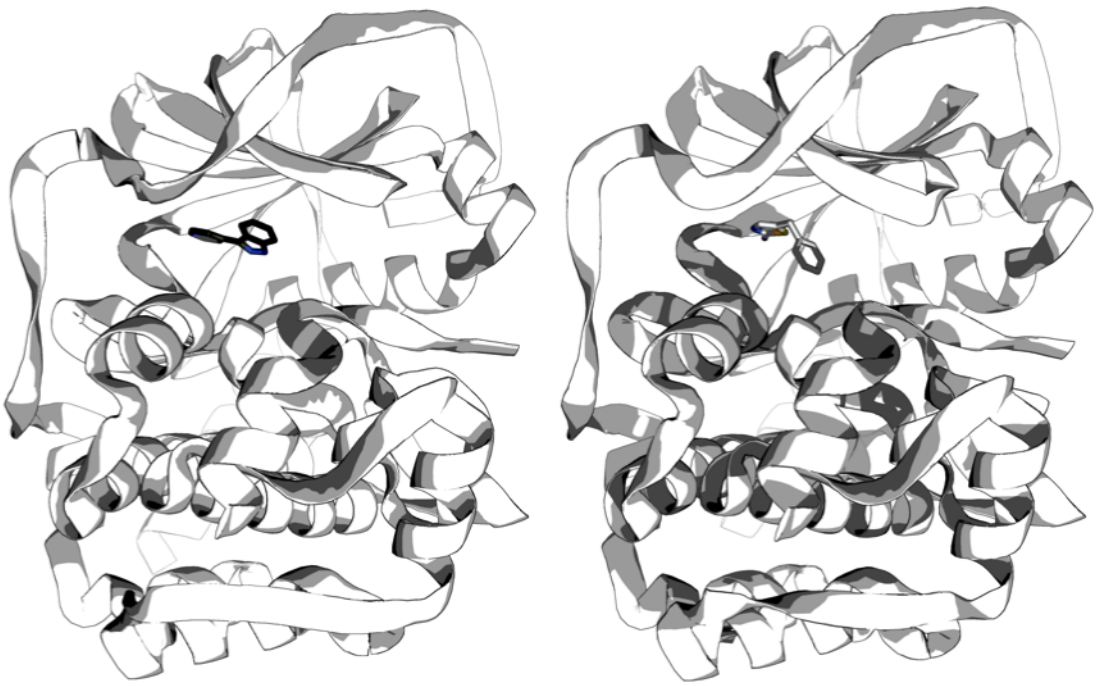


Figure 2.3 – Protein Kinase A complexes PKA/L_A (left, PDB identifier 3dne) and PKA/L_B (right, PDB 3dnd) (Orts et al., 2008b). Proteins are shown in cartoon representation (white) with flat sheets, round helices, and smooth loops for simplicity; ligands are represented as sticks.

In order to provide a general understanding of the influence of internal motions on INPHARMA NOEs, the effect of *uniform* order parameters $S^2 < 1$ of different size for both the protein and the ligands in both the free and bound states was simulated (Figure 2.4). The data were calculated for the system consisting of the PKA/L_A and PKA/L_B complexes, for which the correlation time τ_c was varied artificially between 1 and 1,000 ns to simulate the effect of receptor size. To this end, the INPHARMA method was implemented using the Python scripting language (see Methods section 2.5.4), providing an interface to an efficient routine for the computation of the matrix exponential (see Methods sections 2.5.4 and 2.5.5) using the SciPY library. The implementation devised for this chapter was found

to be more than one order of magnitude faster than a pre-existing implementation in MATLAB® which did not allow for the incorporation of NMR order parameters $S^2 \neq 1$ (Orts et al., 2008b).

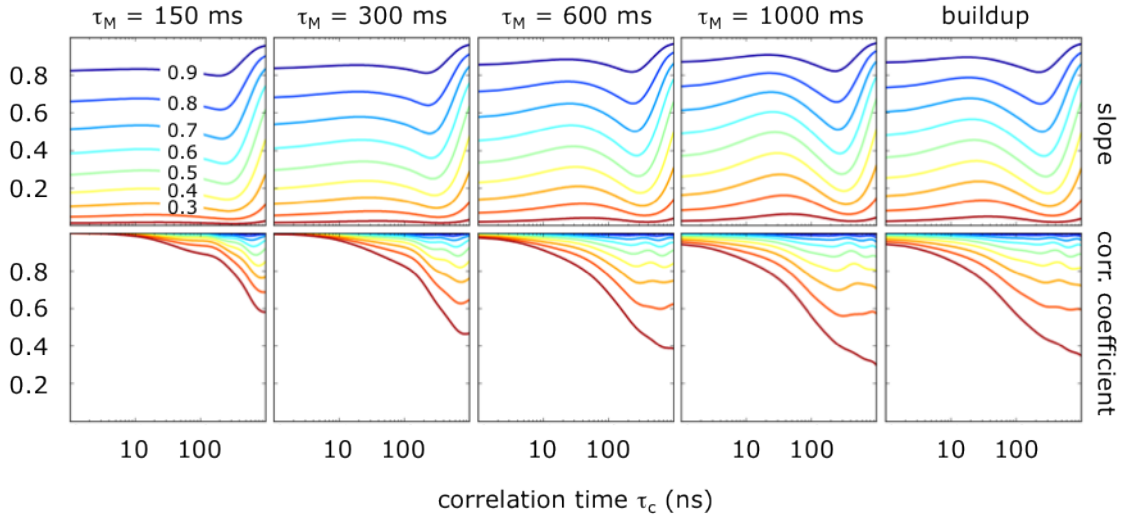


Figure 2.4 – Comparison of the effect of different values of uniform order parameters. Regression analysis was performed for PKA/L_A and PKA/L_B complexes with uniform order parameters $S^2 = c < 1$ versus synthetic reference INPHARMA NOEs calculated for a static model ($S^2 = 1$), for a constant c. Slopes a of the linear regression ($y = ax$, top panels) and Pearson correlation (corr.) coefficients (bottom) of best fit lines are shown in dependence of the complex size (x-axis, logarithmic scale from τ_c =1 to 1,000 ns), and order parameters S^2 (contour lines, colour coded on a rainbow scale from red (0.1) to green to blue (0.9), corresponding to different values c for the constant uniform order parameter, as indicated in the top left panel) for different mixing times τ_M (left to right), as well as the full NOE buildup consisting of combined data from all four mixing times. All combinations of INPHARMA NOEs between the groups of protons of Figure 2.2 were calculated, and data normalised to diagonal peak intensities in a simulated NOESY spectrum with 150 ms mixing time. See methods section 2.5.4 for technical details on the INPHARMA calculation.

For each mixing time τ_M, two parameters were varied simultaneously: τ_c as proxy for the receptor size, and a uniform order parameter $S^2_{kl} = c < 1$ with a constant c for all proton pairs (k, l). The set of INPHARMA NOEs calculated for this specific combination of τ_M, τ_c, and c was then compared to a *synthetic* reference set of INPHARMA NOEs, calculated for the same parameters τ_M and τ_c, but with a *static* model assuming no internal motion, i.e. $S^2 = 1$ for all proton

pairs. The values of the correlation coefficient and slope obtained for linear fits between the two were grouped *per uniform value of* S^2, and presented as functions of the correlation time τ_c, for different mixing times τ_M.

For a medium sized protein (τ_c = 10 to 40 ns), the intensities of the INPHARMA NOEs were found to be very sensitive to internal motion. The effect of order parameters gets more pronounced for larger receptors ($\tau_c > 100$ ns) since as spin diffusion gets more efficient, magnetisation leakage away from the binding site into the protein core increases. For very large receptors ($\tau_c \approx$ 1,000 ns) and long mixing times τ_M, a compensatory effect can be appreciated as the slope values rise again. This is most probably due to the fact that spin diffusion is now so efficient that the protons most distant from the ligand binding site begin to experience magnetisation transfer. As outlined in Methods section 2.5.4, in order to limit computational costs, the INPHARMA calculations were restricted to the 339 protons closest to the ligand binding site, including all receptor protons within 10 Å of any of the two ligands. Therefore, magnetisation leakage away from this outmost proton layer is rendered impossible by the absence of even more distant protons from the receptor model, and the graph of the slope in Figure 2.4 is not necessarily physically meaningful for very large receptors.

This result emphasises the importance of considering internal motions for small and medium sized systems. At the same time, the discrimination power of the ligand pose selection based on the INPHARMA calculations proves remarkably robust with respect to variations of the order parameters for both small and large receptors, as correlation coefficients stay high over a wide range of order parameter values.

Next, it was investigated which group of protons had the strongest effect on the magnitude of magnetisation transfer. To this end, for a fixed receptor size (τ_c = 17 ns) the order parameters S^2 for inter-molecular and intra-molecular NOEs were varied independently in a systematic fashion, and the intensity of INPHARMA NOEs was monitored (data not shown). As expected, the reduction of the order parameters for the inter-molecular NOEs[6] were found to have the largest effect, and reduced the efficiency of the protein mediated, effective magnetisation transfer between the two ligands. This reduction can be compensated for by the presence

[6]that is, the NOEs between ligand and protein.

of $S^2 < 1$ for intra-protein NOEs, which, as explained above, reduces the loss of magnetisation in the protein core. Intra-ligand order parameters $S^2 < 1$ were found to contribute least.

2.3.2 INPHARMA calculations using tailored order parameters

The calculations summarised in Figure 2.4 predict that a uniform order parameter $S^2 < 1$ of about 0.5 would be necessary to cause a three-fold reduction of the magnitude of the INPHARMA NOEs for the PKA system. In order to verify whether a more realistic representation of the protein and ligand dynamics would be able to explain the observed slope of 0.33 in the linear fit between the experimental and back-calculated INPHARMA NOEs for the PKA/L_A and PKA/L_B complexes, an estimation of order parameters for the complexes was obtained from trajectories from Molecular Dynamics (MD) simulations (Brüschweiler et al., 1992). MD simulations of both the PKA/L_A and PKA/L_B complex were performed and a set of order parameters S^2 extracted for all proton pairs within 10 Å of the ligand binding pocket that have less than 6 Å mutual inter-nuclear distance in the starting structure that was obtained from the crystal structures of the two complexes by adding protons (see Methods section 2.5.1). Due to these cutoffs, proton pairs not included in the order parameter extraction would either have a mutual distance beyond NOE detection, or be too distant from the binding pocket for spin diffusion mediated magnetisation transfer. To remove the overall tumbling motion of the molecule during the course of the simulation, each MD frame was superimposed on the crystal structure as a common reference frame. The order parameter was factorised into the radial and angular component according to Equation 2.7, which was found to be a good approximation for our system (correlation coefficient R >0.996 between S^2_{kl} and the product $S^2_{r,kl} \cdot S^2_{\Omega,kl}$). The order parameters derived from the MD simulations were found to have an average value ($\pm$ standard deviation) of 0.62 ± 0.22, in good agreement with proton-proton order parameters derived from MD simulations in another study (Schneider et al., 1999).

Spectral density functions containing the first term of Equation 2.4 were used in the full relaxation matrix to calculate the INPHARMA NOEs for the PKA/L_A

and PKA/L_B complexes with the correct ligand orientations. Disappointingly, the incorporation of internal motions of this functional form into the INPHARMA calculations considerably deteriorated the quality of the fit between experimental and back-calculated data, yielding R = 0.66 (data not shown). However, the slope increased to 0.71, suggesting that internal motions can indeed explain the over-estimation of the INPHARMA NOEs observed in the back-calculation performed for the static complexes. The decrease in the value of the theoretical INPHARMA NOEs of more than two-fold upon inclusion of internal motion[7] indicates that the internal motions primarily affecting the INPHARMA NOEs are of angular nature. A strong effect of radial fluctuations would in fact *increase* the rate of the NOE transfer as $\left\langle r_{kl}^{-6} \right\rangle \geq \left\langle r_{kl}^{-3} \right\rangle^2 \geq \langle r_{kl} \rangle^{-6}$ strictly holds true,[8] and therefore result in a decrease of the slope of the correlation between experimental and theoretical data.

[7] as documented by the *increase* in slope from 0.33 to 0.71.

[8] with equality only applying in the absence of internal motion, i.e. constant r_{kl} over the course of the simulation.

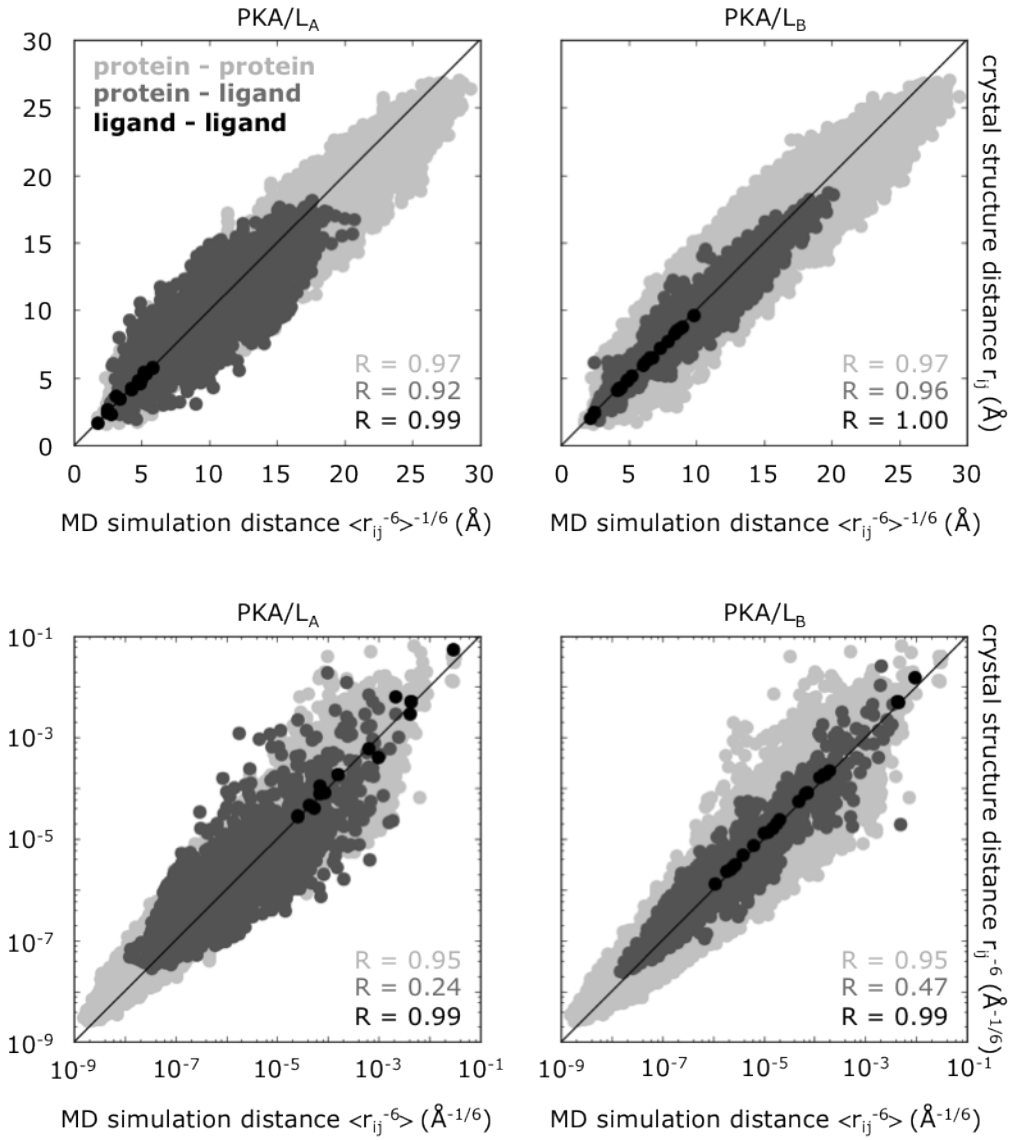


Figure 2.5 – Quality of proton-proton distances from MD simulation. Comparison of inter-nuclear distances of proton-proton pairs in the complexes of PKA/L_A (left panels) and PKA/L_B (right panels) derived from crystal structures (y-axes; PDB identifiers 3dne and 3dnd, after adding hydrogen atoms, see main text and Methods section 2.5.1) and those extracted from MD simulations of the two systems (x-axes). MD simulation distances were evaluated every 1 ps, taken to the power of $-1/6$ (see main text), and then averaged over the entire trajectory (30 ns). Top panels, distances converted to Å, bottom panels, distances to the power of $-1/6$ reported in $Å^{-1/6}$, on a logarithmic scale. Individual proton-proton pairs are shown as circles colored light grey if both protons belong to the protein, black if both protons belong to the ligand, and light grey, otherwise; corresponding Pearson correlation coeffi-

cients (R) of a linear regression are shown in the same colours, in the bottom right of each panel. Not best fit lines but first diagonals ($y = x$) are shown as black lines. There are 339 hydrogen atoms in PKA within a maximum distance of 10 Å of the ligand, and 8 in each ligand L_A, L_B.

The poor fit obtained with the set of order parameters derived from the MD simulation is most likely due to the radial part of the order parameter, that is the distance averaging in the MD simulation contributing to the order parameter according to Equation 2.8. This indicates that MD simulation is not able to reproduce the motional features of the complexes to a high degree of accuracy. This is in line with the notion that obtaining accurate quantitative predictions of NMR relaxation parameters from MD simulations is a challenging task (Trbovic et al., 2008; Markwick et al., 2008; Case, 2002), despite recent improvements in force field parametrisation (Showalter et al., 2007; Showalter and Brüschweiler, 2007). In general, MD simulations are able to reproduce order parameters for backbone NH and methyl group CH bond vectors measured by NMR spectroscopy (Showalter et al., 2007; Showalter and Brüschweiler, 2007; Chen et al., 2004; Ming and Brüschweiler, 2004) which contain only the angular part of Equation 2.7, i.e. Equation 2.8. It is therefore reasonable to assume that in this case the discrepancy between the experimental and back-calculated INPHARMA data can be mainly attributed to the failure of MD simulations to reproduce the distance averages $\langle r_{kl}^{-3} \rangle^2$ or $\langle r_{kl}^{-6} \rangle$ to high accuracy. A similar conclusion has already been reported before (Vögeli et al., 2009).

In the context of the INPHARMA calculations, the situation is aggravated by the inter-molecular ligand-protein distances, which are even more challenging to reproduce theoretically due to current MD force fields being less suitable for simulations of ligands than protein simulations. In order to test the ability of the MD simulation to reproduce the correct distance distribution, average distances $\langle r_{kl} \rangle$ from MD simulations were compared with distances extracted from the crystal structures which can be considered a good approximation of the average state in solution, and usually the most accurate distance information available.

The comparison reveals the poor performance of MD simulation based distance estimation, especially for protein-ligand inter-molecular distances, and in the small distance regime, which is crucial for the INPHARMA methodology (Figure 2.5).

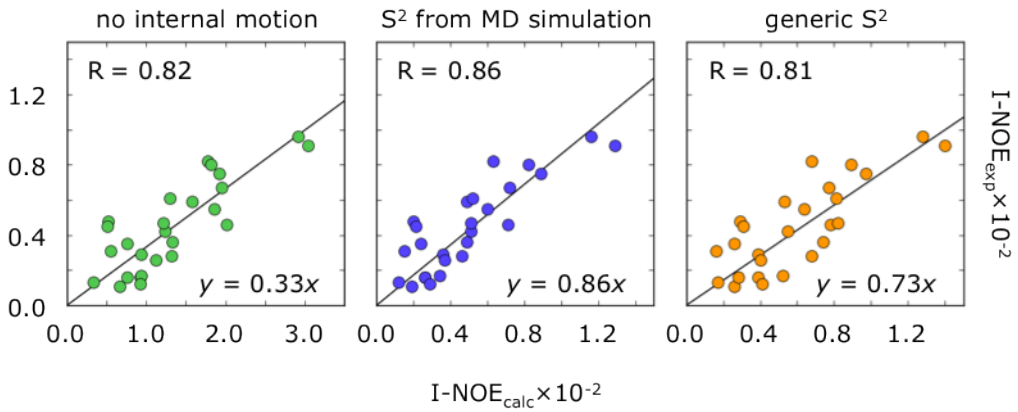


Figure 2.6 – Influence of order parameter model on quality of fit. Linear regression of experimental INPHARMA NOEs (I-NOE$_{exp}$) at misxing times of 300, 450, 600, and 750 ms versus simulated data (I-NOE$_{calc}$) ignoring (left panel, green) and considering (central and right panels) internal motion. In the central panel, tailored order parameters S^2 derived from an MD simulation of the PKA/L$_A$ and PKA/L$_B$ systems (see main text) were used (blue), while in the right panel generic order parameters were used (orange, see Results sections 2.3.3 and 2.3.4). INPHARMA cross peak intensities were normalised to diagonal peak intensities of L$_A$ in a NOESY spectrum at mixing time τ_M = 150 ms. Best fit lines ($y = ax$, black) are plotted after performing a linear regression. See Methods section 2.5.4 for technical details on the INPHARMA calculation and experimental and simulation parameters.

While the overall distance reproduction is good (R = 0.92 to 0.97) in 'real distance space', that is, Euclidean distances (see top of Figure 2.5), in *inverse* distance space such as the one relevant for NOE transfer rates, $\langle r^{-6} \rangle$, small errors amplify considerably and lead to a consistently poorer fit especially for small inter-nuclear distances (see bottom of Figure 2.5). For the two model systems PKA/L_A and PKA/L_B, the quality of the fit for protein-ligand proton pairs drops to R = 0.24 and 0.47, respectively, while it stays at R = 0.95 for protein-protein proton pairs, again illustrating the challenge of simulating protein-ligand complexes with the same quality as free proteins. It should be noted that MD simulation derived distances are consistently larger than the static distances from the crystal structures. This is in agreement with a recent study on perdeuterated ubiquitin (Vögeli et al., 2009) which shows that inaccuracies in order parameter estimations from MD simulations can be attributed to distance effects, and MD simulation derived distances exhibit a poor correlation to distances derived from cross-relaxation experiments by NMR spectroscopy. Inaccuracies in the MD simulations, especially in the short distance range that dominates the average, are expected to have a large effect on $\langle r^{-6} \rangle$, while the effect on $S^2_{r,kl} = \langle r_{kl}^{-3} \rangle^2 / \langle r_{kl}^{-6} \rangle$ (Equation 2.8) can be expected to be less. Indeed, numerical simulations have shown that a 10 % random error on the inter-nuclear distance translates to a 10 % error on the radial order parameter $S^2_{r,kl}$, but to a 25 % error on $\langle r_{kl}^{-6} \rangle$ (data not shown). Therefore, while the averaged distances $\langle r_{kl}^{-6} \rangle$ derived from MD simulation are problematic, MD simulation derived radial order parameters $S^2_{r,kl} = \langle r_{kl}^{-3} \rangle^2 / \langle r_{kl}^{-6} \rangle$ were considered to be closer to the true values, and subsequently used throughout this chapter. In order to obtain robust distance values to be used as an alternative to the term $\langle r_{kl}^{-6} \rangle$ in the spectral density functions (Equation 2.4) governing the NOE transfer rates, alternative distance estimates were considered. To this end, it was attempted to substitute $\langle r_{kl}^{-6} \rangle$ by $r^{-6}_{kl,\mathrm{cryst}}$, i.e. the crystal structure distances, raised to the power of −6. This choice was made following previous work, which had reported a good correspondence between effective distances extracted from NOE data, and distances from crystal structures (Vögeli et al., 2009). In particular, it was shown that for backbone amide inter-proton distances up to 5 Å in perdeuterated ubiquitin, the crystal structure distances were generally found within 5 % of effective averaged distances extracted from NOESY NMR experiments. This result suggests that for most

distances $r^{-6}_{kl,\mathrm{cryst}}$ might be a good surrogate of $\left\langle r^{-6}_{kl} \right\rangle$, as can be expected for internal motions of moderate amplitude.

It should be noted that by using $r^{-6}_{kl,\mathrm{cryst}}$ instead of $\left\langle r^{-6}_{kl} \right\rangle$, true distances are slightly but consistently *larger* than the crystal structure distance surrogates suggest (see Figure 2.5). As the inter-nuclear distance term in Equation 2.4 contributes to the spectral density with a *negative* exponent, however, the rate of magnetisation transfer through space is slightly *under*-estimated by using crystal structure distances as a surrogate. This is equivalent to over estimating the amount of internal motion present in the protein. However, using the more formally correct[9] but less accurate distances derived from the MD simulation, the slope of the linear fit had increased from 0.33 to 0.71, indicating the range of slope values that can be expected from a correct, rigorous distance treatment. The slope that will be achieved using $r^{-6}_{kl,\mathrm{cryst}}$, however, will be a slight over-estimation of the actual slope.

As for the uniform order parameters (see above), artificial inter-ligand NOESY spectra were calculated for the correct crystal complex structures of PKA/$\mathrm{L_A}$ and PKA/$\mathrm{L_B}$ at different mixing times τ_M, using the set of tailored order parameters S^2 from the MD simulation, and using crystal structure distances $r^{-6}_{kl,\mathrm{cryst}}$. These were compared to the experimental data reported previously (Orts et al., 2008b). Compared to the static case neglecting internal motion (R = 0.82), a slight improvement of the correlation coefficient was observed (R = 0.86, Figure 2.6). The slope of the linear fit, however, was found to rise to 0.86 from 0.33. This suggests that internal motions on a fast time scale are indeed responsible for most of the over-estimation of the INPHARMA NOEs in back-calculations using static complexes. Interestingly, setting a uniform order parameter of 0.62, equal to the average of all order parameters extracted by the MD runs, would only result in a slope of 0.60, suggesting that the tailored set of order parameters might capture specific characteristics of the interaction of protein and ligand.

[9] in the sense that their *magnitude* could be expected to be closer to the magnitude of the true distances effecting the NMR observables in a NOESY experiment.

2.3.3 Generic order parameters transferable between model systems

In a routine application of INPHARMA, neither $\langle r_{kl}^{-6} \rangle$ nor S_{kl}^2 would be available for a given protein-ligand system of interest prior to an MD simulation since sets of values depend on an actual binding pose that would be used as a starting structure for an MD simulation aimed at obtaining those values. This is clearly unfeasible if a large number of possible ligand poses needs to be considered. A tangible approach could consist in back-calculating INPHARMA NOEs using static distances from complex structural models[10] in combination with generic order parameters that can be applied to any protein system. In view of these limitations, the possibility of defining such a set of generic order parameters is explored in order to by-pass the demanding task of running MD simulations. This gains additional relevance from the fact that current force fields are designed for proteins while simulations containing small molecule ligands require manual adjustment and extension of the force field. As it is challenging to define a representative and diverse set of protein-ligand complexes from which these generic order parameters must be determined, the following analysis is limited to free proteins only. This is equivalent to making the assumptions that a) a ligand during the course of the MD simulation, and on the time scale relevant for internal motion influencing NMR observables remains firmly bound to the protein, effectively allowing for the ligand to be treated as a protein residue, and b) that a correspondence between protein and ligand atom types can be established.

To generate such generic order parameters containing both angular and radial contributions according to Equation 2.5, the following approach was followed. From MD simulations of different globular proteins, order parameters for pairs of non-exchanging protons were derived. Then, those order parameters were decomposed into contributions from individual protons (see below), and assigned to different chemical groups. Finally, group wise averaged values are proposed to be used as generic order parameters that represent the motional behaviour of equivalent chemical groups in any protein, or protein-ligand system of interest.

[10] ideally, crystal structures, as discussed above.

2.3.4 Order parameter decomposition

In general, oder parameters S^2_{kl} are defined for pairs of nuclei (k, l). To derive generic order parameters for each proton type in a protein, the first step required is the decomposition of S^2_{kl} into proton specific contributions S_k and S_l, called S-factors, such that $S^2_{kl} = S_k \cdot S_l$. It should be noted that S-factors were not devised with the intention that they would correspond to any physical quantity, but rather as a mathematical trick in order to obtain a convenient, manageable set of group wise values that can be sampled with good statistics, i.e. 9 proton groups (see below) instead of $9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1$ proton *pair* groups. In particular, this approach allows for sampling of the motions of rare protons, or rare proton combinations, such as those involving CH_2-ε (see below).

To find optimal values S_k for all protons k, the l^2 norm (Euclidean norm) of the difference $\mathbf{A} - \mathbf{x}^\top\mathbf{x}$, with $A_{kl} = S^2_{kl}$, was minimised to obtain a vector $\mathbf{x}$ with $x_k = S_k$. In other words, the pair-wise order parameter S^2_{kl} was decomposed into contributions from the single protons k and l. By minimising a single target function for all protons at the same time, all available dynamic information for the protein is considered, i.e. for a given nucleus k, S_k contains information about all S^2_{kl} for $k \neq l$.

Non-exchanging protons were classified according to the non-hydrogen atom to which they are covalently bound, yielding 9 groups: C-α, CH_3, CH_2-β, CH_2-γ, CH_2-δ, CH_2-ε, CH_2-proline, CH_1, and aromatic protons. Each of the groups was split into two sub groups, depending on whether the proton k belonging to that group is involved in a dipolar interaction with a proton l belonging to the same or to another residue. While every proton can only belong to one single group, it can belong to both of sub groups, i.e. if it has interaction partners in close distance in the same as well as in other residues. It was found that within a defined distance cutoff of 5 Å, a proton had on average 6 neighbours in the same residue and 12 in other residues. The average S-factor, calculated over S_k for all protons k inside one of those eighteen (sub) groups, represents the generic S-factor for that class. To transfer this information and assign expected motional behaviour to an unknown system, each proton k of that system is assigned an atom type, according to the

grouping scheme described above, along with a pair of corresponding S-factors.[11] The generic order parameter $S^2_{kl,\text{generic}}$ defined for the dipolar interaction between k and any other proton l is then reconstituted by multiplying the respective S-factors S_k and S_l.[12]

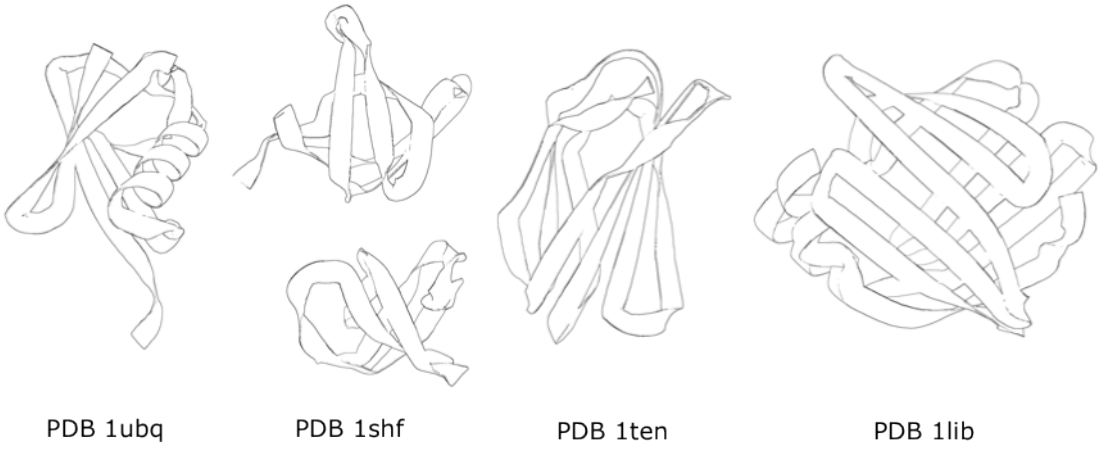


Figure 2.7 – Globular proteins used to estimate generic order parameters. Proteins are shown in cartoon representation (white) with flat sheets, round helices and smooth loop segments for simplicity. Systems from left to right are human ubiquitin (PDB identifier 1ubq), human FYN tyrosine kinase SH3 domain (PDB 1shf), human tenascin, fibronectin type III domain (PDB 1ten), and murine adipocyte lipid binding protein (PDB 1lib). In the case of 1shf, a homo dimer was simulated, but protein dynamics were only investigated for chain A (shown on top of the figure).

S-factors were obtained from 25 ns long MD simulations of four globular proteins for which a high resolution crystal structure is available from the Protein Data Bank (PDB) (Berman et al., 2000): human ubiquitin, PDB identifier 1ubq (Vijay-Kumar et al., 1987); human FYN tyrosine kinase SH3 domain, PDB 1shf (Noble et al., 1993); murine adipocyte lipid binding protein, PDB 1lib (Xu et al., 1993); and fibronectin type III domain from human tenascin, PDB 1ten (Leahy et al., 1992) (Figure 2.7).

[11] one for intra-residue and one for inter-residue interactions.
[12] depending on whether an intra- or inter-residue proton pair is investigated.

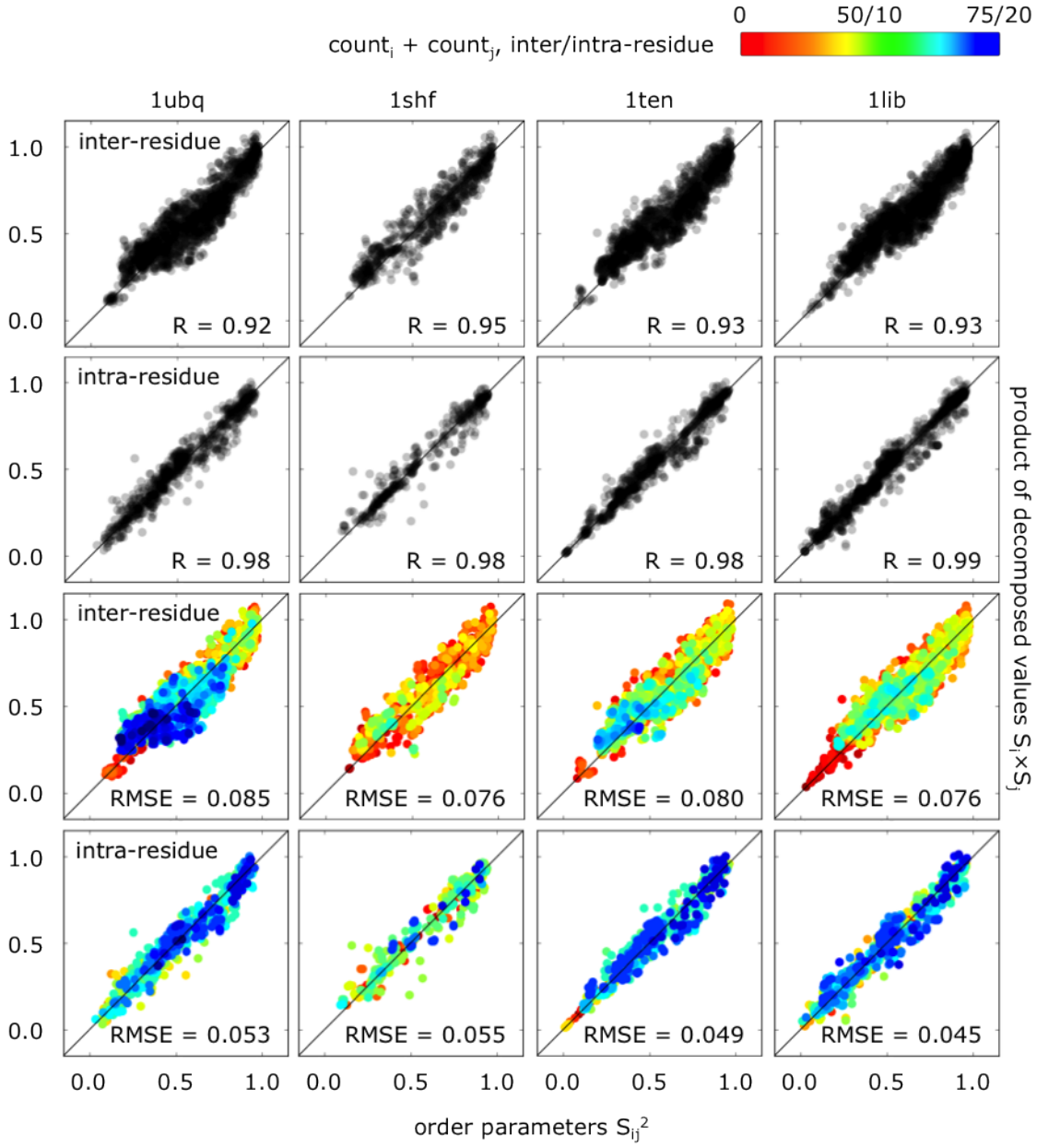


Figure 2.8 – Decomposition quality of order parameters. Values of order parameters S_{ij}^2 (x-axes) are plotted against the product of the decomposed order parameters, or S-factors, $S_i \times S_j$ (y-axes), for inter- (first and third row of panels) and intra-residue proton pairs (second and fourth row) and the set of four globular proteins (identified by their PDB ID, see main text; 1ubq, human ubiquitin, 1shf, human FYN tyrosine kinase SH3 domain, 1ten, human tenascin, fibronectin type III domain, and 1lib, murine adipocyte lipid binding protein). In the bottom two

rows, data are coloured according to the sum of the numbers each proton i and j is observed in order parameter pairs, on a rainbow colour scale from 0 to 75 (for inter-residue proton pairs) or 20 (for intra-residue proton pairs). Values of the Pearson correlation coefficient (R) and root mean square error (RMSE) for a linear regression are indicated in the bottom right of each panel. Diagonals ($y = x$) instead of best fit lines are shown as black lines. Decomposition statistics can be found in Table 2.3.

The abundance of motional information present in these four model systems allows for exclusion on non-converged correlation functions, i.e. unlike for the PKA/L_A and PKA/L_B systems, where every order parameter had to be estimated as it was relevant to track magnetisation transfer, for the estimation of generic order parameters and S-factors, only internal correlation functions that converge are integrated into the data set. For the correlation function $C_{kl}^{\text{internal}}(t)$ (Equation 2.3) to be considered to have converged to S_{kl}^2, it was required to be constant within a range around that plateau value for an extended period of time. A technical description of order parameter extraction and automated convergence tests can be found in Methods sections 2.5.2 and 2.5.6.

Table 2.1 – S-factors for 'inter-' and 'intra-residue' proton pairs extracted from four globular proteins (represented as averaged values). Standard deviations are indicated in parentheses; both angular and radial fluctuations are included in these values, with N the number of protons used to derive the corresponding S factor.

	inter-residue (SD)	intra-residue (SD)	N (inter/intra)
C-α	0.96 (0.01)	0.93 (0.03)	372/297
CH_3	0.62 (0.03)	0.56 (0.03)	611/587
CH_2-β	0.87 (0.02)	0.81 (0.04)	366/307
CH_2-γ	0.78 (0.04)	0.73 (0.04)	154/160
CH_2-δ	0.69 (0.04)	0.60 (0.02)	52/67
CH_2-ε	0.44 (0.04)	0.39 (0.05)	35/48
CH_2-proline	0.82 (0.03)	0.77 (0.04)	62/64
CH_1	0.92 (0.03)	0.79 (0.06)	98/99
Aromatic	0.78 (0.02)	0.88 (0.03)	121/124

As the dynamics of each of the proteins shown in Figure 2.7 has been investigated experimentally by NMR spectroscopy (Best et al., 2004; Mittermaier et al., 2003; Constantine et al., 1998; Lee et al., 1999), MD simulation based order parameter estimation can at least be partially benchmarked against experimental

data. For human ubiquitin, experimental order parameters are reproduced well by the MD simulation (R = 0.87, root mean square error, RMSE = 0.15 for methyl group axes and R = 0.75, RMSE = 0.07 for backbone amides). For the human FYN tyrosine kinase SH3 domain, methyl axis order parameters are in good agreement with experiments (R = 0.76, RMSE = 0.13), while experimental backbone order parameters are not available. For both the murine adipocyte lipid binding protein and the fibronectin type III domain from human tenascin, methyl axis order parameters could be well reproduced (R = 0.84, RMSE = 0.14, and R = 0.71, RMSE = 0.21, respectively), while the reproduction of backbone order parameters was less accurate (R = 0.53, RMSE = 0.14, and R = 0.62, RMSE = 0.05, respectively). For the murine adipocyte lipid binding protein, the lower quality of the fit of backbone order parameters might be attributed to the fact that the experimental dynamics studies were conducted on the human protein, while the crystal structure of the human protein protein was not available when performing the MD simulations. The human and the murine protein, which were used as the simulation structure, differ in 11 residues, which can explain differences in the dynamics of even the conserved residues.[13] Similarly, for the fibronectin type III domain from human tenascin, the dynamics studies have been performed on a construct two amino acids longer,[14] while the crystal structure was truncated. In this context, it has been reported that this C-terminal extension of the protein has a stabilising effect on the protein (Meekhof et al., 1998; Hamill et al., 1998), and therefore, minor differences in protein dynamics can be expected for the different constructs.

Encouraged by the apparent good quality of the MD simulations in reproducing the angular fluctuations, the sets of theoretical order parameters obtained from the MD simulations of all four proteins were used to derive generic order parameters. After decomposing order parameters S^2 to obtain S-factors, these S-factors were averaged for each proton class for all four proteins. Averaged, final S-factors are summarised in Table 2.1, while more detailed values of individual systems can be found in Table 2.2. In order to judge the quality of the decomposition, values

[13] obviously, only pairs of values from conserved residues were compared in the benchmark study.

[14] namely, amino acids 1 to 92.

of S^2_{kl} were compared to the products $S_k \cdot S_l$ for all nuclei (k, l). For the four test systems, average correlation coefficients of R = 0.98 (intra-residue, RMSE = 0.045 to 0.055) and 0.94 (inter-residue, RMSE = 0.076 to 0.085) were obtained, respectively (Figure 2.8 and Table 2.3).

The S-factors obtained reflect the intuition about the dynamic behaviour of a protein: C-α protons are largely static, while methyl group rotations result in low S-factors for CH_3 protons. For different CH_2 protons, an increasing mobility reflected by a *de*creasing S-factor can be appreciated when moving away from the main chain (from β to ε).

Table 2.2 – S -factors for 'inter-' and 'intra-residue' proton pairs extracted from four globular proteins (represented as averaged values *for individual proteins*). Standard deviations are indicated in parentheses; both angular and radial fluctuations are included in these values, with N the number of protons used to derive the corresponding S factor.

	human ubiquitin			human FYN tyrosine kinase SH3 domain		
	inter-residue (SD)	intra-residue (SD)	N (inter/intra)	inter-residue (SD)	intra-residue (SD)	N (inter/intra)
C-α	0.97 (0.05)	0.95 (0.12)	78/64	0.95 (0.08)	0.89 (0.16)	62/46
CH_3	0.62 (0.13)	0.55 (0.12)	146/148	0.56 (0.14)	0.57 (0.12)	81/73
CH_2-β	0.88 (0.10)	0.84 (0.16)	76/67	0.86 (0.12)	0.81 (0.16)	63/49
CH_2-γ	0.79 (0.16)	0.76 (0.16)	41/41	0.82 (0.19)	0.74 (0.16)	15/13
CH_2-δ	0.69 (0.17)	0.57 (0.12)	12/13	0.49 (0.05)	0.67 (0.21)	3/7
CH_2-ε	0.43 (0.18)	0.42 (0.17)	8/14	0.56 (0.00)	0.38 (0.03)	1/2
CH_2-proline	0.84 (0.09)	0.76 (0.12)	18/18	0.84 (0.11)	0.77 (0.12)	12/12
CH_1	0.92 (0.14)	0.75 (0.20)	24/25	0.84 (0.11)	0.62 (0.19)	11/12
Aromatic	0.76 (0.16)	0.85 (0.18)	17/17	0.80 (0.12)	0.89 (0.11)	36/40

	human tenascin, fibronectin type III domain			murine adipocyte lipid binding protein		
	inter-residue (SD)	intra-residue (SD)	N (inter/intra)	inter-residue (SD)	intra-residue (SD)	N (inter/intra)
C-α	0.96 (0.06)	0.93 (0.13)	94/75	0.96 (0.09)	0.93 (0.15)	138/112
CH_3	0.62 (0.12)	0.54 (0.11)	157/141	0.65 (0.12)	0.57 (0.10)	227/225
CH_2-β	0.86 (0.12)	0.83 (0.18)	91/73	0.87 (0.12)	0.78 (0.23)	136/118
CH_2-γ	0.78 (0.17)	0.72 (0.21)	38/36	0.76 (0.17)	0.71 (0.20)	60/70
CH_2-δ	0.70 (0.21)	0.65 (0.20)	8/10	0.70 (0.18)	0.59 (0.23)	29/37
CH_2-ε	0.51 (0.18)	0.40 (0.18)	8/6	0.40 (0.22)	0.38 (0.15)	18/26
CH_2-proline	0.80 (0.08)	0.77 (0.13)	29/30	0.83 (0.20)	0.86 (0.11)	3/4
CH_1	0.92 (0.11)	0.84 (0.11)	25/25	0.94 (0.09)	0.84 (0.15)	38/37
Aromatic	0.83 (0.09)	0.89 (0.17)	22/24	0.74 (0.24)	0.86 (0.22)	46/43

Table 2.3 – Order parameter decomposition statistics for four globular proteins. Numbers of inter- and intra-residue order parameters (corresponding number of individual, unique protons given in parentheses) are given as well as values of the Pearson correlation coefficient (R) and the root mean square error (RMSE) of original (S^2_{ij}) versus the product of decomposed values $S_i \times S_j$ (see Figure 2.8) for each system. For final values of S for individual atom types and all systems see Tables 2.1 and 2.2, respectively.

	human ubiquitin		human FYN tyrosine kinase SH3 domain	
	inter-residue	intra-residue	inter-residue	intra-residue
number of order parameters (unique protons)	2,187 (420)	897 (407)	776 (284)	434 (254)
R	0.92	0.98	0.95	0.98
RMSE	0.085	0.053	0.076	0.055

	human tenascin, fibronectin type III domain		murine adipocyte lipid binding protein	
	inter-residue	intra-residue	inter-residue	intra-residue
number of order parameters (unique protons)	1,816 (472)	918 (420)	2,399 (695)	1,383 (672)
R	0.93	0.98	0.94	0.99
RMSE	0.080	0.049	0.076	0.045

2.3.5 Order parameter validation and performance of generic order parameters in INPHARMA calculations

Next, the set of generic order parameters was used for the back-calculation of magnetisation transfer in the PKA system. Protein protons were assigned to chemical groups as described above, while ligand protons were assigned to equivalent groups as if they would belong to the protein, i.e. the 'aromatic' or 'CH_2' group (compare Figure 2.2). S-factors of individual protons were re-multiplied to retrieve generic order parameters and used in the INPHARMA back-calculation of inter-ligand relayed NOE peaks. Similarly to what has been observed for the set of PKA specific order parameters S^2 (see above), the use of generic order parameters resulted in an increased slope of the best fit line of $a = 0.73$ (see right panel of Figure 2.6). The correlation coefficient was found to be R = 0.81 which is only slightly worse than that obtained with the set of PKA-specific order parameters, and very similar to the value for the rigid model (Figure 2.6).

This result confirms the usefulness of the generic order parameter concept for the INPHARMA back-calculations. To further validate the sets of order parameters, their performance was investigated in a randomisation trial (Figure 2.9). Different sets of random order parameters were created and it was evaluated how well these sets would reproduce experimental data in INHPARMA calculations, in comparison to the PKA-specific, and generic set of order parameters, as well as the rigid model ($S^2 = 1$). There were 1,000 sets of random order parameters created by three different approaches: (1) the set of original, PKA-specific order parameters were shuffled, i.e. each pair of nuclei randomly was assigned an order parameter originally derived for *another* pair of nuclei; (2) random order parameters were drawn from a Gaussian distribution centred at 0.62 and with a standard deviation of 0.22, corresponding to mean and standard deviation of the PKA-specific set of order parameters; and (3) random order parameters were drawn uniformly from $[0, 1]$.

It can be appreciated that both the shuffled set of order parameters and the set drawn from a Gaussian distribution cluster in the same region of quality space,[15] while the uniform set samples a wider range. The sets of PKA-specific and generic

[15] spanned by the Pearson correlation coefficient R and the slope a of a linear fit $y = ax$.

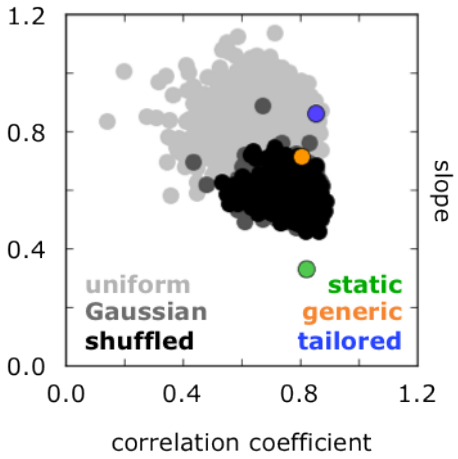


Figure 2.9 – Validation by order parameter randomisation. INPHARMA NOEs which were back-calculated using different sets of (randomised) order parameters, were linearly fitted to experimental data. Each individual set of order parameters is represented as a single dot with its x-coordinate corresponding to the correlation coefficient and the y-coordinate to the slope a of the respective best fit line $y = ax$. Colours are coded as follows: black, 1,000 sets of order parameters randomised by shuffling assignments between proton pairs and order parameters; dark grey, 1,000 sets of order parameters drawn from a Gaussian distribution with mean 0.62 and standard deviation 0.22 to resemble the order parameter set derived from the MD simulation (see main text); light grey, 1,000 sets of order parameters drawn uniformly from $[0, 1]$. Best fit parameters from linear regressions in Figure 2.6 are given for reference, green, static model, blue, order parameters derived from MD simulation, and orange, generic order parameters (see main text).

order parameters however, dominate most solutions according to a Pareto criterion of multi dimensional optimisation (Figure 2.9). This demonstrates that the set of order parameters extracted from the MD simulation is appropriate to describe protein dynamics.

2.3.6 Impact of order parameters on the discrimination power of INPHARMA

The generic order parameter concept allows for using them in INPHARMA calculations of any complex of interest. A relevant question is whether the representation of internal motions through generic order parameters can improve the selection of ligand binding poses, for example by providing a more clear cut discrimination in the linear fit of experimental data to the correct or incorrect binding poses. In order to address this question, the experimental data obtained for the complexes PKA/L_A and PKA/L_B (Orts et al., 2008b) were used to select from among 16 possible pairs of complex structures. These pairs of complex structures resulted from the combination of four different possible binding modes of L_A and four different binding modes of L_B, which all differ from each other by 180 degrees rotations around three orthogonal axes (Figure 2.10).

In Figure 2.11, correlation coefficients and slopes of linear fits of experimental data versus back-calculated data for all 16 complex pairs with and without including generic order parameters into the INPHARMA calculations are shown. When internal motion is ignored, four models show a correlation coefficient $R > 0.80$ and therefore pass the INPHARMA selection criterion as defined previously (Orts et al., 2008b). This made necessary further discrimination between those four passing binding poses by additional criteria, such as the systematic deviation of INPHARMA peaks originating from different structural moieties of the ligands, and the semi-quantitative use of additional weak INPHARMA peaks obtained at a 900 MHz NMR spectrometer.

However, when using the generic order parameters to account for internal motion, a much better discrimination of the binding poses was achieved (Figure 2.11). Both the high correlation coefficient and the slope of the crystal structure pair of PKA/L_A and PKA/L_B are optimal across all solutions, and the quality of the

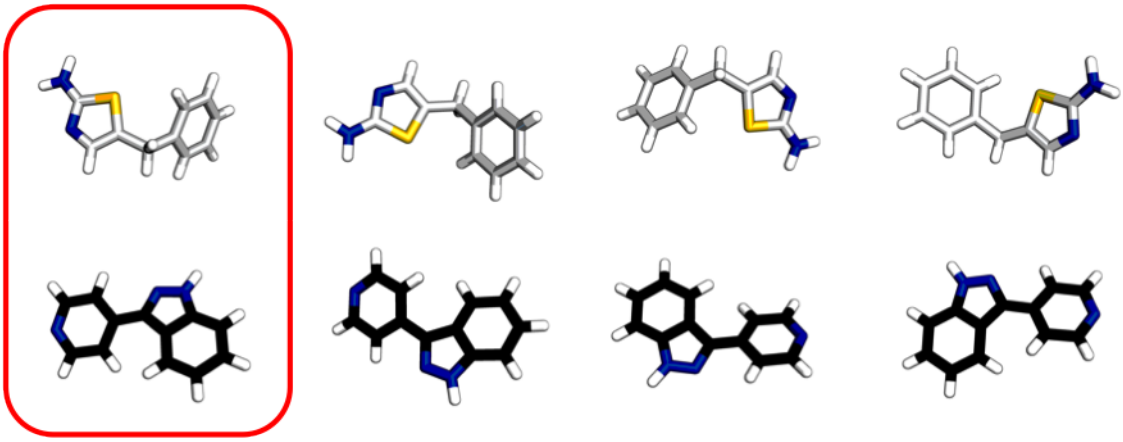


Figure 2.10 – PKA ligand true and decoy conformations. Atoms are represented as sticks (white: hydrogen, blue: nitrogen, yellow: sulphur, black/grey: carbon). True ligand conformations are indicated by a red box and were taken from crystal structures with PDB identifiers 3dne (L_A, upper panels) and 3dnd (L_B, upper panels), respectively. Decoys were generated from the true conformations by rotating the ligands around three orthogonal axes by 180 degrees, and the discrimination power of the method was probed on by its ability to retrieve the true ligand pair conformation from the set of all 16 combinations (see Figure 2.11).

three alternative solutions (see above) falls below the INPHARMA discrimination threshold. In contrast to evaluating the correlation coefficient R and the slope a separately, a composite quality factor of the type $(m(1-R)^2 + n(1-a)^2)^{-1}$ for some positive numbers m and n can be applied to select the pair of binding modes that is in best agreement with experimental data. As evident from Figure 2.11, the composite quality measure achieves a strikingly better discrimination when using a description of internal motions through generic order parameters.

2.4 Discussion

In this chapter, the INPHARMA methodology has been extended by incorporating a rigorous representation of protein internal motion into the underlying physical model, based on the concept of the NMR order parameter S^2. Using uniform order parameters $S^2 < 1$ and synthetic reference data (Figure 2.4), it was found that over a wide range of protein rotational correlation times τ_c, the slope of a linear fit line between spectra calculated for the motional versus a rigid model showed an

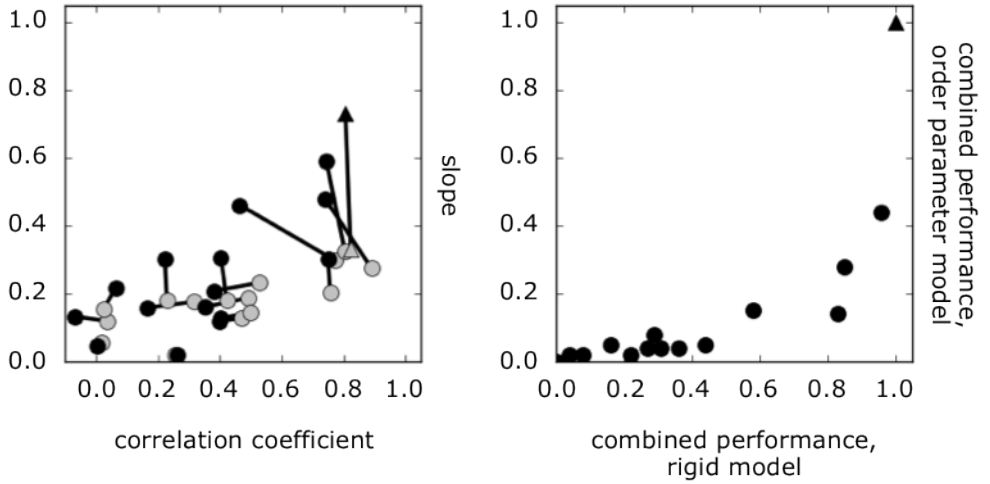


Figure 2.11 – Discrimination power of INPHARMA considering protein internal motion. Comparison of the selectivity of INPHARMA including (black) or excluding (grey) internal motion. For four different binding poses per PKA/L_A and PKA/L_B complex (yielding 16 combinations, see Figure 2.10), INPHARMA NOEs were calculated with and without a protein internal motion model, and compared to the experimental spectra. The *correct* pair of conformations, as observed in the crystal structures of PKA/L_A and PKA/L_B (PDB identifier 3dne and 3dnd, respectively) are indicated by a triangular symbol, other, incorrect combinations as circles. Left panel, slopes and Pearson correlation coefficients for best fit linear regression lines for static and internal motion models, with corresponding combinations connected by straight lines. Right panel, combined quality factor of static versus internal motion model; Pearson correlation coefficients R and linear regression ($y = ax$) best fit slopes a were combined according to the formula $(m(1-R)^2 + n(1-a)^2)^{-1}$ with equal weighting $m = n = 1$, and resulting values normalised to the interval $[0, 1]$.

approximately quadratic response,[16] and the quality of the fit was robust. Order parameters extracted from MD simulations were found to increase the slope of back-calculated versus experimental data to nearly quantitative agreement (Figure 2.6), while inter-nuclear distances derived from the simulations proved to be unreliable in the distance regime d^{-6} that is relevant for NOE mediated magnetisation transfer (Figure 2.5). This problem was addressed by using crystal structure derived distances instead of MD simulation derived distances, but comes at the expense of possible over-estimation of the effect of the order parameter based reduction of the efficiency of magnetisation transfer. A value of the best fit slope between $a = 0.71$, as obtained using MD derived average distances, and $a = 0.86$, as obtained using crystal structure distances, seems realistic. This poses a substantial improvement over the slope of $a = 0.33$ obtained using the rigid model without incorporating internal motion ($S^2 = 1$).

A set of generic order parameters was obtained from the simulation of globular proteins, and these order parameters were mathematically decomposed into S-factors, allowing them to be transferred between different systems of interest. The decomposition preserved the characteristic motions of individual proton pairs with excellent quality (Figure 2.8). It was found that they comply with intuition about the motion and rigidity of different chemical groups, i.e. low amplitude motion of backbone protons, and the amplitude of side chain proton motions increasing with their degree of separation from the main chain (Tables 2.1 and 2.2). The generic order parameters were useful to increase the discrimination power of INPHARMA by an improved ability to identify true ligand conformations of two PKA ligands from a pool of decoy conformations (Figure 2.11).

The following sections will discuss certain aspects of the INPHARMA calculation using the order parameter concept in more detail, before concluding the chapter with outlook on future directions of the work.

[16] that is, for a given order parameter $S^2 = b$, the slope would assume a value of $a = b^2$.

2.4.1 Order parameters improve the quantitative agreement between calculated and experimental reference data

It has previously been discussed that when using uniform order parameters $S^2 < 1$, the quality of the fit largely remains constant over a wider range of possible receptor sizes signified by their corresponding protein rotational correlation coefficient τ_c, with the slope of the linear fit line between motional and rigid model showing an approximately quadratic response (Figure 2.4, see above). For the test system under consideration, PKA and ligands L_A and L_B, the correlation coefficient drops for very large τ_c. This is most likely due to the fact that spin diffusion is now so efficient that the magnetisation can take many different routes within the protein, and that the magnetisation finger print the ligands impose to the protein smears out. More important, however, is the fit between calculated and *experimental* data;[17] as shown in Figure 2.12, the quality of the fit is robust for uniform order parameter values $S^2 \geqslant 0.1$, while the quantitatively correct slope a of the linear regression is reached for uniform order parameters between $S^2 = 0.4$ and 0.5. This is in line with the observation that an order parameter of about $S^2 = 0.4$ is necessary to reach a slope of $a = 0.33$ when *rigid* synthetic reference data are used (top right panel of Figure 2.4, yellow contour line).

2.4.2 Order parameters reflect protein motional behaviour

Both in the Theory section 2.2 and the presentation of the results of the estimation of generic order parameters it has been mentioned that the order parameter S^2 itself represents the plateau value of the correlation function $C(t)$ (Equation 2.2). For the order parameter to have any physical meaning, $C(t)$ must decay to this plateau value and remain stable for a range of time offsets t. Since S^2 represents *fast* internal motion, the time scale relevant for this convergence of $C(t)$ is the nano second time scale. It was found that for the generic order parameters derived from globular proteins, not all correlation functions converge (see below); indeed, closer

[17] in Figure 2.4, spectra calculated using the motional model and different uniform values of $S^2 < 1$ were compared to *synthetic reference data* with $S^2 = 1$.

inspection of individual correlation functions points to slow and medium, larger scale motions that cannot be captured meaningfully by an order parameter S^2.

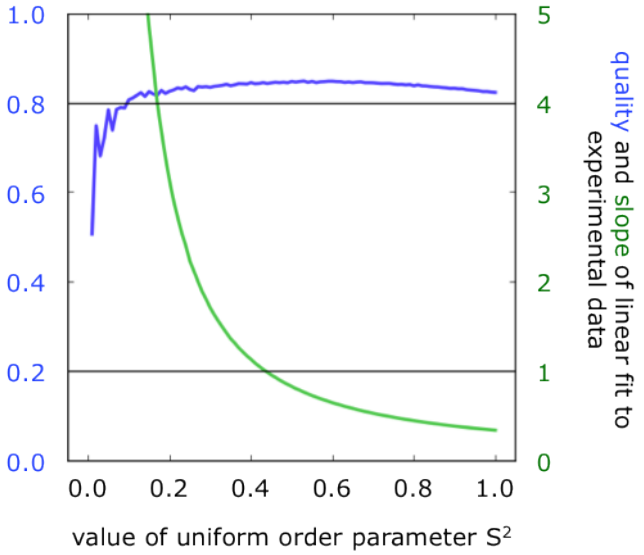


Figure 2.12 – Uniform order parameter fit versus experimental data. Uniform order parameters S^2 (x-axis, $S^2 \in]0,1]$, $\Delta S^2 = 0.01$) were applied to the PKA/L_A and PKA/L_B systems, and linear regression analyses were conducted against experimental INPHARMA data as in Figure 2.6. Both the value R of the Pearson correlation coefficient (blue, y-axis) as well as that of the slope a of the fit line $y = ax$ (green, y-axis) are shown in dependence of the uniform order parameter value. Horizontal black lines indicate values of $R = 0.80$ and $a = 1.0$.

Indeed, for human ubiquitin, 56.6 % of the 5,449 correlation functions C_{kl} converged, while 92.4 % of 489 individual protons of the system were found to have at least one correlation function that converges. For a monomer of human FYN tyrosine kinase SH3 domain, 40.5 % of the 2,992 correlation functions converged with 89.6 % of 345 individual protons having at least one converging correlation function. For fibronectin type III domain, 51.6 % of correlation functions out of 5,295 converge, and 93.5 % of 539 protons possess at least one converging correlation function. For murine adipocyte lipid binding protein, 50.1 % of 7,556 correlation functions converge, and 93.1 % of 796 individual protons have at least one converging correlation function.

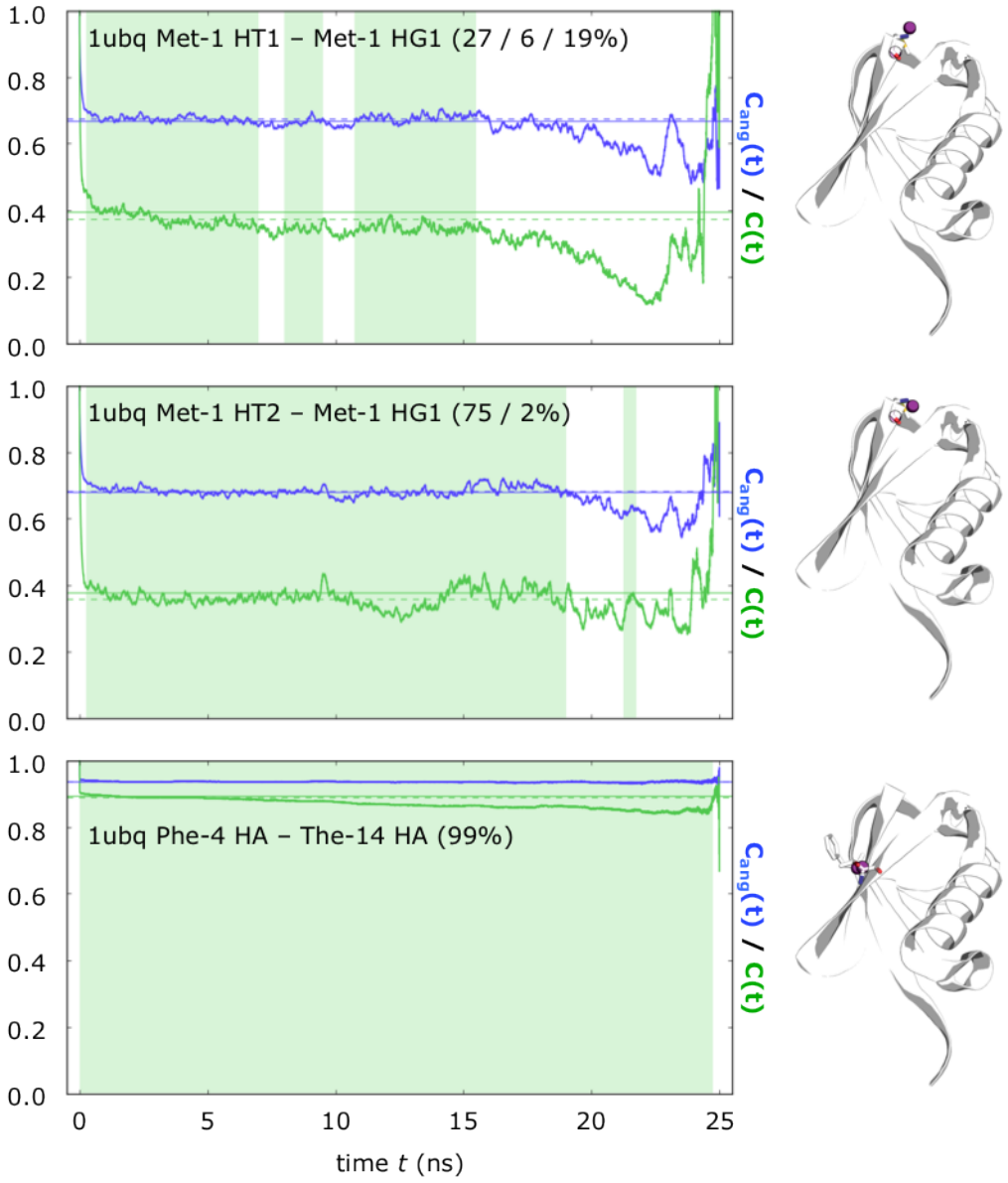


Figure 2.13 – Internal correlation functions of representative proton pairs, part 1. Correlation functions were computed from snapshots of 25 ns long molecular dynamics (MD) simulations of human ubiquitin (1ubq) and murine adipocyte lipid binding protein (1lib). Conformational snapshots were saved every 1 ps, and correlation functions containing only angular motion ($C_{\mathrm{ang}}(t)$, blue line) or angular as well as radial motion ($C(t)$, green line) were computed from the sets of coordinates. Corresponding order parameters S^2 are shown as horizontal lines, and were used to judge convergence of the correlation function. Solid horizontal lines represent the reference order parameter computed as described in the main text, which was

used to judge convergence, and dotted horizontal lines represent the order parameter computed from the average of the correlation functions $C_{\mathrm{ang}}(t)$ and $C(t)$ for 1 ns $\leq t \leq$ 6 ns for comparison. Regions of the correlation function $C(t)$ which were considered as *converged* by an automated method (see main text) are shaded in green, and the fraction of the complete time domain that they represent are indicated in parentheses. The location of the respective proton pair within the protein structure (white, in simplified cartoon representation) is shown on the right, with side chains of residues involved shown as sticks (white, carbon; blue, nitrogen; red, oxygen; yellow, sulphur) and protons under consideration as magenta spheres.

Figures 2.13 and 2.14 as well as Supplementary Figures S2 and S3 in the Appendix show correlation functions and a convergence analysis for select proton pairs within two of the test systems (see Methods section 2.5.6 for details on how to compute correlation functions $C(t)$ from MD trajectories). It can be appreciated that the correlation functions of some proton pairs, such as those located in certain aromatic ring moieties, converge, whereas others do not. Their convergence indeed depends on whether the corresponding ring moiety displays ring flips[18] or not. An analysis of ring flips based on the MD trajectory for two of the four proteins is depicted in Supplementary Figure S4 in the Appendix. It is evident that in the case of frequent ring flips, the correlation function $C(t)$ oscillates considerably, thus making the meaningfulness of any order parameter S^2 void. This outcome is expected from the time scale on which these motions happen (Supplementary Figure S4), namely, high nano second time scale. If enough ring flips are observed during the course of the simulation, $C(t)$ converges; otherwise, the sampling problem inhibits the accurate estimation of an order parameter. The limited length of the MD trajectories (25 ns) might therefore pose a limitation to the accuracy of the estimation of such order parameters (Pfeiffer et al., 2001), leading to an underestimation of the fraction of correlation functions that actually converge. However, in other cases, such as for two glycine residues in the flexible C-terminus of human ubiquitin (top panel of Figure 2.14), large scale motion that is slow compared to the NMR time scale is detected, and no meaningful NMR order parameter can possibly be deduced.

[18] that is, 180 degree inversions of their dihedral angle χ_2 during the course of the MD simulation.

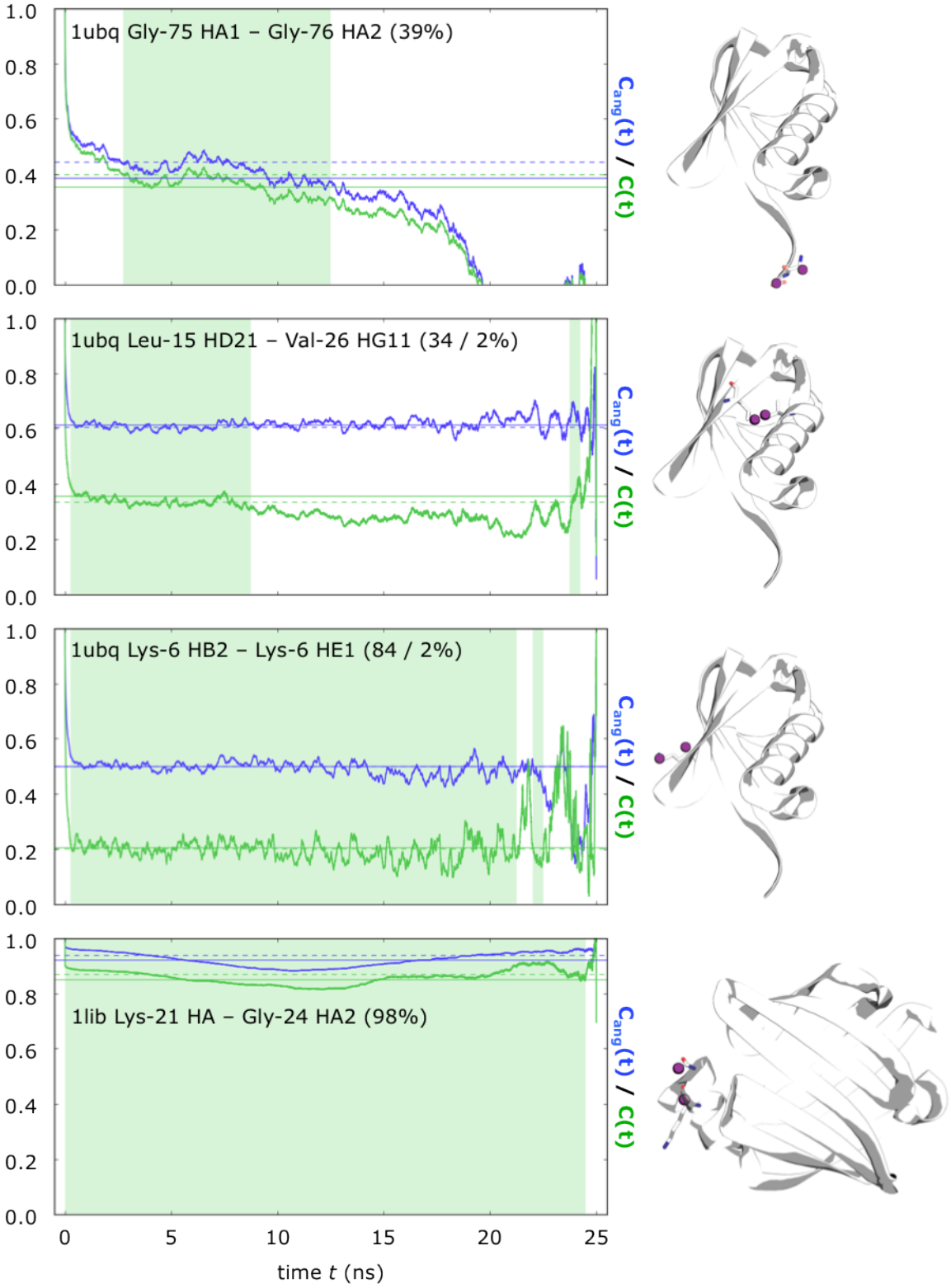

1ubq Gly-75 HA1 – Gly-76 HA2 (39%)
1ubq Leu-15 HD21 – Val-26 HG11 (34 / 2%)
1ubq Lys-6 HB2 – Lys-6 HE1 (84 / 2%)
1lib Lys-21 HA – Gly-24 HA2 (98%)
C_ang(t) / C(t)
time t (ns)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25

Figure 2.14 – Internal correlation functions of representative proton pairs, part 2. See Figure 2.13 for reference.

In the case of the estimation of generic order parameters, instances where convergence could not be detected over the course of the simulation were discarded and did not contribute to the subsequent decomposition. In the case of the derivation of the *tailored* set of PKA/L_A and PKA/L_B specific order parameters, however, convergence could not be taken into account, as order parameters were required for *all* proton pairs, since the entire spin network had to be analysed in the full relaxation matrix approach. An omission of individual, possibly not converged order parameters was anticipated to be worse than the inclusion of an amount of possibly not converged order parameters.

It should be noted that in general, non-converging correlation functions tend to have a lower average value than converging correlation functions. Consequently, the exclusion of non-converging correlation function from the analysis leads to a *higher* generic order parameter and hence higher S-factors, while it possibly also leads to a slight over-estimation of the effect of order parameters on the magnetisation transfer in the PKA systems. Unfortunately, there is no straight-forward way of circumventing this problem. In some instances, however, which seem to have non-converging correlation functions due to a sampling problem, such as protons in rarely flipping ring systems, an order parameter might still be meaningfully estimated from a longer trajectory although it might evade automatic detection of its convergence. Manual inspection has shown that for some of these cases, the apparent order parameter computed from Equation 2.5 still makes sense, as it approaches the mean of the oscillating corresponding correlation function. Most probably, for the proton pairs most relevant for the INPHARMA methodology, namely those close to the ligand binding site, the issue of convergence will be least relevant, as they will likely be situated buried within the protein, close to its hydrophobic core, where large scale motions giving rise to the problems discussed above are unlikely to occur.

Owing to its functional form (Equation 2.15 in Methods section 2.5.6), the correlation function $C(t)$ can be sampled more accurately for small t since more

pairs of MD trajectory snapshots contribute to the average.[19] Therefore, for large t in relation to the overall simulation time, sampling becomes difficult and error prone. Because of that, it has been suggested that reliable correlation functions $C(t)$ are to be estimated up until half of the trajectory length, and order parameters to be determined from the last third of that correlation function (Hornak et al., 2006). The correlation requirement employed in the present work however is purposely more stringent than that (see Methods section 2.5.6), forcing a faster decay of the correlation function $C(t)$ to its plateau value S^2, therefore limiting the analysis to *faster* internal motions. This is based on the assumption that for medium to slow internal motions indicating large scale motions, the progress of the correlation function $C(t)$ depends more strongly on specific characteristics of the protein system, while it is anticipated that fast motions are more comparable between different protein systems. As the generic order parameter approach is intended to be transferred between different protein systems, a possible bias of the convergence filtering towards fast motion that can be extracted from finite MD trajectories in a statistically sound manner is considered beneficial.

2.4.3 PKA flexibility as possible reason for the over-estimation of magnetisation transfer efficiency

It is beyond the scope of this thesis to conduct a detailed conformational analysis of PKA using simulation approaches. Furthermore, large scale motions are unlikely to be sampled by the MD simulations conducted, which aim at extracting pico- to nanosecond timescale motions that are relevant for the NMR order parameter. On the other hand, PKA was found to undergo large scale ligand induced opening and closing motions in solution during its catalytic cycle by experimental methods (see Introduction section 1.2.2). In line with experimental findings, analysis of the MD trajectory for both PKA ligand complexes simulated in this chapter revealed that the N-lobe of the enzyme was more flexible than the C-lobe, and in particular, localised backbone motions of the glycine-rich loop were detected (see Introduction

[19] that finding itself can be used for convergence estimation of the correlation function $C(t)$, i.e. by comparing the difference of values of $C(\infty)$ and C_{tail}, and rejecting instances where that difference is large (Chen et al., 2004).

section 1.2). Furthermore, visual inspection revealed the regulatory spine region and the salt bridge formed by residues Lys^{72} and Gly^{91} to remain intact during the course of the simulation.

The PKA structure used for the back-calculation of INPHARMA NOEs in this chapter was solved by X-ray crystallography in the presence of an inhibitor peptide mimicking the peptide substrate (see Introduction section 1.2). In this ternary protein-ligand-substrate complex, the enzyme assumes a fully closed conformation with both lobes adopting a compact mutual position. Binary complexes and the apo enzyme are known to adopt a more open global conformation, with the N-lobe displaced with respect to the C-lobe in order to facilitate nucleotide access to the binding cleft that appears wider. Importantly, a binary protein-ligand complex in the absence of substrate has been used to generate the experimental NMR data this chapter is based on. It has been suggested by NMR studies in the literature, however, that even in the ternary complex, fast conformational exchange exists between open and closed conformations of the enzyme (Masterson et al., 2010). Exchange rates were estimated as $k_{\text{ex}} \approx 200$ Hz and $k_{\text{open}} \approx 10$ Hz and $k_{\text{close}} \approx 100$ Hz. While this is still fast on the NMR timescale, enzyme opening and closing is slow as compared to the ligand exchange rates which in the case of the low affinity ligands L_A and L_B used in this study is faster than 1,000 Hz (see Methods section 2.5.4 and above). This indicates that in solution, in the presence of a high affinity ligand such as the nucleotide analogue used in the study mentioned above, an equilibrium exists of open and closed enzyme conformations in solution. In the case of the experimental INPHARMA study that has generated the data used in this chapter, a population estimate in the neighbourhood of 10 % of the open enzyme form of PKA existing at any one given point appears to be conservative given the low affinity of the ligands used.

Importantly, repeating the INPHARMA back-calculations of inter-ligand NOE intensities using the *open* enzyme conformation (complexes modelled on the structure of the apo enzyme, PDB identifier 1j3h) revealed a slope of a linear fit between calculated and experimental intensities of $a = 1.95$ to 2.41, depending on whether generic or tailored order parameters were used. At the same time, the quality of the fit diminished considerably to around $R \approx 0.5$. However, assuming the opening and closing chemical exchange of the enzyme to be slower than the ligand binding event,

effective INPHARMA cross peak intensities would be the average of the intensities of the closed form of the enzyme (calculated and described earlier in this chapter, section 2.3.2) and the open form of the enzyme (described in this section), weighted by their relative populations in solution. In particular, each inter-ligand NOE intensity ILOE_i could be calculated as $\mathrm{ILOE}_i = p_{\mathrm{open}} \cdot \mathrm{ILOE}_{i,\mathrm{open}} + p_{\mathrm{closed}} \cdot \mathrm{ILOE}_{i,\mathrm{closed}}$ for $p_{\mathrm{open}} + p_{\mathrm{closed}} = 1$ the relative populations of open and closed enzyme conformation, respectively. Intriguingly, using a model like this and values of p_{open} between 10 and 20 %, a slope of linear fit of $a = 0.93$ to 1.01 can be achieved, while the quality of the fit remains constant over a considerable range of values up to $p_{\mathrm{open}} \approx 0.3$. Compared to the values reported earlier in this chapter (Figure 2.6 in section 2.3.2), assuming $p_{\mathrm{open}} = 0$, where a slope of $a = 0.86$ was reported, assuming a population of the open enzyme conformation with $p_{\mathrm{open}} > 0$ to be present clearly moves the slope of the fit in the desired direction. In summary, the residual overestimation of the efficiency of magnetisation transfer that was described for the PKA system in this chapter, manifesting in a slope of $a < 1.0$ even when using NMR order parameters $S^2 < 1$, could at least in part be explained by the very flexible nature of PKA in solution that has been observed by other research groups before (see Introduction sections 1.2 and 1.2.2). It is therefore important to obtain data on other systems, preferably systems that contain proteins less flexible than PKA (see section 2.6). Along these lines, for protein systems alternative to PKA, preliminary research has indeed indicated the slope of the fit of calculated and experimental data to get closer to 1 (Teresa Carlomagno, personal communication).

2.4.4 Outlook

In the following two sections, extensions of the INPHARMA methodology will be discussed that transcend its original application, the selection of relative ligand orientations with respect to experimental NMR data. In section 2.4.5, the amount of information present about the *receptor* in an INPHARMA inter-ligand spectrum will be investigated, aiming at obtaining a pseudo receptor model from a strictly ligand detected NMR experiment. In section 2.4.6, a direct structure calculation approach will be discussed using INPHARMA data, much like in traditional NMR

based protein structure determination, thus by-passing the need for generating a pool of combinations of ligand orientations from which INPHARMA then selects the best solution, but rather becoming independent of this external step. Both extensions of the methodology have been explored to some degree, and initial results and challenges encountered will be discussed. Given the time constraints on this thesis work, these projects could unfortunately not be pursued any further. However, this presentation will serve as a starting point of future work.

2.4.5 Pseudo receptor construction from INPHARMA data

It is tempting to investigate the amount of information present about the *receptor* in an INPHARMA spectrum. Given the process of spin diffusion, indirect, receptor mediated magnetisation transfer between the ligands, information about the spatial arrangement of receptor protons should be imprinted on the signal between the ligands. If this information could be made use of, it would be of strong immediate interest, as the approach would pose a unique perspective of obtaining receptor information from a ligand detected experiment. As it has been noted before, INPHARMA does not suffer from traditional limitations of protein detected NMR experiments, in that large proteins are actually of advantage for INPHARMA as they amplify the signals of ligands in fast exchange to a greater extent.

Ideally, a natural extension of the INPHARMA approach would allow for determination of the shape, and spatial arrangement of protons within the ligand binding pocket of a protein, as experienced by the ligand throughout an INPHARMA experiment. Clearly, receptor proton positions determine the route of spin diffusion, and therefore modifications of receptor proton positions will result in a different INPHARMA inter-ligand spectrum. If the relative position[20] of the ligands is known, perhaps the positions of the receptor protons could be deduced in a similar fashion as the relative ligand orientation is reconstructed in the original INPHARMA approach, namely, by selecting from a pool of pre-generated possible structures in best agreement with experiment. In this context, this would be the local conformation of the receptor binding pocket, or at least a pseudo receptor

[20] or, superimposition.

model. If this was indeed possible, the approach would amalgamate structure- and ligand based drug design in a novel fashion, providing complementary data.

Preliminary work has been carried out to address the feasibility of this approach. It was anticipated that the ability of the approach to discriminate between different receptor proton arrangements would crucially depend on the number of receptor protons that actually actively contribute to the INPHARMA inter-ligand spectrum by relaying information in the form of magnetisation. Interpreting the receptor protons around the ligand(s) as being part of discrete protons shells, with the first shell in direct contact or close proximity to the ligands themselves, and the second shell in contact with the first shell, and so on, it is plausible that receptor protons in further shells contribute less specifically, and perhaps interchangeably, to the inter-ligand signals. The position of these ambiguous protons would be very challenging to estimate for any approach. To approximate the shape of the ligand binding pocket, however, proton shells closer to the ligand would be most interesting, given they can be determinated to a certain degree of completeness. It is obvious that the possible success of such a method fundamentally depends on how important individual receptor protons are for the reproduction of INPHARMA spectra. If only a few receptor protons contribute to the spectra, this would pose a fundamental limitation to the success of any conceivable method.

For the PKA/L_A and PKA/L_B complex (Figures 2.2 and 2.3), it was estimated by visual inspection that an excess of > 20 protons would be necessary to approximate the shape of the binding pocket (data not shown). It should be noted that L_A and L_B are comparatively small ligands, and that for larger ligands or binding pockets, a higher number of receptor protons might be necessary. Furthermore, it is evident that the ligand protons have to sample the important positions of the binding pocket sufficiently.

Therefore, first the question was asked how important each individual receptor proton was for the generation of the INPHARMA inter-ligand spectrum. To this end, in a *leave on out* approach, reduced receptors were constructed by omitting single receptor protons one at a time, and inter-ligand spectra of these reduced receptors were computed and compared to a synthetic reference spectrum constituted by the full receptor. In this fashion, each receptor proton was assigned an *importance* measure expressing how much its omission would influence the result-

ing spectrum. Artificial spectra from reduced and full receptors were compared by computing the Pearson correlation R coefficient between them, and an importance measure $I = 1 - R^2$ was introduced which would assume a value of 0 if that receptor proton was completely disposable,[21] or 1, if it would singularly determine the spectrum on its own.[22]

Figure 2.15 shows a comparison of receptor proton importance for each of the 339 protons in PKA that are within a maximum distance of 10 Å of the ligands, for different protein rotational correlation times τ_c, and two or five ligands. The two ligands used were the familiar L_A and L_B (Figure 2.2), with three additional ligands used if indicated. It can be appreciated that for small correlation times $\tau_c = 17$ ns, strikingly few ligands impact the spectrum, with all bar two or three receptor protons being assigned a Pearson correlation coefficient of $R > 0.99$ upon their omission. This is slightly alleviated by including the three additional ligands. Going to larger receptor sizes, i.e. artificially increasing the correlation time τ_c to 250 or 1,000 ns, more receptor protons become important, but the mutual differences in importance attenuate, i.e. more receptor protons share the responsibility for magnetisation relay. This is reminiscent of the fact that spin diffusion becomes more efficient in larger receptors,[23] and therefore more receptor protons contribute to the INPHARMA inter-ligand spectrum.

Next, the importance information was used to rank the receptor protons, and construct a complete series of reduced receptors PKA_i consisting only of the i most important protons ($i = 1..339$). Again, inter-ligand spectra for PKA_i/L_A, PKA_i/L_B, .. were constructed, and compared to synthetic reference spectra using the same adjustable parameters, but the *complete*, non-reduced receptor PKA consisting of all 339 protons within a maximum distance of 10 Å of the ligands.

[21] i.e. its omission would not change the INPHARMA spectrum at all.
[22] a merely theoretical possibility.
[23] those with larger τ_c.

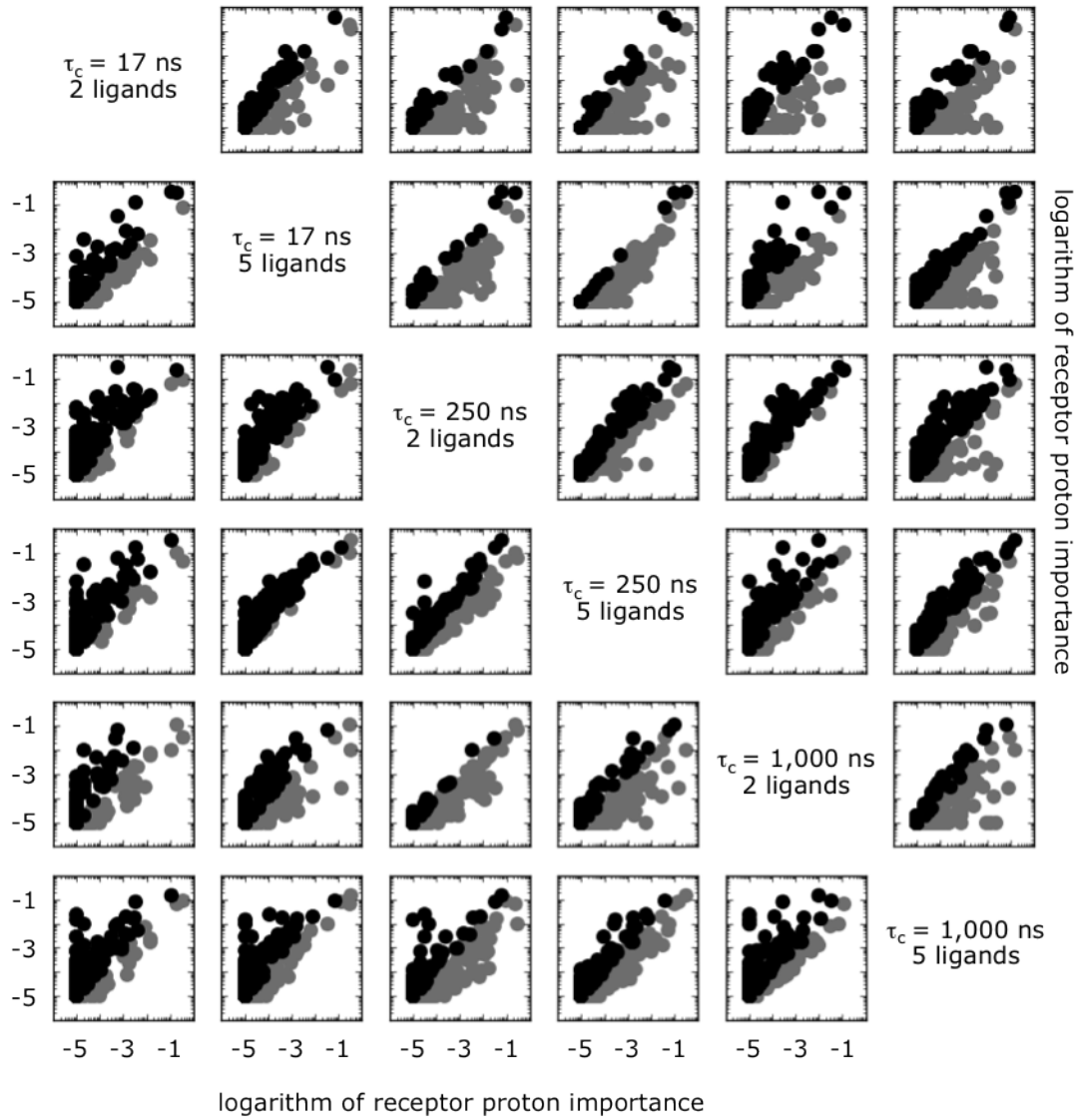


Figure 2.15 – Importance of individual receptor protons. For a given set of parameters (τ_c = 17, 250, or 1,000 ns, and either two or five ligands), a synthetic INPHARMA reference spectrum was calculated using the full relaxation matrix approach, and τ_L = 0.1 ns, ω = 800 MHz, k_{ij} = 1,000 s^{-1} for all ligand pairs $i \neq j$, ligand concentrations of 300 μM for all ligands, and a protein concentration of 30 μM, at a single mixing time of t_m= 300 ms. INPHARMA cross peak intensities were normalised to the diagonal peak intensity of ligand $i < j$ at a mixing time of 150 ms. In all cases, pairwise combinations of all ligands were used to calculate theoretical spectra. Unambiguous assigment of all ligand protons was assumed. Then, using the exact same set of parameters, test spectra were calculated for complexes con-

taining a *reduced* receptor PKA_{N-1}, omitting each receptor proton i out of the 339 receptor protons one at a time. Pearson correlation coefficients R_i were calculated for each test versus the reference spectrum, and the proton i omitted from the reduced receptor was assigned an importance score $I_i = 1 - R_i^2$. In the figure, panels show scatter plots between each of the six systems (τ_c = 17, 250, 1,000 ns) × (2, 5 ligands), with each circle representing one individual (omitted) proton / reduced receptor, and circles below the diagonal coloured grey for clarity. Because of the logarithmic scale employed, before plotting importance values 10^{-5} was added. The order parameter S^2 was set to 1 for all proton pairs.

It was found that for short mixing times (τ_c = 17 ns), about three or four receptor protons were sufficient to reproduce the reference spectrum with a Pearson correlation coefficient of R = 0.8 to 0.9, therefore passing a threshold value previously used to discriminate true from false relative ligand orientation. Clearly, such a low number of relevant receptor proton positions would not allow for a useful estimation of binding pocket shape. Increasing the rotational correlation time τ_c artificially to 250 or 1,000 ns, however, lead to about 30 to 40 receptor protons to be necessary for achieving a correlation coefficient of 0.8 to 0.9, surpassing the objectively useful number of > 20 receptor protons. This is in line with the importance estimation of the individual receptor protons described above (Figure 2.15).

It should be noted that throughout this analysis (Figures 2.15 and 2.16), a uniform order parameter of $S^2 = 1$ was used for all pairs of protons, i.e. a rigid model was employed omitting internal motion. Throughout the remainder of this chapter, however, the influence of order parameters $S^2 < 1$ on the INPHARMA calculations has been described in great detail, and it has been concluded that they make spin diffusion less efficient, therefore making the protein appear smaller, or effectively decreasing τ_c. Therefore, treating internal motion more realistically in a real life scenario would consequently lead to a reduction in the number of protons relevant for the magnetisation relay, thereby aggravating the problem described above. Indeed, repeating the calculations in Figures 2.15 and 2.16 using a uniform order parameter of $S^2 = 0.6 < 1.0$ showed that even for τ_c = 250 ns, fewer than 20 protons were sufficient to reproduce reference spectra with a correlation coefficient > 0.9, and that the correlation time needed for binding pocket shape reconstruction might be closer to 1,000 ns (data not shown).

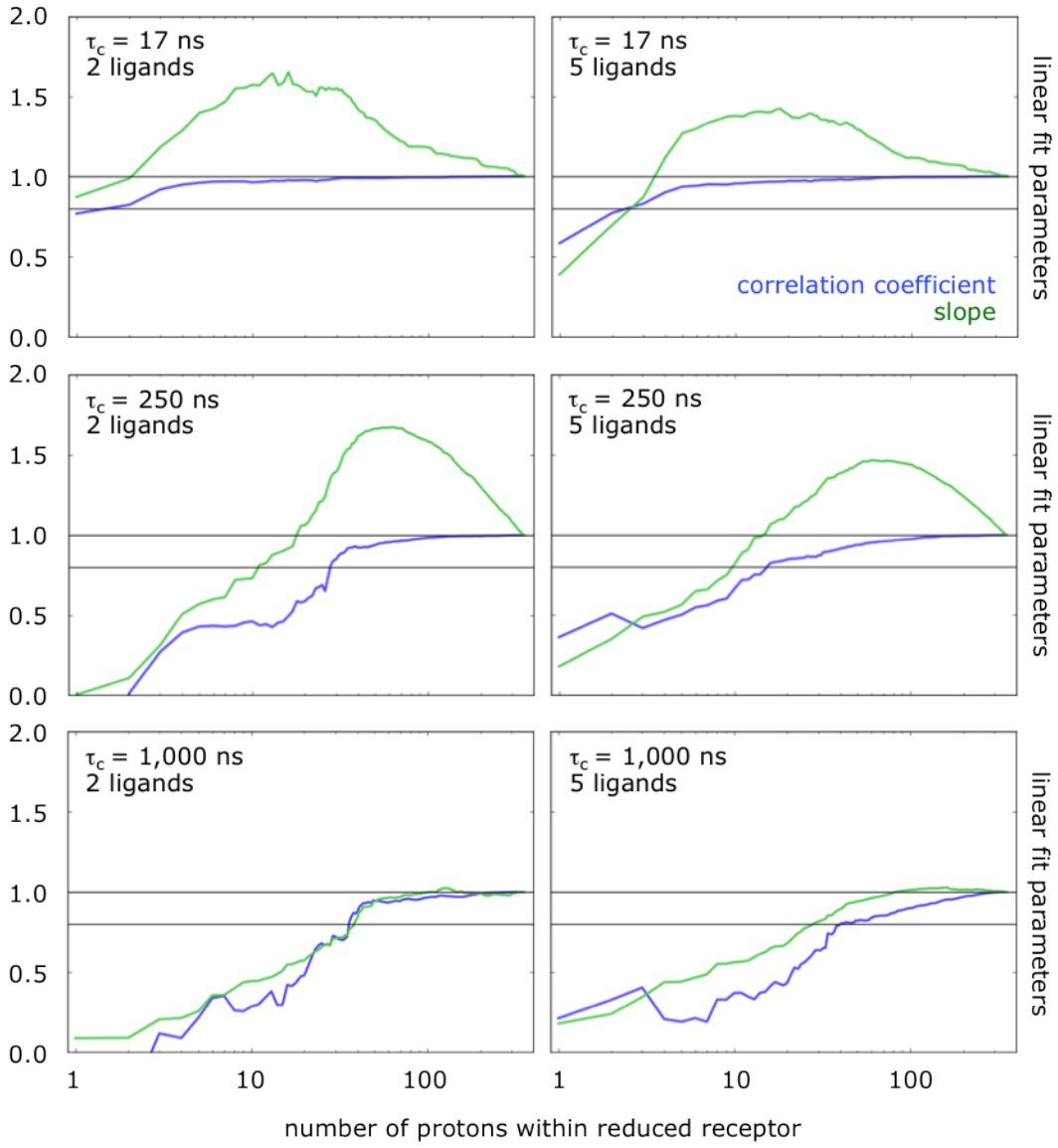


Figure 2.16 – Reproduction of reference spectra from reduced receptors. After ranking receptor protons for their importance with respect to reproduction of reference data (see main text and Figure 2.15), reduced receptors PKA_i were constructed, containing only the i most important protons ($i = 1..339$). For each of these reduced receptors, test spectra were calculated and compared to synthetic reference spectra using the same parameters (see below), but the *complete* receptor PKA_{339}, consisting of all 339 PKA protons within the 10 Å distance cutoff around the ligand. In all cases, pairwise combinations of all ligands were used to calculate theoretical spectra. Linear regression analysis was performed between test (y_i) versus reference (x) spectra, and Pearson correlation coefficients R (blue) and slopes a (green) of the

best fit line $y = ax$ plotted for every set of adjustable parameters. Horizontal black lines at $y = 0.8$ and 1.0 indicate a possible INPHARMA discrimination threshold, or a perfect fit of test and reference data, respectively. The number of protons i within the reduced receptor PKA_i is shown on a logarithmic scale. Adjustable parameters were τ_c and the number of ligands as indicated in the individual panels, and fixed parameters were $\tau_L = 0.1$ ns, $\omega = 800$ MHz, $k_{ij} = 1{,}000$ s^{-1} for all ligand pairs $i \neq j$, ligand concentrations of 300 μM for all ligands, and a protein concentration of 30 μM. A full buildup of mixing times t_m= 300, 450, 600, and 750 ms was used. INPHARMA cross peak intensities were normalised to the diagonal peak intensity of ligand $i < j$ at a mixing time of 150 ms. Unambiguous assigment of all ligand protons was assumed. An order parameter of $S^2 = 1$ was used for all proton pairs.

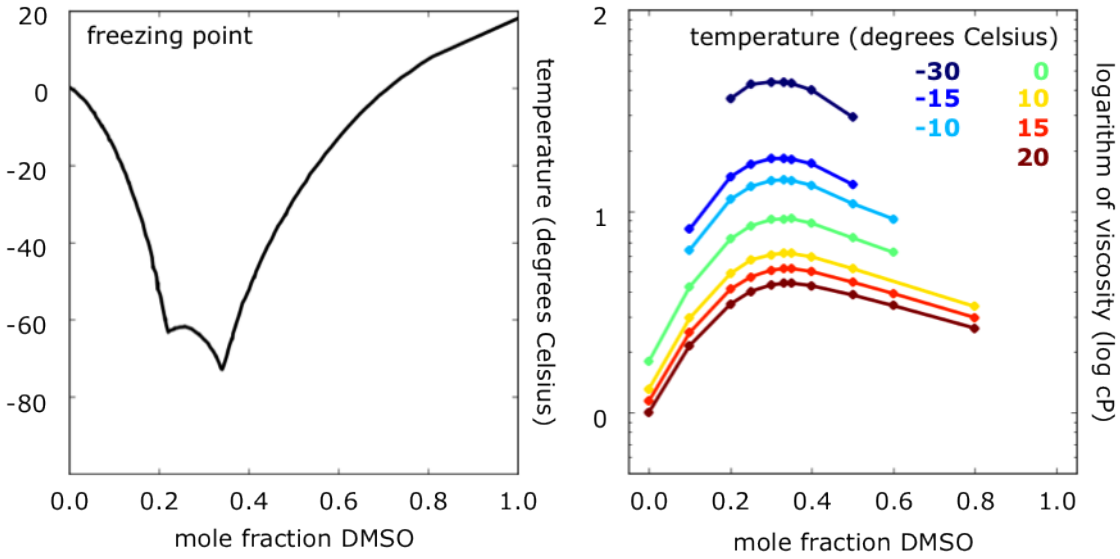


Figure 2.17 – Freezing point (left) and viscosity (right) of DMSO water mixtures. Freezing point data were digitised from a publication (Rasmussen and MacKenzie, 1968) using free graph digitisation software available online (`http://arohatgi.info/WebPlotDigitizer/app/`) by manually picking 142 data points, including experimental and interpolated data. Viscosity data for different temperatures (−30 degrees Celsius to 20 degrees Celsius, dark blue to dark red) were taken from the literature (Schichman and Amey, 1971). Viscosity is quoted in units of centipoise (cP), equivalent to mPa·s.

Therefore, it was concluded that it would desirable to increase the correlation time of the system of interest to $\tau_c > 250$ ns for the approach to be practicable. If the most simplistic hydrodynamic model of the protein is used, representing the

protein complex as a rigid sphere of fixed radius, the rotational correlation time τ_c can be estimated as (Cavanagh et al., 2007)

$$\tau_{c,\text{sphere}} = \frac{4\pi\eta r^3}{3kT}$$

with η the solvent viscosity, r the effective hydrodynamic radius of the protein complex (see below), and k and T the Boltzmann constant and the absolute temperature, respectively. Assuming an average protein density of $\rho = 1.37$ g/cm^3, the hydrodynamic radius can be estimated from the molecular weight M of the complex, $r \approx \sqrt[3]{\frac{3M}{4\pi\rho N_a}} + r_w$ with N_a Avogadro's number, and r_w the hydration radius (1.6 to 3.2 Å) (Cavanagh et al., 2007).

Therefore, τ_c can be regarded as a function of three adjustable parameters, i.e. solvent viscosity, molecular weight, and temperature, which is directly proportional to the former two, and anti-proportional to the latter. It should be noted that usually, viscosity itself decays exponentially as the temperature increases.[24]

Thus, the correlation time τ_c of a protein complex can be increased in different ways, i.e. choosing a solvent with higher viscosity, increasing the molecular weight of the system e.g. by fusing it to another protein, or even an artificial bead, and lastly, decreasing the experimental temperature. Fortunately, the alternative approaches can be used in conjunction which each other, and exhibit a favourable synergetic effect: besides the inverse correlation between sample temperature and viscosity, æqueous buffers are limited by the freezing point of the buffer, i.e. the temperature has to be high enough for the sample not to freeze. Some buffer components, such as dimethyl sulphoxide (DMSO), however, act as cryo protectants, lowering the freezing point of the mixture (Figure 2.17).

At certain molar ratios of DMSO over water (i.e. DMSO:water 1:2 to 3), the mixture exhibits a remarkable freezing point depression (Rasmussen and MacKenzie, 1968), staying in liquid phase even around −60 degrees Celsius. The reduction in temperature alone would increase the correlation time τ_c by a factor of about 1.2 to 1.4 (for temperatures of −20 or −60 degrees as compared to room temperature), but more importantly, the viscosity of the solvent mixture increases markedly by

[24] see (Kestin et al., 1978) for an investigation of the viscosity of liquid water in a temperature range up to 150 degrees Celsius.

a factor of about 44 (DMSO:water 1:2 at a temperature of −30 degrees Celsius, compared to water at room temperature) (Figure 2.17). This effect can be attributed to the formation of DMSO water complexes such as DMSO·2H_2O and DMSO·3H_2O (Kirchner and Reiher, 2002), and can be estimated from the phase diagram of the DMSO water mixtures (Figure 2.17) indicating the presence of stable complexes.

The combined effect of adding a molar fraction of 1/3 DMSO and lowering the sample temperature to about 0 degrees Celsius already results in an increase of the correlation time τ_c by a factor of approximately 10. Additionally, by fusing the protein of interest to a larger protein construct, which could also benefit the purification process, would cumulate in another factor, critically approaching the correlation time of $\tau_c = 250$ to 1,000 ns that was shown to be desirable in the above analysis (Figure 2.16), even starting from a medium sized protein such as PKA.[25] Sample temperatures of −10 to −20 degrees Celsius are easily achievable in NMR experiments using standard equipment.

Taken together, by tweaking experimental parameters correlation times of $\tau_c >$ 100 ns seem within reach, thereby possibly allowing for discrimination between different receptor proton arrangements. It could be shown that the performance of the method is crucially dependent on the correlation time τ_c.

Starting from the super-imposed ligand orientations, and thus inverting the regular INPHARMA work flow, receptor proton arrangements could be generated in a Monte Carlo optimisation like approach, testing them for their agreement with reference data,[26] and improving them in a stepwise fashion until they hopefully converge to the true solution, that is the actual receptor proton arrangement.

Taking the approach multiple steps further, in a true prospective application of this extension of INPHARMA, experimental spectra of one or several ligand combinations could be measured, and relative ligand orientation, as well as receptor proton arrangement, could be optimised simultaneously.

Initial work on synthetic systems, consisting of a receptor with circular and

[25] having a molecular weight of about 41 kDa, and a correlation time of about 17 ns (Orts et al., 2008b).

[26] experimental reference data, or synthetic reference data using the true receptor proton arrangement.

spherical proton arrangements[27] and simplified ligands, were promising, but pointed to a very rough energy surface with local cliffs, making difficult local search strategies (data not shown). It has to be noted, however, that spin diffusion was not incorporated in these initial attempts, but rather magnetisation transfer was treated as a two-step process, i.e. from the first ligand to the protein, and then back from the protein to the second ligand. This is close to experimental conditions if perdeuterated and selectively protonated proteins are used as receptors, since in this setup spin diffusion is effectively eliminated. This approach has already been successfully used to simplify INPHARMA calculations for the PKA/L_A and PKA/L_B system (Orts et al., 2008a), and was initially pursued also for the extension of the INPHARMA approach discussed here and next, applied to the synthetic model systems. However, in light of the experiments on proton importance described here, spin diffusion might actually be beneficial for the estimation of receptor proton arrangements,[28] and therefore, future work will include a rigorous treatment of spin diffusion using the full relaxation matrix approach.

2.4.6 Structure calculation using INPHARMA data

Traditionally, the INPHARMA workflow is as follows: 1. for each ligand, create a set of possible binding poses to the receptor by an external procedure, e.g. automated molecular docking, or manual placement of the ligand in the binding pocket. 2. combine pairs of binding poses for the different ligands to generate a pool of possible, pairwise solutions. 3. for a given pair of solutions, calculate the theoretical indirect magnetisation transfer between the two ligands using the full relaxation matrix approach. 4. rank the solutions according to the agreement of the calculated with the experimental INPHARMA spectrum, that has previously been recorded.

As it can be seen, the success of the method is greatly dependent on the performance of the external procedure producing the complex structures. Specifically, the true ligand binding poses have to be contained within the pool of possible

[27] thereby limiting proton arrangements to one or two dimensions, namely, circle diameter or sphere surface, as evident from coordinate conversion into a spherical coordinate system.

[28] keeping in mind the importance of *effective* spin diffusion as well, i.e. increasing the apparent receptor size by using viscous solvents at low temperatures, and potentially fusion proteins.

solutions. In this chapter, this was guaranteed by the fact that the true, crystal structures of the complexes were among the solutions, together with decoy ligand orientations. From this pool of solutions, the correct one could readily be identified.

It would, however, be desirable to obtain means to directly use INPHARMA data to generate the correct solution in the first place, instead of relying on an external procedure, especially in a real life prospective application, where the potential success of a docking procedure used to generate solutions is challenging to anticipate. This is in line with NMR based structure calculation, where experimentally derived NOE data are used as distance constraints together with a molecular mechanics force field, restricting bond lengths, angles, and dihedral angles to values close to known reference values. In the case of INPHARMA, however, incorporation of experimental restraints into the structure calculation is less straight-forward, since the transfer of magnetisation between ligands is an indirect process. Thus, quantities have to be defined correctly reflecting this indirect process. Programs used for NMR structure calculation such as Xplor-NIH (Schwieters et al., 2003) offer the possibility of defining custom energy functions that can be minimised along with the force field derived constraints described above. They require the definition of an *energy function*, and its derivative to yield force terms. These can then be used in an energy minimisation method such as the *conjugate gradient* method, or molecular mechanics approaches such as MD simulations. In the case of INPHARMA, however, the situation is complicated by the need to account for spin diffusion that can be incorporated rigorously using the full relaxation matrix approach but is not readily differentiable, although an analytical expression of the derivative is known (Yip and Case, 1989). However, in light of recent findings (Orts et al., 2008a), a more simplistic approach was preferred: In case of a perdeuterated, selectively methyl protonated receptor, spin diffusion is greatly attenuated, and magnetisation transfer between ligands is essentially a two step process (see above) (Orts et al., 2008a). In this case, the magnitude of the INPHARMA signal is roughly proportional to the inter-nuclear distance of the ligand protons, $\mathrm{trNOE}_{ij} \propto d_{ij}$. This greatly simplifies the definition of energy and force terms needed for the minimisation process.

Following the approach just outlined, a pseudo energy function was defined,

penalising the deviation of a calculated (calc) from a reference INPHARMA spectrum (obs) for two ligands and pairs of protons i, j,

$$E = \frac{1}{2} K_0 \sum_{ij} \left(\mathrm{NOE}_{\mathrm{calc},ij} - \mathrm{NOE}_{\mathrm{obs},ij} \right)^2 \tag{2.10}$$

with K_0 an adjustable constant. Neglecting spin diffusion in a perdeuterated protein background, and assuming uniform methyl protonation, treating inter-ligand magnetisation transfer as a two step process, for a given pair of complex structures a theoretical INPHARMA spectrum can be calculated as

$$\mathrm{NOE}_{\mathrm{calc},ij} = K \sum_{m} d_{im}^{-k} \cdot d_{mj}^{-k}$$

with K constant, and m receptor methyl groups.[29] For the term to recapitulate an NOE based magnetisaton transfer, the exponent $-k$ is set to -6, but other (even) exponents are also possible. The inter-nuclear distance d_{im} is calculated from the (x, y, z) coordinates as $d_{im} = \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2 + (z_i - z_m)^2}$. The force $\vec{F}_i$ on a ligand nucleus i is now defined as the negative gradient of the energy function, its derivative in all directions,

$$\vec{F}_i = -\nabla E_i = \left(-\frac{\partial E}{\partial x_i}, -\frac{\partial E}{\partial y_i}, -\frac{\partial E}{\partial z_i} \right)$$

For the energy function E defined above (Equation 2.10), this was solved using a mathematical software toolkit capable of performing symbolic operations.[30] For the x component of the force, the following expression was obtained:

$$F_i(x) = -kKK_0 \sum_{j} \Delta\mathrm{NOE}_{ij} \sum_{m} \Delta x_{im} d_{im}^{-(k+2)} d_{mj}^{-k} \tag{2.11}$$

with $\Delta\mathrm{NOE}_{ij} = \mathrm{NOE}_{\mathrm{calc},ij} - \mathrm{NOE}_{\mathrm{obs},ij}$ the difference of cross peak intensities, and $\Delta x_{im} = x_i - x_m$ the relative coordinate difference.

[29] instead of using three methyl proton positions, for simplicity, a dummy atom m can be introduced at the average position of the three methyl protons, or coinciding with the methyl carbon.

[30] a free 2003 version of the MuPAD software (University of Paderborn, Germany) was used before the program was commercialised.

As it becomes evident from the functional form of Equation 2.11, the biggest contribution to the force calculations stems from the inter-nuclear distances raised to a high exponent. If $k = 6$, the distance is approximately raised to an exponent of 14. This results in an extremely rough energy surface, where small coordinate changes result in very large forces. This is especially relevant in the context of balancing the force term with other force terms imposed by the molecular mechanics force field (see above), such as those keeping bond lengths and angles close to equilibrium parameters. If during the minimisation process a balance is not maintained, forces exerted by the INPHARMA potential can virtually destroy the connectivity of the ligand molecule, if they act in opposing directions to e.g. the bond length potential.

Using the *PyPot* interface in Xplor-NIH (Schwieters et al., 2003), providing the possibility to define custom potential energy terms implemented in the Python scripting language, an INPHARMA potential was implemented based on Equations 2.10 and 2.11. As standard, Xplor only supports the treatment of single coordinate sets, i.e. the *CONStraints INTEraction* statements of Xplor had to be used to prevent the two complexes under consideration (e.g. PKA/L_A and PKA/L_B) from interacting with one another. By using these statements, the complexes can be constrained to only interact internally, i.e. atoms within the first complex with each other, and atoms in the second complex with each other, but no direct interaction across the complexes is possible. The only interaction *between* complexes is mediated by Equations 2.10 and 2.11, thereby recapitulating the design of the NMR experiment.

In initial attempts, the newly implemented potential was successfully used to guide synthetic ligands consisting of a single proton to their position completing a geometric arrangement resembling a cube, with protons of the synthetic receptor constituting 7 out of the 8 cube vertices, and the one ligand proton the eighth. In a similar synthetic test system consisting of a receptor made up of two cubes of protons joined by a common surface, and the ligands, of two 'protons' joined by a cube edge, the correct structure calculation could be repeated (data not shown). However, using the approach on PKA/L_A and PKA/L_B as a test system, including force field terms constraining bond lengths, angles, and dihedral angles

to equilibrium positions in order to prevent the newly defined INPHARMA based potential from deforming the ligand, it proved remarkably challenging to define a successful minimisation protocol along with a set of constants to mutually balance the different potentials. This is reminiscent of the high exponents discussed above, leading to very large values of the force term in Equation 2.11 with respect to the molecular force field derived forces. As per design the INPHARMA potential only exerts a force on ligand protons, its force is often directed into a different direction as the other forces also acting on the ligands non-hydrogen atoms, thereby deforming the ligand. This issue can potentially be addressed by applying an additional force to the non-hydrogen atoms of the ligand in the same direction as the net force from the INPHARMA derived potential on the ligand protons, or alternatively, considerably scaling down the INPHARMA derived forces. Also, a non-physical, but lower exponent such as $k = 2$ instead of $k = 6$ might be used in Equations 2.10 and 2.11, making the energy surface smoother.

2.5 Methods

2.5.1 Molecular Dynamics simulation

Proteins and protein-ligand complexes were simulated using NAMD 2.6 (Phillips et al., 2005) and the CHARMM22 force field (MacKerell et al., 1998) in a periodic cubic box of explicit TIP3P water (Jorgensen et al., 1983) and counter ions for charge neutralisation, with a box side length of the maximum inter-nuclear distance of the respective protein atoms (~ 45 Å for 1ubq,[31] ~ 50 Å for 1shf,[32] ~ 47 Å for 1ten,[33] and ~ 45 Å for 1lib[34]), plus a padding of 25 Å to avoid mutual interaction of the protein images. Crystal structure coordinates, with hydrogen atoms added using REDUCE (Word et al., 1999), were used as starting points of the simulations. Total atom counts were 32,260 (1,231 protein atoms plus 10,343 water molecules) for 1ubq, 39,586 (twice 917 protein atoms plus 12,584 water molecules) for the homo dimer of 1shf, 35,087 (1,397 protein atoms plus

[31] human ubiquitin.
[32] human FYN tyrosine kinase SH3 domain.
[33] fibronetin type III domain from human tenascin.
[34] murine adipocyte lipid binding protein.

11,230 water molecules) for 1ten, and 32,018 (2,057 protein atoms plus 9,987 water molecules) for 1lib. After 10^4 discrete steps of initial energy minimisation, systems were heated step wise from 0 to 298 K with a temperature increment of 3 K per 1 ps, followed by an equilibration phase of 5 ns. Both minimisation and equilibration were carried out in an NPT ensemble with a Langevin thermostat and constant pressure control to allow for volume adjustment, using a damping constant of 5 ps^{-1}, hydrogens uncoupled from the bath, and a piston target pressure of 1 atm. The time step integrator was set to 2 fs, requiring bond lengths to be fixed throughout the course of the simulation using the SHAKE algorithm (van Gunsteren and Berendsen, 1977).[35] The equilibration phase was followed by 25 ns of production run in an NVT ensemble, taking the latest unit cell dimensions from the equilibration phase. Production run unit cell dimensions were (68.16 Å)^3 for 1ubq, (72.95 Å)^3 for the 1shf dimer, (70.07 Å)^3 for 1ten, and (67.95 Å)^3 for 1lib. Force field parameters for the ligands of Protein Kinase A (PKA) (equilibrium bond lengths, angles, dihedral angles, and non-bonded interactions) were assigned analogous to known compounds following previous description (Vanommeslaeghe et al., 2010). Analogous substructures were extracted from thiazole, tiophene, histidine, indole, and aminopyridine moieties, and corresponding parameters assigned to unknown ligand parameters. In case of MD simulations of PKA, the PKI inhibitor peptide fragment present in the structures with identifiers 3dnd and 3dne was excluded from the analysis, but phosphate groups at positions Thr^{197} and Ser^{338} were retained, in order to recapitulate the (NMR-)experimental setup.

For all systems, after MD simulation, each frame of the trajectory was superimposed to the crystal structure conformation for reference, minimising protein non-hydrogen atom RMSD, utilising VMD (Humphrey et al., 1996). Water coordinates were deleted and one snapshot per 1 ps was subjected to further analysis. Trajectories were routinely subjected to convergence analysis based on the development of the values of force field derived energy terms.

[35] it has been shown previously that the application of SHAKE has a negligible effect on simulated order parameters (Pfeiffer et al., 2001).

2.5.2 Order parameter extraction

Distance-dependent NMR order parameters S^2 (Lipari and Szabo, 1982a,b) were extracted directly from the MD trajectories utilising a custom tcl script in VMD (Humphrey et al., 1996) according to Equation 2.5 (Brüschweiler et al., 1992). Purely radial and purely angular order parameters were calculated as in Equations 2.8 and 2.9 (Brüschweiler et al., 1992). As for the implementation of the angular order parameter calculation, an alternative functional form was preferred, following an approach outlined previously (Bremi et al., 1997): expression of the spherical harmonics in terms of Cartesian coordinates allows for the order parameter to be expressed as

$$S^2 = 1 - {}^3\!/_4 \left(3\sigma^2_{z^2} + \sigma^2_{x^2-y^2} + 4(\sigma_{xy})^2 + \sigma^2_{yz} + \sigma^2_{xz}\right) \tag{2.12}$$

where $x(t)$, $y(t)$, and $z(t)$ are the Cartesian components of the inter-nuclear unit vector $\mathbf{e}_i$ at MD snapshot i and $\sigma^2_f = \langle f^2 \rangle - \langle f \rangle^2$ is the variance of the real function $f(\mathbf{e}_i)$ (Bremi et al., 1997). It is straightforward to show that Equation 2.12 is invariant with respect to permutation of x, y and z. The benefit of this functional form lies in its easy implementation and efficiency as it allows for computation of S^2 in a single loop over the frames of the trajectory.

2.5.3 Decomposition of generic order parameters

Distance-dependent order parameters for all pairs of non-exchanging protons with mutual distance less than 5 Å were extractd from the trajectories of MD simulations of four globular proteins as described in Methods section 2.5.2. For the order parameter of proton pair (k, l) S^2_{kl}, a matrix $\mathbf{A}$ with $A_{kl} = S^2_{kl}$ was constructed. By applying a conjugate gradient method as implemented in the *fsolve* routine[36] of MATLAB® (2007a, The MathWorks, Natick, MA, USA), a vector $\mathbf{x}$ was determined minimising the l^2-norm $\left\|\mathbf{A} - \mathbf{x}^\top \mathbf{x}\right\|$. Ths vector holds the S-factor S_k for all nuclei k, approximating $S^2_{kl} \approx S_k \cdot S_l$ for all pairs of nuclei (k, l) belonging to the same residue ('intra') or different residues ('inter'). Protons were classified accord-

[36] with default parameters and a maximum of 2,500 function evaluations. Convergence as indicated by the exit flag was achieved in every case.

ing to the carbon atom they are attached to: C-α, CH_3, CH_2-β, CH_2-γ, CH_2-δ, CH_2-ε, CH_2-proline, CH_1, and aromatic protons. For each group, the S-factors of the proton were averaged over all pairs in the group and over the four globular proteins to yield the generic S-factor for this group. To restore generic order parameters from the S-factors, protons were assigned generic S-factors according to their chemical connectivity, and two S-factors were multiplied to retrieve the generic order parameter.

2.5.4 INPHARMA calculations

As described previously (Orts et al., 2008b, 2009), INPHARMA NOEs between the two exchanging ligands for Protein Kinase A (PKA) are computed employing the full relaxation matrix approach (Nilges et al., 1991; Kalk and Berendsen, 1976; Keepers and James, 1984; London, 1999) to account for all possible pathways of spin diffusion, thus allowing for rigorous, quantitative treatment of the NOE transfer. The differential equation

$$\frac{d\mathbf{M}}{dt} = -(\mathbf{R} + \mathbf{K}) \cdot \mathbf{M}(t) \tag{2.13}$$

is solved for a given NOESY mixing time τ_M, yielding $\mathbf{M}(t)$, the magnetisation matrix at time t, as

$$\mathbf{M}(\tau_M) = \exp\left(-(\mathbf{R} + \mathbf{K})\tau_M\right) \cdot \mathbf{M}(0) \tag{2.14}$$

The kinetic matrix $\mathbf{K}$ represents the chemical exchange according to the kinetic model $\mathrm{T} \cdot \mathrm{L_A} \longleftrightarrow \mathrm{T} \cdot \mathrm{L_B}$ with T being the target protein and $\mathrm{L_A}$ and $\mathrm{L_B}$ the respective ligands. This model assumes the target protein is never found in the unbound state due to the ligands being present in large excess. The relaxation matrix $\mathbf{R}$ contains the auto relaxation rates $R_{kk} = \rho_k$ and cross relaxation rates $R_{kl} = \sigma_{kl}$ for all nuclei (k, l). The spectral density function used to estimate these rates has the form of the first term of Equation 2.4 as described in the Theory section. A more thorough theoretical treatment of INPHARMA can be found elsewhere (Orts et al., 2009).

For this chapter, the INPHARMA approach was reimplemented using the

Python scripting language. Software versions used in earlier studies (Orts et al., 2008b) represented less efficient MATLAB® implementations of the approach that did not allow for modification of the NMR order parameter. Equation 2.14 was solved using the matrix exponential routine[37] of the SciPY library in Python 2.4.3, for mixing times τ_M of 150, 300, 450, 600, and 750 ms, unless stated otherwise. Adjustable parameters were set to ω = 800 MHz proton resonance frequency; τ_c = 17 ns correlation time for the protein; τ_L = 0.1 ns correlation time of the free ligands; k_{AB} = 3,000 s^{-1} and k_{BA} = 1,000 s^{-1} exchange rates of the respective ligands according to the kinetic model; L_A = 450 μM and L_B = 150 μM respective ligand concentrations; and 25 μM (for the NOESY experiments with 450 and 750 ms mixing time) or 30 μM (else) protein concentration, to recapitulate the experimental setup. INPHARMA NOEs were normalised to the intensities of the diagonal peaks of L_A in a NOESY spectrum at 150 ms mixing time. Normalised INPHARMA NOEs for mixing times 300 to 750 ms were compared to normalised experimental intensities obtained at the same mixing times, and conditions, and a simple linear regression was performed to yield the Pearson correlation coefficient and the slope a of the regression line $y = ax$.

To limit computational expenses, all calculations were performed on a 'reduced' receptor consisting of the 339 protons of PKA within a maximum distance of 10 Å around the ligands. At realistic mixing times τ_M and for medium sized proteins, as measured by their rotational correlation time τ_c, receptor protons in greater distance can be assumed not to contribute significantly to the spin-diffusion mediated magnetisation transfer. In applications of the INPHARMA method to cases where a very high number of possible ligand conformations and their combinations needs to be discriminated amongst, an iterative approach can be used to filter pairs of conformations. This is advisable as INPHARMA relies on an expensive matrix operation, namely the computation of the matrix exponential (see above) performed on rather large matrices, the size of which scales with twice the number of receptor protons incorporated into the model. It is conceivable that this matrix operation has a computational complexity not significantly smaller than that of a standard matrix multiplication, which is in $O(N^3)$, with N the number of rows and columns. In this iterative approach, working in multiple

[37] see Methods section 2.5.5.

cycles of ranking pairs of conformations, in early cycles, only the receptor protons closest to the ligands can be taken into account, e.g. those within a distance of 6 Å, the experimental detection limit of a standard NOE. Excluding a high proportion of the ligand pairs giving the worst fit to experimental data, most of the high number of false ligand conformation pairs can be excluded in an efficient manner, as bad ligand conformations will score bad even using comparatively coarse receptor models. In subsequent steps, more detailed calculations can be carried out, incorporating receptor protons within 7 Å. Again, excluding the worst solutions according to their fit with experimental data, fewer combinations remain to be back-calculated in even greater detail, reaching 8, 9, and eventually 10 Å as a distance cut-off. This cut-off should suffice for medium sized proteins such as PKA.

2.5.5 Computation of the matrix exponential

The matrix exponential $e^{\mathbf{A}}$ of a square $(n \times n)$ matrix $\mathbf{A}$ can be computed in analogy with the power series definition of the exponential, $e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + ..$, as

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{6}\mathbf{A}^3 + ..$$

If $\mathbf{A}$ is decomposed into $\mathbf{A} = \mathbf{PDP}^{-1}$ for some matrices $\mathbf{P}$, $\mathbf{D}$, the matrix exponential can be calculated as $e^{\mathbf{A}} = \mathbf{P}e^{\mathbf{D}}\mathbf{P}^{-1}$ since

$$\begin{aligned} e^{\mathbf{A}} &= \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + .. \\ &= \mathbf{I} + \mathbf{PDP}^{-1} + \frac{1}{2}\mathbf{PDP}^{-1}\mathbf{PDP}^{-1} + .. \\ &= \mathbf{I} + \mathbf{PDP}^{-1} + \frac{1}{2}\mathbf{PD}^2\mathbf{P}^{-1} + .. \\ &= \mathbf{P}\left(\mathbf{I} + \mathbf{D} + \frac{1}{2}\mathbf{D}^2 + ..\right)\mathbf{P}^{-1} \\ &= \mathbf{P}e^{\mathbf{D}}\mathbf{P}^{-1} \end{aligned}$$

using $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ and $\mathbf{DI} = \mathbf{D}$.

A good choice for $\mathbf{D}$ is a diagonal matrix since the exponential of a diagonal matrix $\mathbf{D}$ with elements $D_{ii} = d_i$ is a diagonal matrix $\mathbf{E}$ with elements $E_{ii} = e^{d_i}$ and therefore easy to compute (Moler and Van Loan, 2003). In general, real symmetric matrices (such as those used in the INPHARMA calculation, see Methods section 2.5.4) are diagonalisable by orthogonal matrices, that is, matrices whose row and column vectors are orthogonal unit vectors. This diagonalisation approach is implemented in the *linalg.expm2* routine of SciPY.

2.5.6 Correlation function convergence

The existence of an order parameter S^2_{kl} requires the internal correlation function $C_{kl}(t)$ to converge to a plateau value. After removing overall tumbling motion by superimposing each snapshot of the MD simulation trajectory on the initial structure as reference (Methods section 2.5.1), normalised internal correlation functions were computed directly from the MD simulation trajectory as (Schneider et al., 1999)

$$C_{kl}(t_i) = c_1(N-i)^{-1} \sum_{j=1}^{N-i} \frac{P_2\left(\theta^{\text{lab}}_{kl}(t_j) \cdot \theta^{\text{lab}}_{kl}(t_{i+j})\right)}{r^3_{kl}(t_j) \cdot r^3_{kl}(t_{i+j})} \tag{2.15}$$

with $C_{kl}(t)$ containing both angular and radial fluctuations and $P_2(x) = 1/2(3x^2 - 1)$ the second Legendre polynomial, $\theta^{\text{lab}}_{kl}(t_i)$ / $r_{kl}(t_i)$ the inter-nuclear unit vector orientation/distance at discrete time step i, respectively, N the finite length of the MD simulation trajectory, and $c_1^{-1} = (N-i)^{-1} \sum_{j=1}^{N-i} r^{-6}_{kl}(t_j)$. It should be noted that the choice of the normalisation constant c_1 is arbitrary and was set to correspond to the definition of the order parameter S^2 (Equation 2.5) for convenience. For fast internal motions, C decays rapidly to a plateau value S^2 with a characteristic internal correlation time. As MD trajectories are of finite length, estimation of $C(t)$ is not precise for $t_i \to t_N$ since $N - i \to 0$, only few MD simulation snapshots contribute to the average in Equation 2.15, and ergodicity can no longer be assumed due to this sampling problem.

For $C(t)$ to converge, it was required to stay within a certain range of the plateau value S^2, without large fluctuations, for an extended period of t. To this end, an error function $\varepsilon(t) = |C(t) - S^2|$ was defined, and the longest interval $[t_i, t_j]$ with $i < j$ was estimated such that the mean $\tilde{\varepsilon}_{ij} = \langle\varepsilon\rangle_{ij}$ in this time interval, and

the corresponding standard deviation σ_{ij} do not exceed 0.05, as well as requiring $\tilde{\varepsilon}_{i'j'} \leq 0.05$ and $\sigma_{i'j'} \leq 0.05$ for all sub intervals $[i', j'] \in [i, j]^2$ with $i, j = 1..N$ such that $i' < j'$. If $2(j - i) < N$, i.e. if C is close to S^2 consecutively for at least half of its domain of definition, C is considered to have converged.

Values of $\tilde{\varepsilon}$ and σ can be computed efficiently using *dynamic programming* and on-the-fly computation of means and standard deviations in a single pass, online algorithm (see Methods section 2.5.7). However, since $1/2N(N+1)$ values of $\tilde{\varepsilon}$ need to be computed and N is in the range of 2.5×10^4 discrete MD simulation time steps, and $\geq 10^4$ internal correlation functions need to be examined, it was decided to divide $[1, N]$ into 10^2 non-overlapping stretches of equal size, to compute average values of ε on these intervals, and use the 10^2 averages instead of the 2.5×10^4 original values for further analysis. This has the additional advantage of smoothing the data, without changing the characteristic course of a particular correlation function.

2.5.7 Single pass computation of sample variance

The variance (and mean) of a sample $\mathbf{x}$ of length N can be computed in a single pass using the following recursive formulæ:

$$\begin{aligned} \bar{x}_i &= \bar{x}_{i-1} + \frac{x_i - \bar{x}_{i-1}}{i} \\ M_{2,i} &= M_{2,i-1} + (x_i - \bar{x}_{i-1})(x_i - \bar{x}_{i-1}) \\ \sigma_N^2 &= \frac{M_{2,i}}{N} \end{aligned}$$

with the auxiliary value $M_{2,i}$, $i = 1..N$, $\bar{x}_1 = x_1$ and $M_{2,1} = 0$, and $\bar{x}_N$ and σ_N^2 the sample mean and variance, respectively. The algorithm is efficient and numerically stable (Knuth, 1997; Welford, 1962).

2.5.8 Linear regression

Pearson correlation coefficients between samples $\mathbf{x}$ and $\mathbf{y}$ and slopes of the best fit line $y = ax$ were calculated as $R = \mathrm{cov}(x, y)/(\sigma_x \sigma_y)$ and $a = \sum_i x_i y_i / \sum_i x_i^2$, respectively, with σ the sample standard deviation.

2.6 Summary

In this chapter, modifications to the NMR based INPHARMA method for the determination of the relative orientation of two competitive ligands binding to the same protein receptor have been discussed. By incorporating a rigorous representation of protein internal motions, using the order parameter concept, into the underlying model representing the physics of the magnetisation transfer on which the method is based, a more realistic theoretical estimation of signals measured experimentally could be achieved. Importantly, a previously observed over-estimation of the magnitude of magnetisation transfer could be attributed to the rigid model ignoring protein internal motions used before, and thereby accounted for. By incorporating NMR order parameters S^2 derived from MD simulations, for the PKA test systems and two ligands, an improvement in calculating theoretical INPHARMA spectra in better agreement with the experiment, both quantitatively and qualitatively, could be achieved.

Next, the novel concept of order parameter decomposition was developed and introduced to provide a formalism representing motion as a *proton wise* quantity, as opposed to the proton *pair* wise quantity that is the NMR order parameter S^2. These values, called S-factors, were obtained by mathematically decomposing order parameters from globular proteins into individual proton contributions, and can later be re-combined by multiplication to recover order parameters S^2. The decomposition was found to retain almost all information present in the original order parameters, and allows for rapid application of the order parameter concept to INPHARMA calculations of unknown systems. This was taken advantage of in discriminating true ligand orientations in the PKA test system from decoys, eliminating false positive solutions previously observed that had hampered the performance and discrimination power of the method. Performing an INPHARMA back-calculation on a modelled, open conformation of the PKA-ligand complexes, it was found that a minor population of the open enzyme conformation, which is likely to be present in solution according to previous experimental work described in the literature, can account for some of the residual over-estimation of magnetisation transfer efficiency that remained when using INPHARMA in conjunction with NMR order parameters.

Furthermore, preliminary results for possible extensions of the INPHARMA methodology were presented. To this end, it was investigated how much information about the receptor proton arrangement is present and can possibly be deduced from the INPHARMA inter-ligand spectra. It was found that for small receptors with low rotational correlation time τ_c, only very few receptor protons participate in the the magnetisation relay, and therefore, any method aiming at the reconstruction of the shape of the ligand binding pocket of the receptor based on INPHARMA quantities is likely set to fail. It was found, however, that for very large receptors with $\tau_c \approx 1{,}000$ ns, a sufficiently high number of receptor protons contributes significantly to the magnetisation transfer, and can therefore, possibly, be probed by a Monte Carlo approach. Strategies of increasing the *apparent* receptor size to feasible regions have been discussed.

Lastly, an analytical expression for an INPHARMA based pseudo energy function, along with its derivative yielding force terms, have been presented, fit for the incorporation into structure calculation work flows, thereby potentially by-passing the dependence on external means to generate complex structures to be scored by the INPHARMA method, and instead using the INPHARMA concept *directly* to obtain complex structures. Initial attempts have highlighted the importance of the challenging task of balancing this novel force term with existing molecular mechanics force field based terms constraining bond lengths, angles, and dihedral angles to equilibrium parameters.

As a concluding remark, it should be noted that most of the results obtained in this chapter come from the analysis of a single protein system, namely PKA and its two ligands L_A and L_B. Much care was taken to eliminate the bias of general statements derived from this single system. However, future work should explore additional systems, and ideally repeat some of the analyses presented here for those. Although purely computational analyses are easy to conduct for novel systems even in the absence of experimental information, it is exactly the experimental data that ultimately have to serve for validation of all models. Unfortunately, experimental INPHARMA data is difficult to obtain, as the sensitivity problem fazing NMR spectroscopy is aggravated by an additional order of magnitude when using INPHARMA, since all magnetisation transfer between ligands is indirect. However, alternative systems to PKA have recently been explored us-

ing INPHARMA,[38] and hopefully in the future, the method will be more widely adopted into the routine of the NMR spectroscopy based investigation of protein ligand interactions, perhaps aided by the insights provided in this chapter.

[38] for reference, see (Sanchez-Pedregal et al., 2006; Bartoschek et al., 2010; Erdelyi et al., 2010; Kubicek et al., 2010; Krimm, 2012).

Chapter 3

Characterisation of Xenon Binding Sites of Proteins

3.1 Introduction

In this chapter, a computational knowledge-based pseudo-potential of xenon interacting with proteins will be devised. The noble gas xenon is now widely used as a standard heavy atom for *ab initio* phase determination in X-ray crystallography. In this section, the *phase problem* in X-ray crystallography will be illustrated, together with means to solve it, following the outline and arguments presented in a standard crystallography textbook (Rupp, 2009). The term 'phase problem' refers to the fact that in crystallography, the amplitudes of beams diffracted by a crystal can be measured, but not their absolute phases. Phase information, however, is crucial for the reconstruction of an electron density map, and ultimately, an atomic model, the determination of which is the most important aim of crystallographic studies. The attractiveness of xenon as a heavy atom for phasing is discussed, and results in a plethora of structural knowledge about xenon-protein interaction to be present in the Protein Data Bank. This structural information is made use of for deriving the predictive method. Known to interact with preformed hydrophobic cavities of proteins, the empirical notion that xenon is often found in, and is therefore potentially useful to identify, ligand binding sites of proteins[1]

[1] based on largely anecdotal evidence (Prange et al., 1998).

is corroborated systematically, and a knowledge-based methodology is devised to predict xenon binding sites in proteins by computational means. In addition to characterising xenon binding sites of proteins statistically, the goal of the predictive method is the computational detection of protein cavities that are able to bind drug-like small molecule ligands. This would render xenon even more attractive for usage in crystallographic studies, providing it with an additional utility as a possible novel standard fragment in *fragment based drug discovery* (FBDD, see Introduction section 1.1).

3.1.1 The Phase Problem in Crystallography

Crystals are regular arrangements of structural motifs such as atoms, ions, or molecules, their regular macroscopic form already indicative of the regular microscopic arrangement of their constituents. The basic structure of a crystal can be formally described by the *space lattice*, a pattern of abstract points representing the locations of these repeating structural motifs, thereby defining the basic structural layout of the crystal. Based on the lattice, the *unit cell* can be defined as a parallel-sided figure representing the repeating three-dimensional pattern from which the entire crystal can be reconstructed by using only translations.[2] It is the primary objective of crystallography to determine the contents of the unit cell at atomic resolution.

This can be achieved by exploiting the periodic organisation of the crystal that leads to measurable scattering of X-ray beams by the crystal. In order for radiation to interact with objects, the wavelength of that radiation has to be comparable to characteristic spatial properties of these objects. For biomolecules, the spatial scale of the atomic bond length (about 1.5 Å) coincides with the X-ray region of the electromagnetic spectrum (0.01 to 10 nm) and is therefore far outside the region of visible light (about 400 to 700 nm). Biomolecules thus evade direct observation by optical means, but in analogy to the phenomenon that larger objects scatter visible light that can be focused by a lens and projected onto a screen[3] for detection, in the absence of physical means available to focus scattered

[2] for a given lattice, different unit cells can be defined.

[3] this can be an artificial lens made of glass or plastic, or the lens of the human eye focusing scattered rays of light onto the retina.

X-ray beams by a physical lens, mathematical procedures can be used to retrieve information contained in patterns of scattered X-rays. It turns out that the *Fourier transform* is the necessary tool for achieving this, a formalism apt to link real space and the reciprocal space that contains the experimental observables in X-ray crystallography. In the following section, the process by which crystals diffract X-rays will be reviewed in a concise fashion, giving rise to signals that can be measured experimentally. It will be shown that a crucial component of information cannot be detected, namely the phase of the scattered beams. This is known as the *phase problem* in crystallography, and ways to overcome it will be sketched. An emphasis will be put on using heavy atom, anomalous scatterers, such as xenon, which is the primary interest of this chapter.

X-ray scattering by individual molecules does not give rise to measurable signals, the signals being far too weak to detect. However, the periodic arrangement of features within a crystal amplifies signals. X-ray beams can be represented by waves of defined amplitude, wavelength and absolute phase. When many rays interact with equivalent 'features' of the crystal, constructive interference of these waves can occur, and signals can be detected experimentally. Constructive interference enhances the amplitude of the ray and gives rise to sharp, discrete peaks of intensities, termed reflections, that can be collected on a detector device. It turns out that for constructive interference to occur, stringent requirements have to be fulfilled which are formalised by *Bragg's law* (Equation 3.1). If these requirements are not fulfilled, destructive interference occurs, and no signal can be detected. The periodic 'features' mentioned above that X-ray beams interact with turn out to be sets of equivalent, imaginary two-dimensional planes inside the three dimensional arrangement of space lattice points within the crystal grid, called *lattice planes*.

Bragg's law states that constructive interference occurs if

$$\lambda = 2d \sin \theta \tag{3.1}$$

with λ the wavelength of the incident radiation, d the distance between the two planes under consideration, and θ the glancing angle of the radiation.[4]

[4]the glacing angle being the angle between the beam and the surface, as opposed to the angle

When two (parallel) incident rays of the same wavelength λ are reflected by two adjacent planes of the lattice with inter-planar distance d, they acquire a path length distance that depends on the incident angle θ, and d. If that path length difference is an integer multiple of the wavelength λ, the reflected rays remain in phase and interfere constructively. For angles not satisfying the Bragg condition, rays emerge out of phase, and destructive interference prevents detectable signals. The reflection of rays by a single set of planes gives rise to a single reflection that can be measured experimentally. Furthermore, from measuring the reflection angle θ, the plane spacing can be determined. As d and $\sin\theta$ are inversely proportional, it follows that high-resolution features of a crystal[5] have large reflection angles.

Individual planes are referred to by discrete integer indices h, k, l, the *Miller indices*, representing their separation measured in reciprocal fractions of unit cell lengths. It should be noted that the usage of Miller indices represents a transition from real, continuous space of inter-molecular or -atomic distances to the reciprocal, discrete space of reflections.

It can be shown that the overall amplitude of a wave diffracted by plane hkl is

$$F_{hkl} = \sum_j f_j e^{i\Phi_{hkl}(j)} \tag{3.2}$$

summing over all atoms j within the unit cell, with f_j the atomic *scattering factor*,[6] i the imaginary unit, and $\Phi_{hlk}(j) = 2\pi(hx_j + ky_j + lz_j)$ the phase difference between the hlk reflections of two atoms.[7] The atomic scattering factor f_j directly depends on the number of electrons of an element. Therefore, heavy atoms diffract more strongly than light atoms (see below). F_{hlk} is called the structure factor, and is related to the intensity I_{hlk} that can be measured at the detector by $I_{hlk} \propto |F_{hlk}|^2$. An individual structure factor, and therefore a single reflection, is thus a Fourier sum of wave contributions from diffraction by every single atom in the crystal.

Instead of summing over all atoms (Equation 3.2), the electron density ρ can

between the beam and the *plane normal* that is quoted more customarily ('angle of incidence').

[5] those with small inter-planar distance d.

[6] a measure how strongly a given element scatters a ray.

[7] with atom coordinates defined respective to an origin as fractions $x_j = \frac{X_j}{|a|}$, $y_j = \frac{Y_j}{|b|}$, and $z_j = \frac{Z_j}{|c|}$ for three base vectors $\vec{a}$, $\vec{b}$, $\vec{c}$.

be integrated over unit cell volume V to obtain structure factors:

$$F_{hkl} = \int_{xyz} \rho(x, y, z) e^{2\pi i(hx+ky+lz)}\, dx\, dy\, dz \tag{3.3}$$

As the Fourier transform is reversible, it provides a means of deriving the electron density if the structure factors are known:

$$\rho(x, y, z) = V^{-1} \sum_{hlk} F_{hkl} e^{-2\pi i(hx+ky+lz)} \tag{3.4}$$

(note the sign change the exponent in Equation 3.4 with respect to Equation 3.3, signifying the *inverse* Fourier transform). It is the electron density ρ that is the objective to be determined by X-ray crystallography, rendering Equation 3.3 especially important.

It should be noted that the absolute phases of the waves are implicit in the formulation of Equation 3.3. A more directly useful, explicit formulation is given by Equation 3.5 below. As the structure factor F is a vector quantity in the plane of complex numbers, uniquely determined by amplitude $|F|$ *and* phase angle α as $F = |F| \cdot e^{i\alpha}$, it follows

$$\rho(x, y, z) = V^{-1} \sum_{hlk} |F_{hkl}|\, e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})} \tag{3.5}$$

for some phase angle $\alpha = 2\pi\alpha'$.

In this equation, $I_{hlk} \propto |F_{hlk}|^2$ can be measured by the detector, and hkl are available from indexing the respective reflection.[8] The only unknown that cannot be determined in a straightforward manner through experiment is the phase α of each reflection. The ambiguity of α is known as the *phase problem*, and it can be shown that the information contained in the phases is equally, if not more important than that in the amplitudes of the reflections. In the following, common strategies to experimentally access phase information are described.

[8] that is, their position on the detector screen as a function of the ϕ angles that describe the orientation of the crystal within the X-ray beam.

3.1.2 Molecular replacement, isomorphous replacement, and anomalous scattering

There are several ways of obtaining crystallographic phase information. First, it should be noted that Equation 3.2 provides a means of calculating phases from the atomic positions of a molecule. Therefore, in a process called *structure refinement*, iteratively, a structural model with atomic resolution can be modified and improved, manually or automatically, within an electron density map, and phases can be calculated from this model by the application of Equation 3.2. In the next step, a new, improved electron density can be calculated from these phases, driving the refinement process towards an optimum defined by the agreement between calculated and observed structure factors. In other words, structure refinement alternates between using equations 3.3 and 3.4, interspersed with manual or automatic adjustment of the atomic model. The definition of Equations 3.2 and 3.3 allows electron density and atomic positions to be used interchangeably, and the incorporation of additional restraints into the refinement process by employing protein force fields facilitates the stepwise improvement of phases and atomic models. It follows that it is sufficient to determine *initial phases* that are sufficiently good to obtain an interpretable initial electron density, into which an initial atomic model can be built. As this model will implement the additional constraints in the form of knowledge about the molecule of interest, e.g. the amino acid sequence of a protein under investigation, and equilibrium atomic bond lengths and -angles from a protein force field, initial phases provide a starting point for the structure determination process. Notably, from Equation 3.2 it can be appreciated that the placement of individual atoms, including water molecules and ions, has an effect on all structure factors, and therefore, all phases, and consequently, the entire, back-calculated electron density. Given the high information content of phases (see above), it is evident that wrongful refinement can lead to a potentially problematic *model bias*, especially when using a molecular replacement strategy (see below), especially in cases where subtle differences between protein conformations, e.g. in the context of ligand binding, are investigated. This has to be accounted for during refinement cycles by a process called *cross validation*.

A straightforward source of initial phases is obtained by employing a *molecular*

replacement strategy. In this approach, in the context of protein crystallography, the atomic model of a closely related protein can be used to obtain initial phases, together with the new intensities (amplitudes) recorded for the protein under investigation. In Chapter 4, a molecular replacement strategy has been used for the phasing of Hsp90-NTD. Interestingly, recent progress in *ab initio* protein structure prediction, paired with a crowd-sourcing approach, has resulted in a remarkable success story: a model of a protein of unknown structure that was created manually by players in a multiplayer online computer game was successfully used for obtaining initial phases, that could then be used in the independent, experimental structure determination of that protein (Khatib et al., 2011; Cooper et al., 2010).

Alternatively, certain properties of heavy atoms can be used to obtain initial phases. In a process called *isomorphous replacement*, heavy atoms are soaked into crystal to obtain *heavy atom derivatives* of that crystal. The term 'isomorphous' refers to the fact that the heavy atom derivatisation must not interfere with protein structure, and importantly, crystal parameters, as diffraction data from the derivative crystals has to be compared to data collected from native crystals. As stated in Equation 3.2, every single atom in a crystal contributes to every structure factor. In the isomorphous replacement strategy, the addition of one or few atoms to specific sites that are uniform across all unit cell instances causes a slight perturbation of the diffraction pattern. For the perturbation to be large enough to measure, the atom must diffract X-ray beams strongly (see factor f_j in Equation 3.2). Therefore, atoms with high atomic number such as heavy metals are used to derivatise protein crystals. A common alternative is the substitution of cystein and/or methionine residues by their selenocystein and selenomethionine bioisosteres by site-directed mutagenesis.[9] The isomorphous replacement strategy commences by comparing crystallographic data sets collected from native and derivative crystals. By using the *difference* diffraction pattern, one hopes to be able to determine the positions of the heavy atoms alone within the unit cell. If this is successful, phases can be calculated from the heavy atom positions alone, using equation 3.2. It should be noted that although individual atoms contribute to *all* structure factors, they contribute to some of them more strongly than to others. Therefore, phase estimation based on heavy atom positions provides accu-

[9] selenium has an atomic number of 34 that is higher than that of sulphur (16).

rate phases for some reflections, but inaccurate phases for others. Often, however, the set of phases is of sufficient quality to be used as initial phases for atomic refinement.

Heavy atoms can be located within the unit cell by using the *Patterson synthesis* instead of the Fourier synthesis (Equation 3.5). The Patterson synthesis results in a Patterson map instead of an electron density map ρ. The Patterson function is defined as $P(u,v,w) = V^{-1} \sum_{hkl} |F_{hkl}|^2 e^{-2\pi i(hu+kv+lw)}$. Comparing this equation to Equation 3.5 it becomes evident that (a) the intensity $I_{hkl} \propto |F_{hkl}|^2$ is used directly as coefficients in the summation, and that (b) it does not contain any phase terms. However, instead of an electron density ρ on real space coordinates (x,y,z) it results in a Patterson map that is defined on *coordinate differences* (u,v,w).[10] This prevents Patterson maps from being interpreted intuitively. However, using a trial and error strategy, heavy atom positions can be derived, given that a sufficiently small number of them are located within the unit cell. From these heavy atom positions, phases for the difference diffraction pattern can be computed. Together with the *amplitudes* from the diffractions collected from the native and derivative data sets, it turns out that these phases can be used to reduce the ambiguity of the initial phases (the phases of the native data set) to two possible values, instead of the whole range of phases of 0 to 2π. By collecting a second data set from a different heavy atom derivative, this residual ambiguity can be resolved. However, both types of heavy atoms have to bind to different locations of the protein. Otherwise, the phase information they provide is redundant. This approach is termed MIR (multiple isomorphous replacement).

An additional property of heavy atom facilitates an alternative approach to obtain phases from a difference diffraction pattern. Heavy atoms have the capacity to absorb X-rays of specific range. As a result of this absorption, *Friedel's law* does not hold any more: reflections hkl and $\overline{hkl}$ are no longer equal in intensity, and the phase α_{hkl} is no longer equal to $-\alpha_{\overline{hkl}}$ as they are in absence of absorption. This property of absorption induced inequality of symmetry-related reflections is called *anomalous scattering*. While for most elements and wavelengths, the *absorption edge*, that is, the wavelength at which the absorption characteristics of an element

[10] for each pair of atoms within the unit cell, the Patterson map contains a peak at the coordinate difference of these atoms.

suddenly changes, cannot be reached, for heavy atoms, using tunable synchrotron radiation, the anomalous signal is appreciable even at remote energies. In analogy to isomorphous replacement, heavy anomalous scatterers can be located within the unit cell of a derivatised crystal by the Patterson method (see above), and phases can be computed using equation 3.2. Again, the phase is two-fold degenerate, but this time, the asymmetry between the two Friedel partner reflections can be used to resolve this residual ambiguity, with the help of a third data set, recorded from the derivatised crystal at a different wavelength chosen to maximise anomalous scattering. Again, selenomethionine labelling of the protein by site-directed mutagenesis can be used to introduce anomalous scattering. Together with the amplitudes of the native and derivative data sets, phases for the native reflections can be deduced.

3.1.3 Xenon as heavy atom anomalous scatterer

Using xenon as a heavy atom to derivatise protein crystals for phase determination by isomorphous replacement has become increasingly popular in recent years. An additional benefit of xenon is that for wavelengths used at a synchrotron anomalous scattering can be observed (see also Chapter 4), providing an independent experimental method of verifying xenon positions in protein structures (Panjikar and Tucker, 2002a). In classic experiments on sperm-whale myoglobin, xenon has been found to specifically interact with the protein (Schoenborn et al., 1965). Additionally, xenon possesses direct medical relevance, as it interacts with biological molecules and acts as a potent anæsthetic (Cullen and Gross, 1951).

Xenon is known to bind to specific sites in macromolecules (Tilton et al., 1984; Sauer et al., 1997). Crystals can be derivatised by pressurising native crystals with xenon gas within a pressure chamber. In general, modest xenon pressures (around 10 atm) are sufficient to achieve a good occupancy of xenon sites, and if multiple binding sites with differential affinity are present within a protein, the number of bound xenons can be influenced by altering the xenon pressure applied to the pressure chamber. Xenon is not toxic, and no further modifications of the protein system have to be carried out, e.g. replacing biochemical buffers, making xenon derivatisation cheap, fast and uncomplicated. Typically, xenon binding sites

differ from heavy metal binding sites, which is important for the MIR approach (see above). Xenon interacts weakly with proteins. Its association is mediated by inherently weak dispersive forces and does not involve electrostatic interactions (Sauer et al., 1997). As noble gases can be expected to bind in pre-formed hydrophobic cavities of the protein, there is a smaller likelihood of disrupting or altering crystal contacts, so the isomorphism between native and derivative crystals can typically be expected to be high (Panjikar and Tucker, 2002b,c). Xenon can safely be assumed to have a limited effect on pH or the ionic strength of the buffer (Sauer et al., 1997), furthermore contributing to the isomorphism of the derivatised crystal. For porcine elastase, binding is achieved within less than 20 minutes (Sauer et al., 1997) inside a pressure chamber. The retention time of xenon within a protein crystal is in the region of minutes (Sauer et al., 1997). For freeze-trapped crystals, no significant loss of protein-bound xenon was observed in a study carried out on sperm-whale myoglobin (Sauer et al., 1997). In non-frozen crystals, xenon occupancy was monitored at the same time as data were collected from myoglobin crystals mounted within a pressure cell. Xenon dissociation was found to happen on the timescale of minutes, and the half life of the off-rate was found to be in the order of 20 seconds (Soltis et al., 1997). Based on these advantages, xenon has been used for ab initio phasing exploiting both its anomalous signal or by isomorphous replacement, or combinations of the two complementary approaches (Vitali et al., 1991; Bourguet et al., 1995; Cianci et al., 2001; Panjikar and Tucker, 2002c; Olczak et al., 2003). However, xenon binding sites tend to be less numerous than binding sites of other heavy atoms (Panjikar and Tucker, 2002c), so it has been suggested that novel xenon binding sites could be introduced into proteins by mutagenesis (Quillin and Matthews, 2002); it was found that indeed xenon can bind to engineered cavities, obtained by large-to-small mutations within the protein core.

In this introductory section, the phase problem in crystallography, the concept of heavy atom anomalous scatterers, and the general relevance of xenon have been illustrated, all of which are important for the study conducted in this chapter. Additional concepts such as molecular replacement and crystallographic structure refinement have been introduced which are of particular interest for the next

chapter (Chapter 4). The aim of the remainder of this chapter is to formalise the empirical notion that xenon binds to preformed hydrophobic protein cavities, and that many of these cavities overlap with binding pockets for small drug-like organic molecules. By characterising xenon-binding sites structurally, a computational method is devised to predict xenon binding sites in a prospective manner, given a protein structure.

3.2 Results

In this chapter, a knowledge-based *xenon likeness score* is developed based on structural information available in the Protein Data Bank (PDB), aiming to discriminate positions in protein structures that can bind xenon atoms from those that do not. After describing the general properties of the data set of xenon binding proteins, their spatial coordinates are retrieved from the PDB, and the propensity of xenon to bind to ligand binding sites is investigated. Next, the environments of the xenon atoms are characterised by radial distribution functions (rdfs). Special care is taken to correct for a protein sequence bias present in the data set, and boundary effects arising at crystal unit cell interfaces and between multiple copies of the asymmetric unit. The rdfs are combined into a knowledge-based potential, or *xenon likeness score*. This score is benchmarked on the positions of xenon atoms against those of buffer ions/water molecules present in the data set, and an analysis is carried out to determine possible score cutoff values to be used to discriminate the former from the latter.

3.2.1 Xenon-binding proteins in the Protein Data Bank

The Protein Data Bank (PDB) (Berman et al., 2000) contains 99 protein structures solved by X-ray crystallography that bind at least one xenon atom.[11] In total, these protein structures consist of 177 individual protein chains, 160 of which bind xenon, representing 64 unique UniProt (UniProt Consortium, 2012) identifiers, and 470 individual xenon atoms. To characterise the xenon binding protein chains

[11] as of April 2012. See Supplementary Table S2 in the Appendix for details.

and general properties of the data set, protein sequences, molecular masses and experimental resolutions were retrieved from the PDB records.

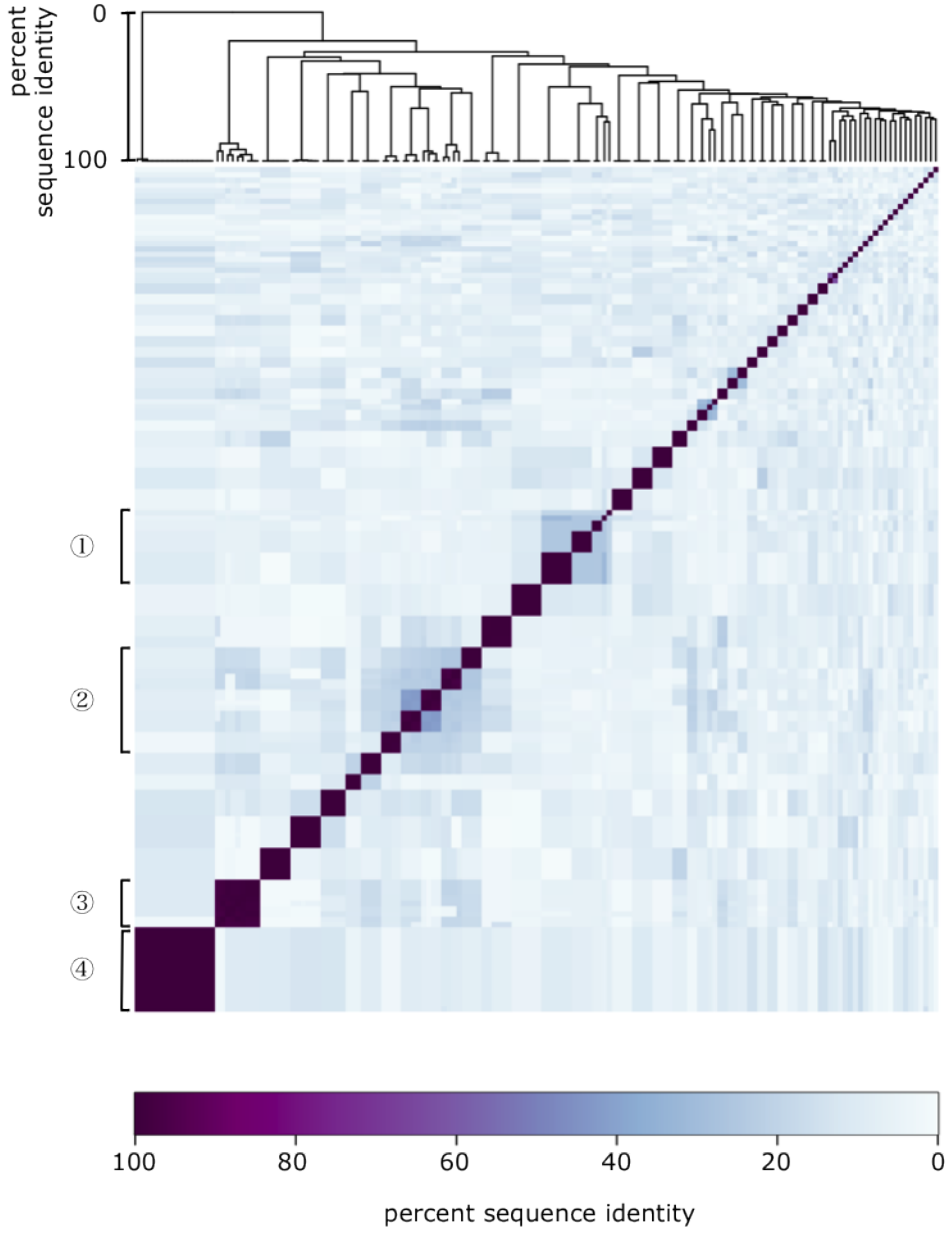


Figure 3.1 – Sequences of 160 protein chains binding xenon were aligned in a pairwise fashion to yield a 160 × 160 sequence identity matrix (shown colour-coded from purple to blue to white, see colour bar) using the global sequence alignment tool *needle* of the EMBOSS suite of programs (see Methods section 3.4.3). Select individual sequence clusters from hierarchical cluster analysis (dendrogram

shown on top; see Methods section 3.4.3) are indicated by circled numbers ①-④ with ① being amine oxidases (UniProt identifiers: P46883, Q43077, P12807), ② globin proteins: cytoglobin (Q8WWM9), myoglobin (P02185), hemoglobin α-subunit (P69905), hemoglobin β-subunit (P68871), globin-1 (P02214), from top to bottom; ③ lysozyme (P00720), and ④ uricase (Q00511).

When the sequences of the xenon-binding protein chains in the PDB were compared against each other, some redundancy was found to be present in the data set (Figure 3.1); for instance, there are 9 lysozyme chains (UniProt identifier P00720), as well as 16 uricase chains (Q00511), with $\geq$ 98 % mutual sequence identity. Overall, the data set contains 74 unique clusters at 100 % sequence identity,[12] 60 clusters at 70 % sequence identity, and 59, 58, and 54 clusters at sequence identity levels of 50, 40, and 30 % sequence identity, respectively (based on results from a hierarchical cluster analysis; see below). This sequence bias will be accounted for later in this chapter. It should be noted that the alignment-based sequence similarity analysis was conducted without the assumption of an evolutionary relationship between all proteins.

The majority of the experimental structures of the xenon-binding proteins present in the data set were found to be typical globular proteins with structures solved at medium to high resolution (median 1.9 Å), and have low to medium molecular mass (median 26.3 kDa) (Figure 3.2), with molecular masses reported for each individual protein chain in the asymmetric unit. It was further observed that a fraction of xenon atoms were located towards the limits of the asymmetric unit, thus potentially interacting with atoms in crystal images (Figure 3.3) and influencing the calculation of solvent accessible surface areas (Figure 3.4.4, Methods section 3.4.4) of individual xenon atoms. Since the knowledge-based potential to be developed throughout this chapter is based on the spatial distribution of atoms around xenon positions, and treatment of isolated unit cells would lead to problematic boundary effects, the crystal packing was reproduced by a unit cell expansion around the xenon atoms (Methods section 3.4.5) and the expanded structures were further used to derive the knowledge-based potential.

Further technical details can be found in Methods sections 3.4.1, 3.4.2 and 3.4.3. Next, it was investigated whether xenon atoms were frequently found in the

[12] clusters at i % sequence identity are considered homogeneous if each pair of sequences in that clusters shares at least i % sequence identity.

ligand binding sites of proteins.

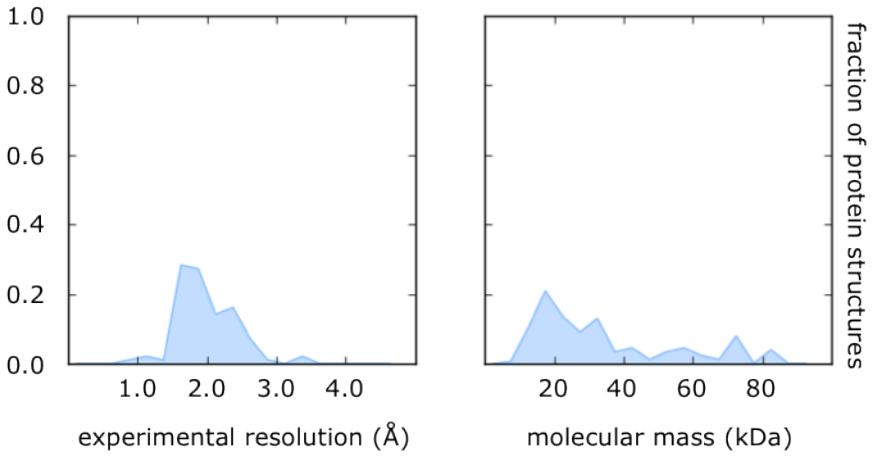


Figure 3.2 – Distribution of experimental resolution (left, in Å, N=99, bin width 0.25 Å) and molecular mass (right, in kDa, N=177, bin width 5 kDa, all values stated are bin centres) of crystal structures of xenon-binding proteins in the Protein Data Bank. Median values are 1.9 Å and 26.3 kDa, respectively.

3.2.2 Xenon atoms are found in ligand binding sites

To investigate the propensity of xenon atoms to bind to ligand binding sites of proteins, ligands that sterically overlap with the positions of xenon atoms were extracted from the Protein Data Bank (PDB).

To achieve this, first, for each of the 470 xenons in the data set, the UniProt identifier of the protein chains they were bound to was used to query the PDB in order to identify other instances of the same protein where an experimental structure is known (see Methods section 3.4.2). This could be achieved for 426 out of 470 xenon atoms. The remaining xenon atoms were either bound to synthetic protein chains that had no UniProt identifier associated to them or there was no other PDB record available with the same UniProt identifier that also contained a ligand.

Of the PDB entries identified, only those which were both solved by X-ray crystallography, and contained a ligand were retained. As it was aimed to investigate druggable protein binding sites, the analysis should be limited to drug-like

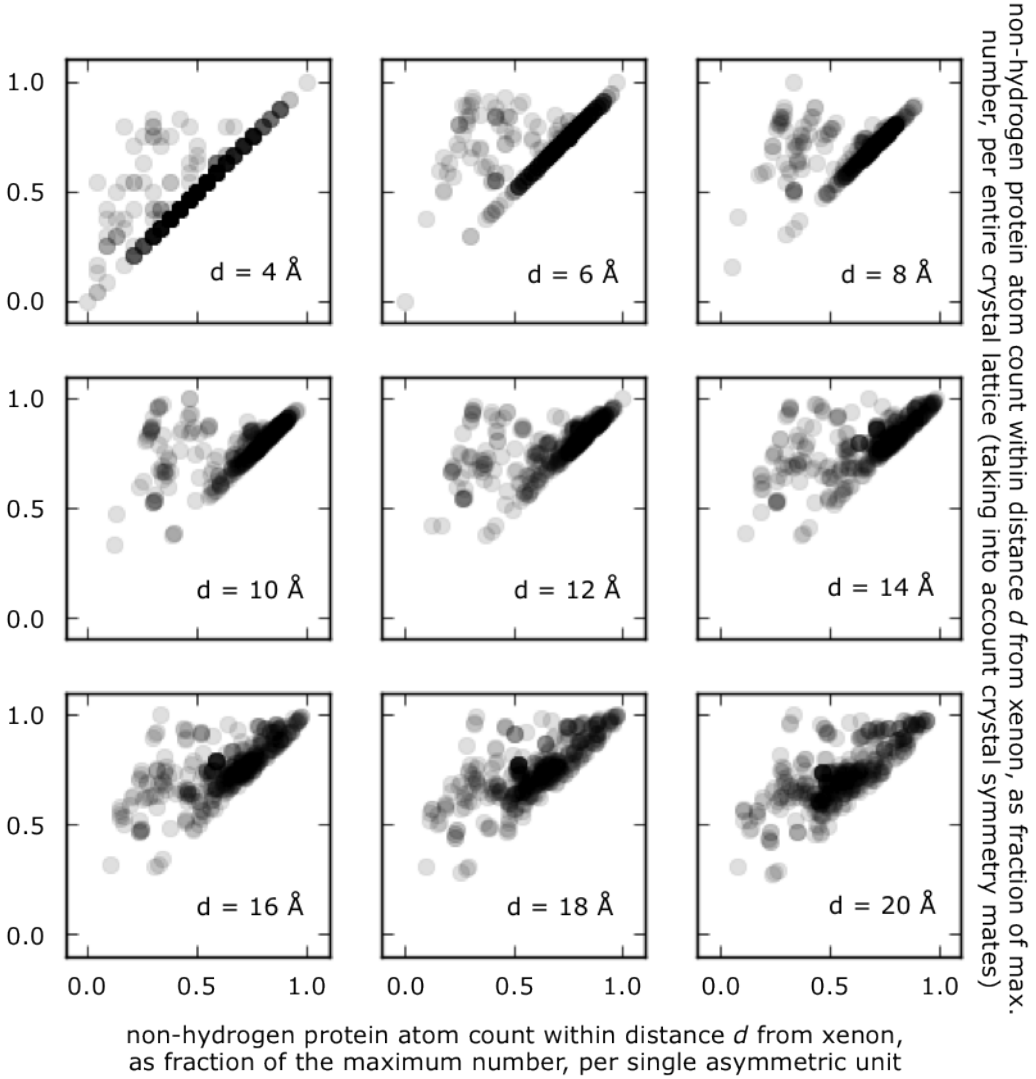


Figure 3.3 – Enhancement of the number of protein non-hydrogen atoms proximal to xenon by using crystal symmetry information. The number of protein atoms (C, N, O, S) around individual xenon atoms in the asymmetric unit (x-axes) is plotted against the number of protein atoms around the same xenon atom in the entire crystal lattice (y-axes, by taking into account crystal symmetry mates) within specific distances (bottom right corner of each panel). Values are represented as a fraction of the maximum number of atoms found within that distance around *any* xenon atom in the data set. Each black circle represents one xenon atom ($N = 470$), and circles above the diagonal indicates the presence of neighbouring protein atoms in another instance of the asymmetric unit.

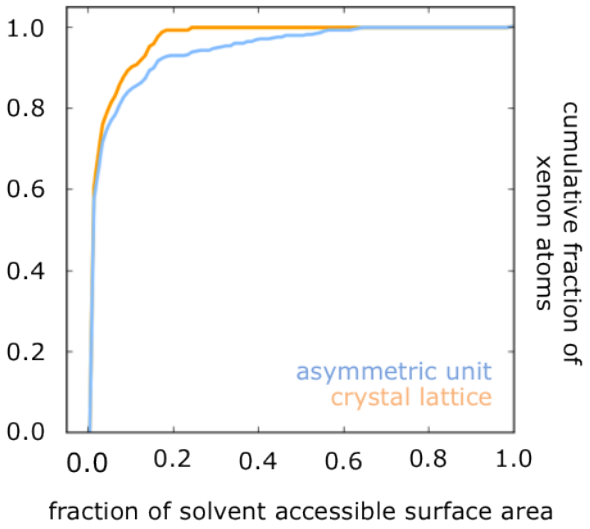


Figure 3.4 – Cumulative histogram of relative solvent accessible surface areas of xenon atoms in the Protein Data Bank, assuming a *van der Waals* radius of $r_{vdW,Xe}$ = 2.2 Å and a probe radius of r_{probe} = 1.4 Å. The analysis was performed before (light blue) and after (orange) expansion of the unit cell, i.e. taking into account symmetry related objects by exploiting crystal symmetry (see Methods section 3.4.4).

molecules. Therefore, to exclude solvent molecules and ions from the analysis, only ligands with ≥ 5 non-hydrogen atoms were considered. Furthermore, heme groups were excluded from the analysis as they were found to constitute roughly 31 % of the total number of ligands. Next, pairs of ligand-binding protein chains and xenon-binding protein chains were super-imposed, minimizing the *root mean square deviation* (RMSD) between the C^{α} atoms of the two protein chains (see Methods section 3.4.6 for details). Ligands were considered to overlap with xenon if at least one of their non-hydrogen atoms was found within ≤ 3 Å of the xenon atom in the super-imposed pair of structures. In turn, a xenon atom was considered to be in a ligand binding site if it was found to overlap with at least one ligand.

By following this approach, a total of 120 out of 426 xenon atoms (28.2 %) were found to be located in ligand binding sites, covering 22 of the 64 proteins represented by their UniProt identifier.[13] For a further 30 xenon atoms ligands

[13] 8 of 64 UniProt identifiers did not have any other PDB record associated to them.

were found in a distance between 3 Å and 4 Å. To coarsely characterise xenon binding from the perspective of the respective ligand with which it was found to potentially overlap, SYBYL atom types (Clark et al., 1989) were assigned to the atom closest to the xenon position in each ligand (see Methods section 3.4.7 for details). The majority of these atoms were either sp^3 hybridized or aromatic carbons, contributing 36.4 % and 27.4 % of all closest atoms, respectively. Halogens made up an additional 11.6 % (chlorine: 3.7 %, fluorine: 3.5 %, bromine: 2.5 %, iodine: 1.9 %). Moreover, sp^2 hybridized carbons and sp^3 hybridized oxygen atoms constituted an extra 6.8 % and 5.2 %, respectively. Trigonal planar nitrogen atoms contributed 3.0 %, while *sp* hybridized carbon atoms were found in 2.4 % of cases. All other atom types, except sp^2 hybridized oxygen (2.4 %), selenium (1.7 %) and sp^2 hybridized nitrogen (1.3 %), contributed less than 1 % each (see Table 3.1 for a summary).

Table 3.1 – Frequency distribution of SYBYL ligand atom types for atoms overlapping with xenon. Only the closest ligand atom to the xenon position was considered. Values are given as percentage of the total number of closest atoms over 1996 ligands.

C.3	sp^3 carbon	36.4 %
C.ar	aromatic carbon	27.4 %
C.2	sp^2 carbon	6.8 %
O.3	sp^3 oxygen	5.2 %
Cl	chlorine	3.7 %
F	fluorine	3.5 %
N.pl3	trigonal planar nitrogen	3.0 %
Br	bromine	2.5 %
C.1	*sp* carbon	2.4 %
O.2	sp^2 oxygen	2.4 %
I	iodine	1.9 %
Se	selenium	1.7 %
N.2	sp^2 nitrogen	1.3 %
N.ar	aromatic nitrogen	0.9 %
S.3	sp^3 sulphur	0.9 %
N.3	sp^3 nitrogen	0.2 %
B	boron	0.2 %
N.4	sp^3 positively charged nitrogen	0.1 %
N.am	amide nitrogen	0.1 %

3.2.3 A knowledge-based potential for xenon-protein interaction

To describe the binding of xenon to proteins, a knowledge-based potential was derived from the *radial distribution function* (rdf) of xenon and different *hetero*-atoms of the protein. The rdf describes the environment of a given particle in a distance-dependent manner by counting the number of instances a specific atom type is found in a given distance from the reference particle, and normalising that number to the *expected* number of atoms in that distance bin, assuming a *perfect gas* background distribution. Thus, the rdf describes the enrichment of that atom type at that distance over the background. The rdf can then be converted into a *pseudo* potential of energy in a straightforward manner, assuming that this distribution has the same characteristics as a Boltzmann distribution. In this work, an individual rdf (and later, an individual pseudo energy function) is derived for each *type* of protein atom (see below).
The rdf for atom type *het* is given by

$$g_{\text{het}}(r) = \frac{\text{count}_{\text{het}}(d_r)}{\rho V_{\text{shell}}} \tag{3.6}$$

with $count_{\text{het}}(d_r)$ representing the number of atoms of type *het* found in distance bin d_r between $r + 1/2\delta r$ and $r - 1/2\delta r$ of the xenon atom, $\rho = N_{\text{het}}/V_{\text{all}}$ the *number density* with N_{het} the number of atoms of type *het* found in the entire spherical volume $V_{all} = 4/3 \cdot \pi r_0^3$ under consideration, and $V_{\text{shell}} = 4/3\pi\left((r + 1/2\delta r)^3 - (r - 1/2\delta r)^3\right)$ the volume of the spherical shell that is the distance bin. In this work the environment of xenon atoms was investigated up to a distance r_0=20 Å, using a bin width of δr=0.25 Å. It should be noted that $g(r)$ is dimensionless. The distance of 20 Å is comparable to the *radius of gyration* (R_{gyr}) of medium sized proteins such as ubiquitin (R_{gyr}=11.7 Å, based on the NMR structure with the PDB identifier 1d3z (Cornilescu et al., 1998)), or corresponding to the radius of a (spherical) protein with a molecular weight of about 27.2 to 28.9 kDa (assuming an average protein density of 1.43 to 1.35 g/cm^3)[14]. It follows that for proteins in that size

[14] average protein density values taken from (Fischer et al., 2004), and references therein, and using 1 g = $6.02213665168 \cdot 10^{23}$ kDa.

range, no single unit cell dimension will be smaller than 20 Å and thus, after *unit cell expansion* (see above), each xenon atom will have an entire sphere of protein atoms of > 20 Å around it, thereby avoiding periodic artifacts.

Radial distribution functions for each of the 470 xenon atoms with all non-hydrogen protein atoms were calculated, with protein atoms being assigned to and represented by one of 25 CHARMM22 atom types (MacKerell et al., 1998), yielding 470×25 preliminary rdfs.[15] CHARMM atom types were used because they were expected to represent protein atoms more accurately than the more general SYBYL atom types used for the ligand analysis above.

Preliminary rdfs were then averaged over the xenon atoms to yield the final 25 atom type wise rdfs. Since the initial sequence analysis had indicated redundancy in the data set (Figure 3.1), the averaging of the preliminary rdfs was done in a stepwise manner while performing a *hierarchical clustering* (see Methods section 3.4.3) of the amino acid sequences of xenon-binding protein chains (for a dendrogram see the top of Figure 3.1). Whenever two sequence clusters were merged by the clustering method, and the distance between those two clusters passed a certain threshold value (see below) for the first time, the preliminary rdfs assigned to the respective sequences were averaged to yield an averaged rdf now representing the newly formed sequence cluster. This procedure was repeated until all clusters were merged, and for distance thresholds corresponding to sequence identities of 100 %, 95 %, 90 %, 70 %, 50 %, 40 %, 30 %, and 0 %. The sequence identity I (in percent) and the distance D are related by $D = (100\% - I)/1\%$, and mutual cluster distances were calculated with a *complete linkage* approach as the maximum distance of any two sequences in non-identical clusters.[16]

In other words, firstly rdfs of xenon atoms bound to protein chains sharing 100 % mutual sequence identity were averaged to yield a reduced set of rdfs, the resulting rdfs representing disjoint sets of xenon environments at 100 % sequence identity level. Next, some sets were merged by averaging the intermediate rdfs obtained at 95 % sequence identity level,[17] thereby further reducing the number

[15] an overview of atom types can be found in Methods section 3.4.8.
[16] which is equivalent to *single linkage* clustering on the sequence identities.
[17] those sharing ≥ 95 % sequence identity.

of rdfs. This procedure was then iterated at sequence identity levels of 90 %, 70 %, 50 %, 40 %, 30 %, and 0 %, with the ultimate averaging resulting in a single, averaged rdf for each protein atom type. These rdfs were then smoothed by calculating a weighted average over adjacent values of the radial distribution function $g(r)$ (Equation 3.6) for the discrete distance bins r_i and r_j with $|i-j| \leq 2$, that is, distance bins not further separated than $2\delta r$, or 1 Å. The weighting factor was chosen as $2^{-|i-j|}$, thus assuming values of 1 for the central bin under consideration, $1/2$ for neighbouring bins, and $1/4$ for bins separated by one interjacent bin. Therefore, instead of using $g(r_i)$, the weighted sum $\sum_{k=-2}^{2} 2^{-|k|} g(r_{i+k})$ was used. The smoothing procedure was repeated once. Radial distribution functions were then scaled to have the same integral as before the smoothing.

The resulting, final rdfs are described in Table 3.2 and Figure 3.5. To obtain a *xenon likeness score* for a given point in three dimensional space, a scoring function was defined as the sum of the negative natural logarithms of $g_{\text{het}}(r)$ for all atom types and instances of that atom type,

$$\textit{xenon likeness score} = -\sum_{\text{het}} \sum_{A \in \text{het}} \ln(g_{\text{het}}(r_A) + \epsilon) \tag{3.7}$$

summing over all instances A of all types of protein atoms *het* with a distance of r_A from xenon, and $\epsilon = 10^{-6}$ to avoid $\ln(0)$ for small r.
Subsequently, the scoring function was validated by a *receiver operating characteristic* (ROC) analysis, and score cutoff values were determined suitable to discriminate xenon atoms from non-xenon atoms.

3.2.4 Validation of the scoring function

In order to validate the xenon likeness scoring function, the data set used to derive it (470 xenon atoms in 99 PDB records) was split into non-overlapping training and test data sets such that each xenon atom would appear exactly once in each test data set. To achieve this, the data were evenly distributed over five bins, where one bin was used to test a scoring function derived from the other four bins, following the steps outlined above. Thus, each bin, and hence each xenon atom, and other hetero atom present in any PDB record, was used for testing exactly

Table 3.2 – Radial distribution functions $g_{het}(r_0)$ of xenon environments for 25 CHARMM22 protein atom types (MacKerell et al., 1998), with $r_0 = r_{vdw,Xe} + r_{vdw,het}$ and $r_{vdW,Xe} = 2.2$ Å (Cohen et al., 2006; Liu et al., 2010). For reference, also van der Waals energy parameters ε_{het} are given (see section 3.3.7).

atom type *het*	$r_{vdw,het}$ (Å)	$g_{het}(r_0)$	ε_{het} (kcal mol^{-1})
S	2.00	4.01	−0.45
CT3	2.06	3.98	−0.08
CA	1.99	3.18	−0.07
CT1	2.28	2.00	−0.02
CY	1.99	1.80	−0.07
NY	1.85	1.60	−0.2
CT2	2.17	1.51	−0.055
CPT	1.80	1.47	−0.09
OH1	1.77	1.18	−0.1521
CPH2	1.80	1.16	−0.05
NR1	1.85	1.05	−0.2
CP1	2.28	1.01	−0.02
C	2.00	0.94	−0.11
NR2	1.85	0.90	−0.20
CP2	2.17	0.89	−0.055
O	1.70	0.79	−0.12
CC	2.00	0.78	−0.07
NH1	1.85	0.78	−0.2
NH2	1.85	0.59	−0.2
CP3	2.17	0.56	−0.055
CPH1	1.80	0.54	−0.05
N	1.85	0.51	−0.2
OC	1.70	0.45	−0.12
NC2	1.85	0.33	−0.2
NH3	1.85	0.27	−0.2

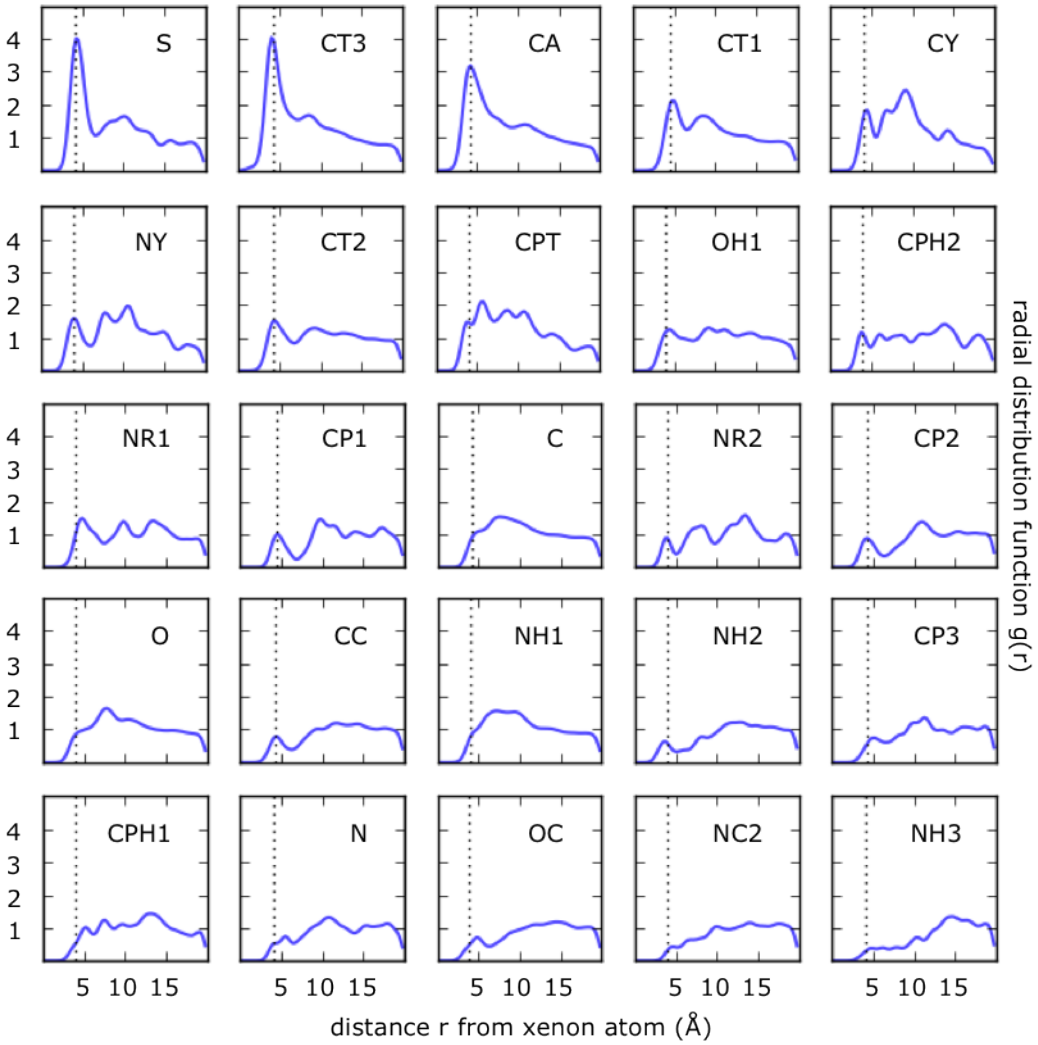


Figure 3.5 – Radial distribution functions of protein xenon environments of 25 CHARMM22 protein atom types (see Methods section 3.4.8 and Table 3.6 for atom type definitions). Atom types *het* are given in the panels and sorted by decreasing value of $g_{\mathrm{het}}(r)$ at distance $r_0 = r_{\mathrm{vdw,Xe}} + r_{\mathrm{vdw,het}}$ (indicated by vertical dotted lines) with $\mathrm{r_{vdW,Xe}} = 2.2$ Å and $\mathrm{r_{vdw,het}}$ taken from the CHARMM22 force field (MacKerell et al., 1998) and the literature (Cohen et al., 2006; Liu et al., 2010). Values of $r_{\mathrm{vdw,het}}$ and $g_{\mathrm{het}}(r_0)$ are tabulated in Table 3.2. Functions $g(r)$ for all atom types are tabulated in Supplementary Table S3 in the Appendix.

once, and test sets were never part of the training procedure, but exclusively used to report the performance of the classifier in an unbiased way. It should be noted that the method aims to discriminate xenon atoms from non-xenon atoms for a given position, and not xenon-binding proteins from non-xenon binding proteins; positive and negative samples are thus necessarily present in the same protein/PDB record. Training data contained all xenon atoms present in the PDB records in that respective bin, while test data contained xenon atoms *and* non-xenon atoms (see below) in the other four complementary bins.

The partitioning of data sets was achieved by randomly selecting one UniProt identifier and assigning all of its associated xenon atoms to one bin. This was then iterated until approximately 20 % of the 470 xenon atoms were assigned to that bin.[18] The xenon atoms of the next randomly selected UniProt identifier were then assigned to another bin, until all PDB records and xenon atoms were assigned to exactly one of the five bins. The bins in turn would constitute the test sets, and corresponding training data sets were simply their logical complement. From those five training sets five different xenon likeness scoring functions were then derived as described above. A comparison between the rdfs derived from the (smaller) training data sets and the full reference data set is shown in Figure 3.6. It can be appreciated that rdfs for atom types occuring more rarely in the data set, such as S (sulphur) or CPT (aromatic carbon atoms between the 5- and 6-membered tryptophan heterocycles) display more variability sets than those most frequently occuring, such as CT3 (methyl carbon atoms), or C (carboxyl carbon atoms).

The five scoring functions were then applied to score every hetero atom position in their associated test data set. Depending on a variable scoring cutoff t, hetero atoms were predicted to belong to either the class 'xenon' (if their score was $\leq t$), or 'other' (else). Hetero atoms considered for this validation were xenon atoms, the most common ions found in buffers used in biochemistry (sodium, potassium, chloride, magnesium, and calcium), and water molecules.[19] Xenon atoms scored

[18] ideally, exactly 94 xenon atoms would be assigned to each bin. However, more relaxed termination criteria were employed where a bin was considered complete if it contained between 92 and 98 xenon atoms. A randomly selected PDB record was rejected if incorporation of its xenon atoms into the current bin led to more than 98 xenon atoms in that bin, and an alternative UniProt identifier fulfilling this requirement was then selected. The procedure was restarted in case this could not be achieved.

[19] more precisely, the position of water molecule oxygen atoms.

$\leq t$ were considered *true positives* (TP), 'other' atoms scored $> t$ *true negatives*, xenon atoms scored $> t$ *false negatives* (FN), and 'other' atoms scored $\leq t$, *false positives* (FP).

The quality of these predictions was evaluated using common performance measures from the field of *machine learning*: the area under the curve of the *receiver operating characteristic*, or ROC curve, called AUC; Matthews Correlation Coefficient (MCC); Youden's index; and the minimal distance of the ROC curve to the point of perfect classification.[20] Technical details of the ROC analysis and the definition of MCC and Youden's index can be found in Methods section 3.4.9. Results for the validation are summarised in Table 3.3. The AUC was found to be excellent (values of 0.97 to 0.99), with high values of MCC (0.58 to 0.84) and Youden's index (0.80 to 0.92), and a low minimum distance to the point of perfect classification (0.06 to 0.16), suggesting a high discrimination power of the classifier.

Since the rdfs derived from the training data sets were found to not differ much from the rdfs derived from the entire data set (Figure 3.6), and the validation performed suggests a high quality of the classifier, the full scoring functions derived from the entire data set was used to score the test data sets for comparison (Figure 3.7). The high correlation of the full scoring function with the five training scoring function suggests the absence of over-fitting of the data; therefore, it was used throughout the remainder of this chapter.

3.2.5 Determination of discrimination threshold values

The characteristics and quality of a binary classifier are crucially dependent on the discrimination threshold value used. If the threshold value is set too low, the classifier might be too *conservative*, being specific but not sensitive enough;[21] on the other hand, with too high a threshold value specificity would be sacrificed to increase sensitivity.

A *receiver operating characteristic* (ROC) analysis, as performed for the validation of the xenon likeness scoring potential described above, and described in more technical detail in Methods section 3.4.9, investigates the mutual balance

[20] where $specificity = 1$ and $sensitivity = 1$.

[21] assuming *true positive* samples score *lower* than *true negative* samples.

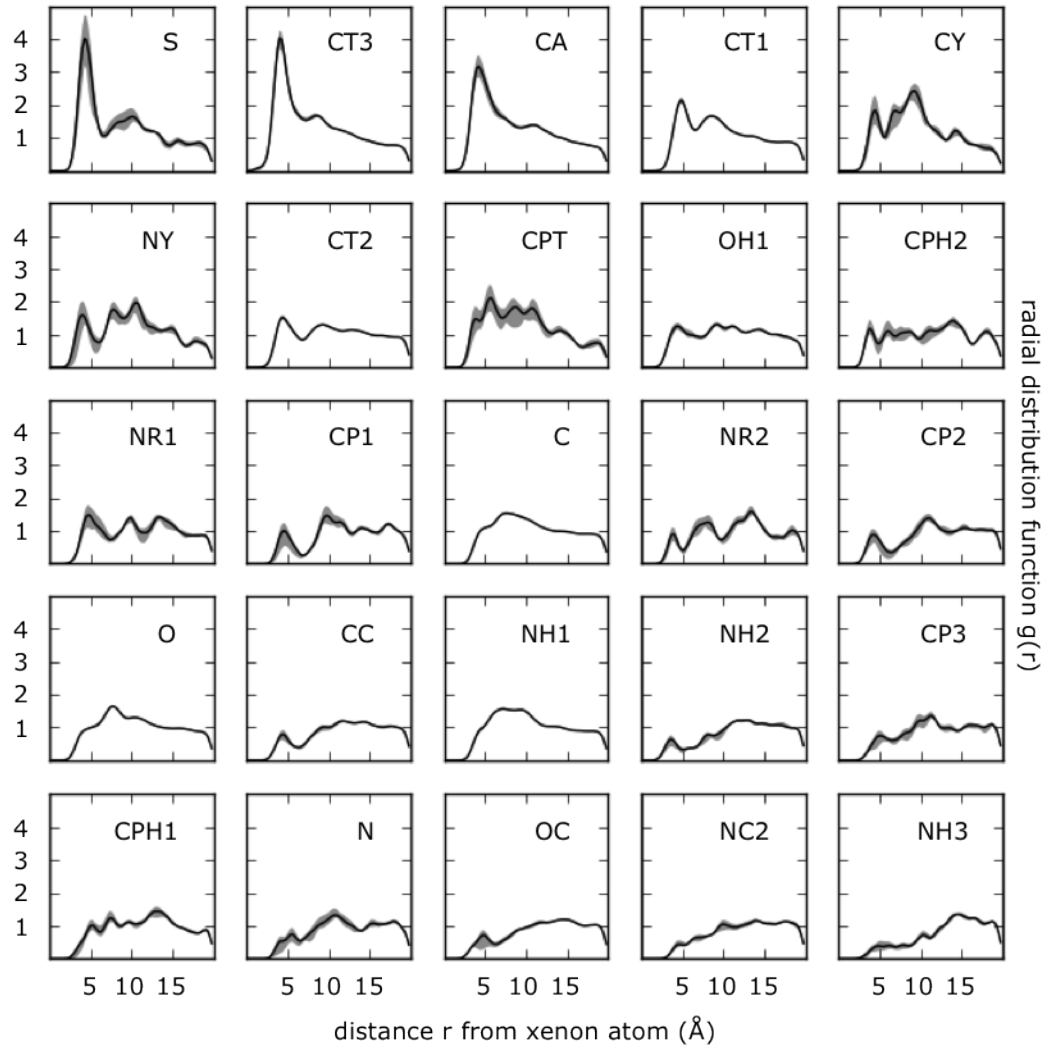


Figure 3.6 – Radial distribution functions (rdfs, black) for different atom types are shown in comparison to rdfs $g_{\mathrm{het},i}$ derived from reduced training data sets (grey, later used to validate the scoring function), consisting of approximately 80 % of all xenon atoms. For simplicity, instead of five individual training data rdfs $g_{\mathrm{het},i}$ for $i = 1, 2, 3, 4, 5$ being shown, the area between the minimum and maximum of their five local values $min_i/max_i(g_{\mathrm{het},i}(r_0))$ for a given protein atom type *het* at distance r_0 is shaded grey.

Table 3.3 – Validation of the xenon likeness score.

	partition 1	partition 2	partition 3	partition 4	partition 5
training set size[a]	377 (88)	377 (76)	377 (80)	373 (84)	376 (68)
test set size[b]	93 / 4,024	93 / 7,171	93 / 6,504	97 / 5,361	94 / 6,010
AUC[c]	0.99	0.96	0.97	0.97	0.99
MCC[d]	0.84	0.73	0.58	0.72	0.80
Youden's index[e]	0.88	0.87	0.80	0.85	0.92
ROC-distance d[f]	0.09	0.12	0.16	0.11	0.06
MCC Y[g]	0.69	0.61	0.43	0.35	0.53
MCC d[h]	0.45	0.61	0.28	0.38	0.53

[a] number of xenon atoms (proteins) in training set.
[b] number of xenon atoms/other hetero atoms in test set, that is, the number of positives/negatives in that data sub set.
[c] area under ROC curve.
[d] maximum value of Matthews Correlation Coefficient (MCC).
[e] maximum value of Youden's index.
[f] minimum distance to the point of perfect classification.
[g] MCC at the maximum value of Youden's index (above[e]).
[h] MCC at the minimum distance to perfect classification (above[f]).

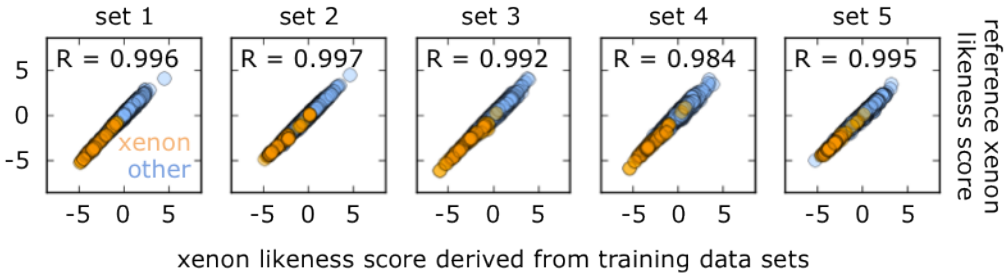


Figure 3.7 – The scoring functions based on the radial distribution functions derived from training data sets 1 to 5 (left to right, x-axes) are compared to the scoring function derived from the entire data set (y-axes). For comparability, values are expressed as *z-scores*, that is $z = (x_0 - \mu_x)/\sigma_x$ with x_0 the original score, μ_x the sample mean, and σ_x the sample standard deviation. Sample means and standard deviations were computed from the entire test and reference set(s), respectively. Scores for xenon atoms are represented as orange circles, those of 'other' hetero atoms in light blue. Pearson correlation coefficients R were calculated and are reported in the panels. For a discussion on the validity of the z-scores, see Figure 3.8 and main text.

of *sensitivity* and *specificity*, and is thus a possible technique to locate reasonable threshold values. Figure 3.8 shows the distribution of scores for xenon atoms and other hetero atoms, including the most common ions, and water molecules in the entire data set, spanning 99 PDB records. There are 470 xenon atoms (*positives*) and 29,080 'other' hetero atoms (*negatives*) in the data set.

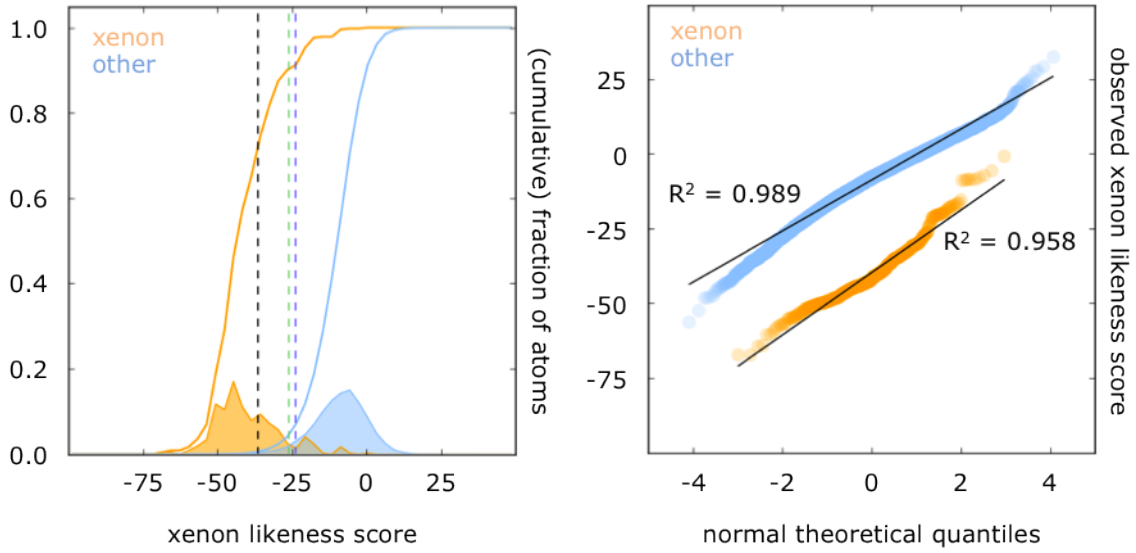


Figure 3.8 – Distribution of xenon likeness scores of hetero atoms in xenon binding proteins in the Protein Data Bank. Left side, xenon likeness scores of xenon atoms (orange) and other hetero atoms (light blue), including ions of sodium, chlorine, potassium, magnesium, and calcium, and water molecules present in the 99 protein structures that were previously used to derive the scoring potential. Data were binned into 50 bins of equal width spanning the interval delimited by the minimum and maximum scores encountered in the analysis, and normalised histograms (shaded areas) and cumulative normalised histograms (solid lines) are shown. Orange and light blue areas represent 470 and 29,080 atoms, respectively. Vertical dashed lines indicate possible threshold values from a subsequently performed *receiver operating characteristic* analysis (see main text) coinciding with the maximum of *Matthews Correlation Coefficient* (black, −36.6), *Youden's index* (green, −26.1), or the smallest distance of any *sensitivity/specificity* pair to perfect classification (blue, −23.8), from left to right (see Methods section 3.4.9 for details). Right side, quantile-quantile (Q-Q) plot of xenon likeness scores of xenon (orange) and other atoms (blue) against quantiles from a theoretical standard normal distribution (x-axis). Best fit lines from linear regression analysis (black) and squared values of the Pearson correlation coefficient (R^2) for that regression line are indicated.

To determine possible discrimination threshold values, a ROC analysis was

carried out (Figure 3.9), yielding an *area under (ROC) curve* (AUC) of 0.98, and a *negative accuracy* of close to the optimum of 1 over almost the entire range of *specificity* values. Threshold scanning analysis suggests three possible values for the discrimination threshold, based on the maximum of Matthews Correlation Coefficient (MCC), the maximum of Youden's index, and the minimal distance to the point of perfect classification: −36.6 (MCC of 0.72), −26.1 (Youden's index of 0.86), and −23.8 (minimal distance to point of perfect classification of 0.10), as summarised in Table 3.4. Values of MCC at maximum Youden's index and minimal distance to the point of perfect classification are 0.51 and 0.44, respectively. The performance of the overall classifier is thus very similar to the characteristics of the classifiers derived from the smaller training sets (see Results section 3.2.4 and Table 3.3 for comparison).

The average score for a *positive* sample was found to be −39.8 with a standard deviation of 10.6 and a median value of −42.0, while the average score for a *negative* sample was −8.8 (standard deviation 8.6, median −8.1). The average score over all positive and negative samples was −9.3 (standard deviation 9.5, median −8.3). Thus, the three possible score thresholds expressed as *z-scores*, that is $z = (x_0 - \mu_x)/\sigma_x$ with original score x_0, sample mean μ_x and standard deviation σ_x, i.e. the number of standard deviations from the sample mean, were −2.87 (MCC), −1.77 (Youden's index) and −1.53 (minimal distance to point of perfect classification), respectively. The assumption of the data to be normally distributed was found to hold true to good approximation (see quantile-quantile (Q-Q) plot at the right of Figure 3.8), allowing for meaningful computation of σ and z-scores. Figure 3.10 provides a further discrimination threshold dependent analysis of the positive/negative coverage/accuracy.

3.2.6 Probabilistic interpretation of xenon likeness scores

An alternative useful interpretation of the xenon likeness score can be obtained by converting a given score into the probability that a xenon atom is found at a certain position within a protein structure: given a xenon likeness score s_0, the probability $P(\mathrm{XE}|s < s_0)$ can be estimated using *Bayes' theorem* which states

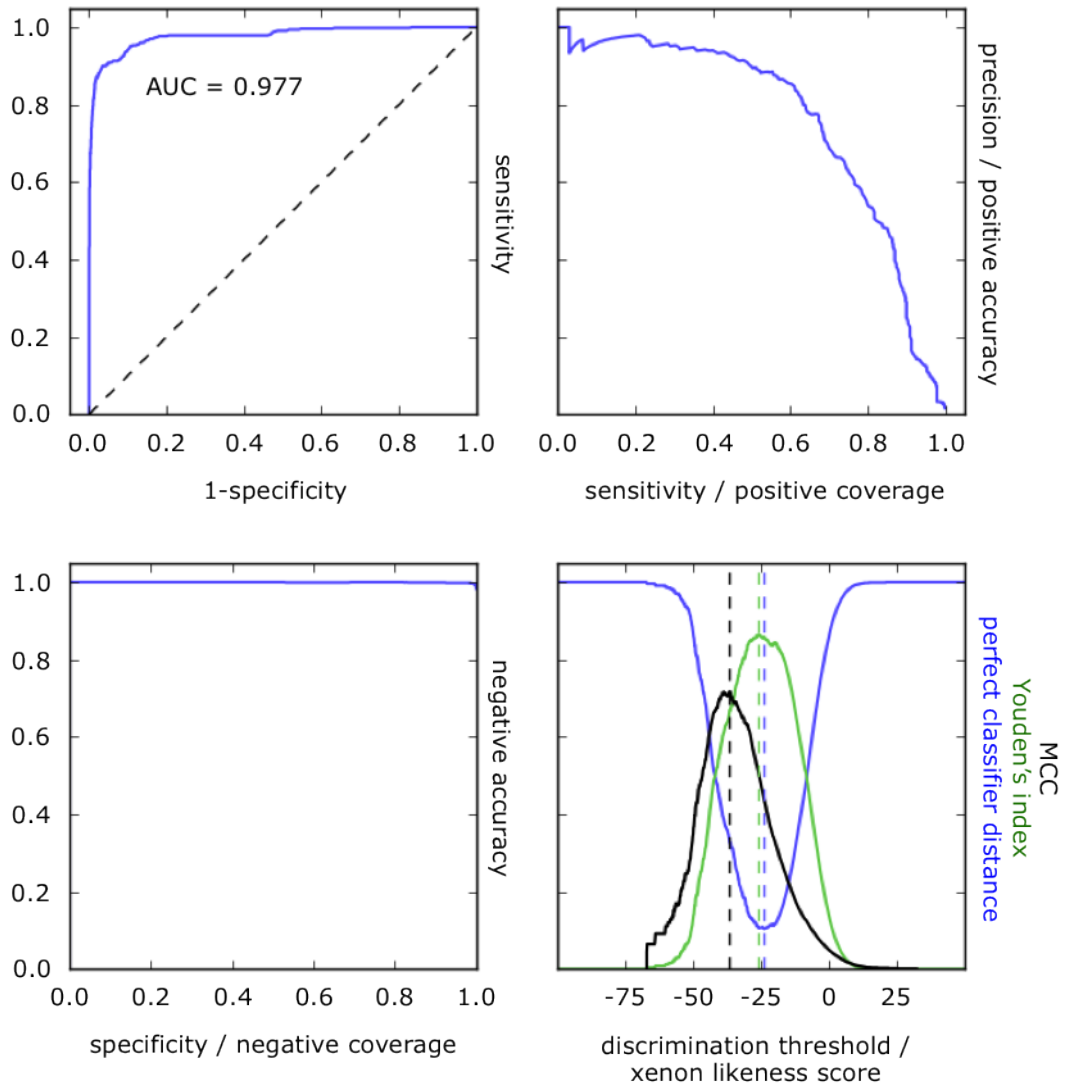


Figure 3.9 – *Receiver operating characteristic* analysis of the xenon likeness score (top left) yields an *area under curve* (AUC) of 0.977. Top right, plot of *sensitivity* against *precision*. Bottom left, *negative coverage* against *negative accuracy*. Bottom right, score threshold scan analysis and optima (dotted vertical lines) of Matthews Correlation Coefficient (MCC, black, −36.6), Youden's index (green, −26.1) and distance to the point of perfect classification (blue, −23.8). Definitions of all quantities are given in Methods section 3.4.9.

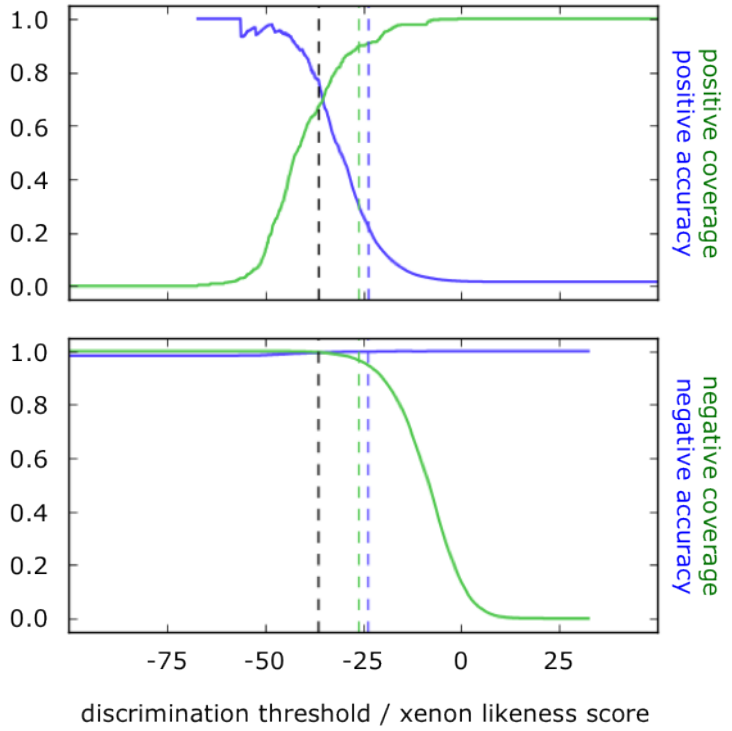


Figure 3.10 – Score dependence of positive (top panel) and negative (bottom panel) coverage (green) and accuracy (blue). Optima of Matthews Correlation Coefficient (black), Youden's index (green), and the distance to the point of perfect classification (blue), as determined by a threshold scan, are indicated as vertical dotted lines at positions −36.6, −26.1, and −23.8, respectively (see also Table 3.4 and Methods section 3.4.9). It should be noted that due to the definition of positive and negative accuracy, with either *only* positive or negative predictions (i.e. values of the scoring function *below* or *above* the variable discrimination threshold value) in the denominator of the formula, their value is not defined for all possible threshold values.

Table 3.4 – Discrimination threshold analysis of the xenon likeness score. *Sensitivity*, *specificity* and the contents of the *confusion matrix*, consisting of the number of *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN), are shown for three different score threshold values. Those values were obtained by scanning possible thresholds for the maximum of Matthews Correlation Coefficient (MCC), Youden's index, and the minimum distance to the point of perfect classification (min distance). Xenon atoms (other atoms) scoring below the threshold value are considered TP (FP), and xenon atoms (other atoms) scoring above the threshold are considered FN (TN).

criterion	threshold	sensitivity	specificity	TP	TN	FP	FN
MCC	−36.6	0.67	1.00	316	28,978	92	154
Youden's index	−26.1	0.90	0.97	422	28,089	981	48
min distance	−23.8	0.91	0.95	427	27,601	1,469	43

$P(A|B) = P(B|A)P(A)/P(B)$, thus

$$\begin{aligned} P(\mathrm{XE}|s < s_0) &= \frac{P(s < s_0|\mathrm{XE})P(\mathrm{XE})}{P(s < s_0)} \\ &= \frac{P(s < s_0|\mathrm{XE})P(\mathrm{XE})}{P(s < s_0|\mathrm{XE})P(\mathrm{XE}) + P(s < s_0|\neg\mathrm{XE})P(\neg\mathrm{XE})} \end{aligned} \tag{3.8}$$

with $P(s < s_0|\mathrm{XE}) = TP/(TP + FN)$, $P(s < s_0|\neg\mathrm{XE}) = FP/(FP + TN)$, $P(\mathrm{XE}) = P/(P + N)$, $P(\neg\mathrm{XE}) = N/(P + N)$, and TP, TN, FP, FN, P, and N, true positives, true negatives, false positives, false negatives, positives, and negatives, respectively.

Figure 3.11 shows the correspondence between xenon likeness scores and probabilities calculated according to Eq. 3.8 on the data set of xenon binding proteins corresponding to the ROC analysis carried out before (Figure 3.9). Probabilities for the suggested score threshold values of −36.6, −26.1 and −23.8 are 0.78, 0.30 and 0.23, respectively, with the *break even point* of $P(\mathrm{XE}|s < s_0) = P(\neg\mathrm{XE}|s < s_0)$ at a score of about −30.8. For $s_0 \to +\infty$, $P(\mathrm{XE}|s < s_0)$ approaches $P/(P + N) \approx 0.02$.

The sigmoidal shape of the functions shown in Figure 3.11 suggests them to be well approximated by a cumulative distribution function (CDF) of a probability distribution, possibly that of a normal distribution. The CDF Φ of a standard normal distribution with probability density function $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is $\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2} dt$ which is closely related to the error function erf, stating

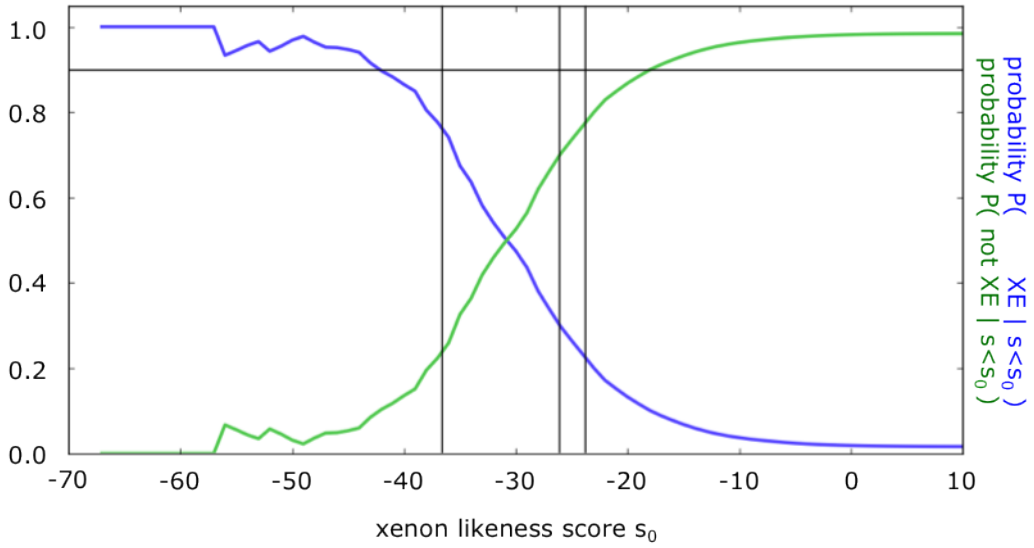


Figure 3.11 – Probability of observing a xenon atom in dependence of a given xenon likeness score. Probabilities of finding a xenon atom (blue) or no xenon atom (green) at a given position were calculated from Eq. 3.8. Vertical black lines indicate xenon likeness score thresholds suggested by a threshold scanning analysis (see main text). The horizontal black line indicates a probability of 0.9, corresponding to a xenon likeness score of about −44.7 (blue line).

the probability that a random variable drawn from a standard normal distribution falls into the range $[-x, x]$ can be computed as $erf(x) = \frac{1}{\sqrt{2}} \int_{-x}^{x} e^{-t^2} dt$ and for which a approximate closed formula exists (see below). It can be seen that $\Phi(x) = \frac{1}{2}\left(1 + erf(\frac{x}{\sqrt{2}})\right)$.

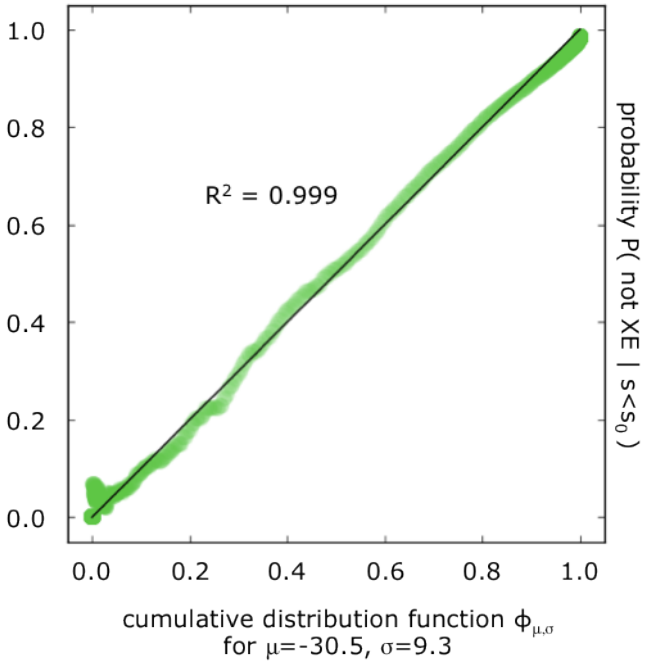


Figure 3.12 – Fit of the probability density function for xenon binding with a sigmoidal curve. The cumulative distribution function (CDF) $\Phi_{\mu,\sigma}$ for a normal distribution ϕ with mean μ and standard deviation σ (x-axis) was fitted against the empirical probability of *not* observing a xenon atom at a given protein position, given a specific score (see Figure 3.11 and main text, y-axis). The fit was obtained minimising the *sum of squared differences* of all pairs of values sharing the same xenon likeness score (see main text). Best fit parameters were found to be $\mu = -30.5$ and $\sigma = 9.3$, yielding a Pearson correlation coefficient (R) of $R^2 = 0.999$. The back line indicates the function $y = x$ for $x \in [0, 1]$.

Formulæ for generic normal distributions with mean μ and standard deviation σ are obtained by substituting x by $(x-\mu)/\sigma$, arriving at $\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$. A *grid search* was employed to determine values for μ and σ such that the corresponding CDF $\Phi_{\mu,\sigma}$ minimises the *sum of squared differences* of all score pairs $\Phi_{\mu,\sigma}(i)$ and $P(\neg\text{XE}|s < i)$ for scores $-70 \leq i \leq +10$ and $\Delta i = 0.01$.

By scanning values of μ around the *break even point* (see above) and an initial

estimate of σ of around 10,[22] and progressively using more fine grained search intervals, optimal values of $\mu = -30.5$ and $\sigma = 9.3$ were obtained. Figure 3.12 shows the goodness of fit of this approximation. The Chebyshev fitting of the error function was used (see Methods section 3.4.11).

3.3 Discussion

The derivation and validation of a knowledge-based xenon likeness score has been presented that can be used to discriminate between xenon atoms and non-xenon atoms, based on their environment in three dimensional space, given a protein structure and a position to evaluate. The scoring function is based on radial distribution functions (rdfs) for pairs of xenon and instances of 1 of 25 different protein atom types. The rdfs were derived from 470 individual xenon atoms covering 160 protein chains falling into 64 classes of unique UniProt identifiers. Some xenon atoms are located close to interfaces between instances of the asymmetric unit, influencing the number of neighbouring non-hydrogen protein atoms, and their solvent accessible surface area (Results section 3.2.1). A high number of xenon atoms (28.2 %) was found to partially overlap with protein ligand binding sites. The majority of ligand atoms found in close proximity of xenon in super-imposed protein *holo* structures are aliphatic, i.e. sp^3 hybridised or aromatic carbon atoms, but also halogen atoms were frequently found (Results section 3.2.2). Consistently, computation of rdfs between xenon and 25 different protein atom types suggests that xenon atoms prefer an aliphatic and therefore hydrophobic environment, i.e. rdfs show an enrichment of small pairwise distances of xenon and sulphur and sp^3 hybridised/aromatic carbon atoms, while polar nitrogen and oxygen are more frequently found in larger distances (Results section 3.2.3).

Validation of the scoring function suggests robustness with respect to excluding a number of randomly chosen xenon atoms from the derivation of the scoring function, and consistent scoring between those reduced scoring functions and the reference scoring function derived from the entire data set. By employing a *receiver operating characteristic* analysis, the discrimination power of the scoring function

[22] corresponding to standard deviations observed for *positive* and *negative* samples in the threshold scan analysis conducted in Results section 3.2.5.

was found to be excellent as documented by cumulative performance measures such as the *area under curve* (AUC, values ≥ 0.96), Matthews Correlation Coefficient (MCC ≥ 0.58), Youden's index (≥ 0.80), and the minimal distance to the point of perfect classification (≤ 0.16) (Results section 3.2.4). The score distributions of xenon and non-xenon atoms were found to be well separable and little overlapping. Depending on which criterion was used to determine a score cutoff value, values of -36.6, -26.1, and -23.8 were found, in decreasing conservativeness, corresponding to -2.87, -1.77, and -1.53 standard deviations from the mean score, respectively (Results section 3.2.5).

3.3.1 Accounting for redundancy in the data set

An initial sequence analysis had suggested a potential bias in the data set inasmuch as 16 out of 160 xenon-binding protein chains were found to be uricase, with an additional 9 instances of lysozyme, and several proteins from the globin family; some further clusters with high mutual sequence identity were identified as well (Figure 3.1). As the intention was to devise a generally applicable scoring function allowing the application to protein classes not present in the training data, this sequence bias had to be accounted for. It is conceivable that the over representation of certain protein classes in the training data could lead to a specialised classifier, performing well on the over-represented protein classes, but potentially poorly on others less well represented.

Instead of proposing a simple selection criterion, such as selecting only the most highly resolved crystal structure for a given protein and omitting the rest, a more robust approach was chosen in order to make use of data redundancy and correct for data imperfections at the same time, but without excluding any of the limited training data. To this end, a weighted averaging procedure was designed to combine the rdfs of different xenon atoms, based on the mutual similarity (yet to be defined) of the protein chains they were bound to. The procedure would iteratively average rdfs, starting with those of xenons bound to identical proteins and terminating with maximally dissimilar proteins, thus averaging highly similar data first to exploit redundancy, and averaging highly dissimilar data later to reflect diversity.

One possible choice for a similarity measure would be the percentage of protein amino acid sequence identity, making the assumption that proteins with 100 % sequence identity would also be structurally equivalent, and that in general sequence similarity corresponds to structural similarity. Global sequence identity was preferred over local sequence identity because of the benefit of having a single measure of similarity (i.e. global sequence identity) as opposed to a composite measure (i.e. local sequence identity plus alignment length). As a consequence, alignment significance did not have to be examined in detail, either, since, given the forced length of the alignments, high scoring alignments occurring by chance could safely be ruled out.

3.3.2 Unit cell expansion prevents boundary effects

It has been observed that some xenon atoms were located close to symmetry related protein interfaces, that is, in proximity to unit cell boundaries, or multiple instances of the asymmetric unit (Results section 3.2.1). In consequence, the number of neighbouring protein atoms for a given xenon atom might be underestimated if just atoms within a single asymmetric unit were to be considered (Figure 3.3). It is evident from the functional form of the xenon likeness score (Equation 3.7) that each protein atom within a distance of 20 Å to any xenon atom contributes to the xenon likeness score. Thus, neglecting 'image atoms', i.e. atoms in copies of the original asymmetric unit, would influence both score derivation and scoring. To circumvent this, the crystal lattice was reproduced by generating images of the asymmetric unit as determined by the crystal space group, and then translating these unit cell contents, leading to an increased number of protein hetero atoms in the neighbourhood of the xenon atoms, and to a reduction in their solvent accessibility (Figure 3.4). It is conceivable that in an experimental setting when the crystal is derivatised with xenon gas, a xenon atom might be attracted to the interface area between multiple unit cells, or multiple instances of the asymmetric unit, interacting with protein chains in each one of them. The same xenon atom would perhaps not interact with one of the proteins in that interface alone, given that the interactions contributed by the other unit cell or asymmetric unit copies could be missing entirely, thereby partially exposing the xenon atom to a

hypothetical vacuum, or buffer. Expanding the unit cell computationally therefore leads to a more realistic representation of the experimental crystal lattice and avoids boundary effects.

3.3.3 Protein ligand binding sites are susceptible to xenon binding

It was found that roughly 28 % of xenon atoms are situated in ligand binding sites, partially overlapping with, or within 3 Å of non-hydrogen ligand atoms (Results section 3.2.2) in other related protein complexes. This value increases to over 35 % if the inter-nuclear distance cutoff is increased from 3 Å to a more generous 4 Å. These cutoffs were chosen to guarantee a partial overlap between the xenon atom and ligand under consideration so that they would effectively compete for the same binding site within a given protein, and simultaneous binding would not be possible. They seem reasonable given an assumed xenon van der Waals radius of about 2.2 Å (Cohen et al., 2006; Liu et al., 2010) and the fact that most non-hydrogen protein atoms were modeled as having a van der Waals radius between 1.85 and 2 Å (Table 3.2). However, slightly larger cutoffs could also be considered, since in the analysis protein hydrogen atoms have not been included explicitly. In any case a more generous cutoff would lead to the inclusion of an identical or higher number of ligands but never to a lower number.

This finding underlines the relevance of predicting xenon binding, as for a given xenon there appears to be a likelihood of about one third of being located in a ligand binding site. Thus, a method with high discrimination power used to determine xenon binding sites could potentially also be employed to predict druggable protein ligand binding sites. It has been found that the active catalytic site of an enzyme is in most cases found among the three largest cavities in the surface of the protein (Weisel et al., 2007); the xenon binding prediction could therefore serve as complimentary information in the detection of potentially druggable protein ligand binding sites. As evident from the distribution of the atom types of the ligand atom found most closely to xenon (Table 3.1), a wide range of ligand classes could be covered, as most drug-like ligands will contain aromatic or sp^3 hybridised carbon atoms, or halogens.

3.3.4 Radial distribution functions represent protein atom type characteristics

The derivation of the rdfs is in intuitive agreement to the complimentary information obtained from the ligand analysis (see above, and Results sections 3.2.2 and 3.2.3); xenon atoms are most frequently found close to protein hetero atoms that are of predominantly uncharged, hydrophilic nature such as sulphur, aromatic, sp^3 hybridised and other aliphatic carbon atoms, and less often close to polar oxygen or nitrogen atoms (Table 3.2). The peaks of the rdf functions in the cases of favourable interaction coincide well with the sum of the *van der Waals* radii of the two atoms (Figure 3.5), indicating that interactions are predominantly dispersive. There are subtle but visible differences between the rdfs for co-occurring but non-identical atom types, such as CY and NY, two atoms in the tryptophan hetero cycle, or CP1, CP2 and CP3 carbon atoms of the sterically restricted proline cycle.

It is conceivable that the way of assigning atom types will influence the characteristics of the scoring function, and anticipating the optimal atom typing scheme *a priori* is rather challenging. Too crude an atom typing might not represent necessary subtleties well enough to discriminate between different environments,[23] while too detailed an atom typing sceme might require too many different parameters to be estimated, and lead to sampling problems: the concept of a knowledge-based potential requires each relevant atom to appear in the relevant, preferred distance context often enough that the characteristics of its interaction can be captured by observing and sampling it multiple times. In case of too few instances of a given atom type, spurious distance occurrences could influence the distance distribution too much.[24] Using the 25 CHARMM22 atom types (Table 3.6 and Methods section 3.4.8) seemed a good compromise of not being too detailed yet still realistic in the sense that they roughly correspond to certain hybridisation states of protein hetero atoms, and have been successfully used in molecular dynamics studies, suggesting that they could provide the right level of abstraction for this type of analysis.

In order to reduce the number of atom types, and parameters to estimate,

[23] e.g. an atom typing scheme just grouping atoms by their chemical element.

[24] e.g. an atom typing scheme that groups per atom type *and* per amino acid type, for instance, splitting sulphur into two atom types, with one atom type for sulphur atoms contained in cystein and another one for sulphur atoms in methionine amino acid side chains.

certain atom types which frequently co-occur close to each others in the same amino acid side chain could be grouped to form a combined atom type. However, as can be seen for atom types CP1, CP2 and CP3,[25] they seem to possess certain individual features that lead to non-identical rdfs. On the other hand, non-identical charges of the same atom type in different amino acid side chains as suggested by CHARMM22 could lead to a finer atom typing with more classes; however, charges of instances of the same atom type do not vary wildly across different amino acid side chains, and as mentioned previously, further segregation of atom types into more subtypes could induce a sampling problem hampering the derivation of the potential. Thus, it was decided to not modify the atom typing provided by the CHARMM22 force field and rather directly use it in this study.

3.3.5 Discrimination of xenon atoms from water molecules and ions

The validation of the xenon likeness score suggests a high discrimination power based on a clear separability of positives and negatives and little overlap of the distribution of their scores (Results section 3.2.4 and Tables 3.3 and 3.4). It also suggests the robustness of the approach to derive the score, and the absence of overfitting of the data. The performance of five scoring functions derived from a reduced set of training data, and tested on independent, non-overlapping training data is consistently good (Table 3.3). However, as can be seen in the plot of *precision* versus *sensitivity*, if high sensitivity is requested, precision decreases markedly (see Figure 3.9). This is also visible in the *precision* values computed from the *confusion matrix* at certain threshold values given in Table 3.4: at a threshold value of -36.6 (corresponding to a sensitivity value of 0.67), the classifier has a precision of 0.78, while at threshold values of -26.1 (sensitivity 0.90) and -23.8 (sensitivity 0.91) the precision drops to 0.30 and 0.23, respectively.

Precision is defined as $TP/(TP+FP)$, while sensitivity is defined as $TP/(TP+FN)$ (see Methods section 3.4.9). In developing a classifier both precision and sensitivity are to be maximised, thus minimising the number of FP and FN simultaneously. Usually, however, these quantities are negatively correlated. In a

[25] sterically restrained carbons in the secondary amine hetero cycle of proline amino acids.

scenario where $N \gg P$, it is likely that $FP > FN$, and thus precision declines with increasing sensitivity. This is the case for the data set of xenon binding to proteins, where there are 470 positive and 29,080 negative samples, corresponding to a factor of more than 60 times more negative than positive samples. In Figure 3.10, the trade off between sensitivity (positive coverage) and precision (positive accuracy) is shown in a discrimination threshold dependent manner. While at a score threshold of around -36.6 the sum of both quantities is close to an optimal value, for a more conservative classifier (i.e. moving to smaller threshold values), precision rises but sensitivity declines (Results section 3.2.5). Conversely, a less conservative classifier allows for more sensitivity, but at the expense of precision. The relatively high number of false positives compared to true positives at less conservative, higher xenon likeness scores is mainly a result of the high number of negative samples, and is likely to contain a certain amount of 'false false positives', i.e. water molecules/ions obtaining favourable xenon likeness scores and that can indeed be replaces by xenon atoms, potentially at higher xenon concentrations, or different experimental conditions such as longer exposure of the crystal to xenon gas than those used to solve the structures under consideration. Therefore, the *precision* of the method is likely to be an underestimate.

A scoring threshold value of around -36.6, which was found by maximising the Matthews Correlation Coefficient (MCC), seems beneficial in terms of simultaneously optimising precision and sensitivity. The other two threshold values determined from *receiver operating characteristic* (ROC) derived quantities sacrifice precision for the benefit of a higher sensitivity. This is not surprising since the MCC due to its functional form is a *balanced* performance measure that is able to correct for very different sizes of classes. From this point of view, sensitivity and threshold values of around 0.7 each, or a sensitivity of 0.67 and a specificity of 0.78 at threshold -36.6, are the optimum of what this classifier seems able to achieve.

3.3.6 The xenon likeness score captures more information than a trivial atom count classifier

Next, the performance of the classifier is compared against a trivial classifier, basing the classification solely on the number of neighbouring non-hydrogen pro-

tein atoms within a defined distance. It can be appreciated that xenon atoms tend to possess more neighbouring non-hydrogen protein atoms than 'other' hetero atoms/ions and water molecules do (see top of Figure 3.13), leading to them being more buried and having a smaller solvent accessible surface area (data not shown). It would be tempting to classify xenon from non-xenon hetero atoms by such simple means since it would only require two parameters, i.e. a distance cutoff of non-hydrogen protein atoms to be considered, and the number of atoms as a discrimination threshold, thereby simplifying all the analysis and parametrisation necessary for an atom type based approach.

However, a systematic performance analysis of this class of trivial classifiers as a function of the distance cutoff used to define neighbouring atoms around an atom of interest indicates inferior performance to the classifier based on the xenon likeness score (bottom of Figure 3.13). Based on the cumulative performance measures employed, classification performed best when neighbouring protein atoms within a distance of 11 or 12 Å were considered: the classifier then achieved a maximum MCC of 0.22, a maximum of Youden's index of 0.65 to 0.66, and a minimum distance to the point of perfect classification of 0.24 to 0.25. A receiver operating characteristic analysis (Figure 3.14) revealed an area under curve of 0.88 to 0.89. This indicates a discrimination power higher than random, yet clearly inferior to the performance of the xenon likeness score (AUC 0.98, MCC 0.72, Youden's index 0.86, minimal distance to point of perfect classification 0.10; see Results section 3.2.5). The relatively poor value for MCC in comparison to the relatively high value of the AUC, and other performance measures is caused by the imbalance of the data set ($N \gg P$) as discussed above. Thus, clearly a classification based on the number of atoms around xenon is neither optimal nor sufficient to discriminate between xenon and non-xenon.

However, it does show that xenon prefers to be buried from the bulk solvent. This is reflected in the xenon likeness scoring function by the fact that each instance of a protein atom type *het* with $g_{\mathrm{het}}(r_0) > 1$ contributes to a favourable xenon likeness score.[26] The more of those atoms found in proximity to xenon, the more beneficial for the xenon likeness score. At the same time, estimation of the radial distribution function (Equation 3.6) of xenon and water, which has not been used

[26] i.e. *decreasing* it.

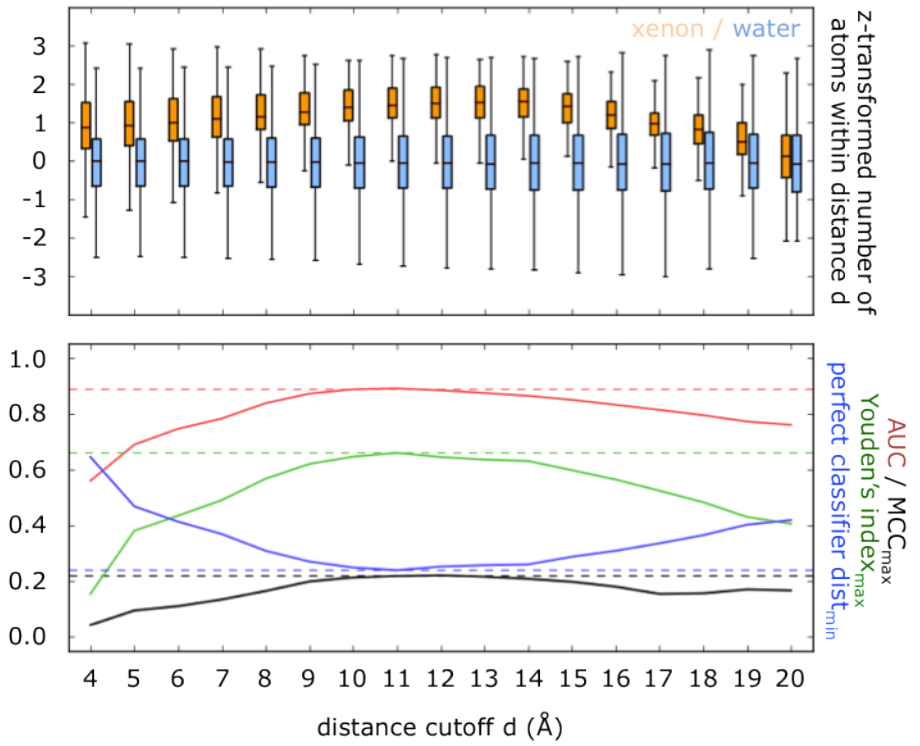


Figure 3.13 – Comparison of the number of neighbouring atoms of xenon and water molecules within different distance cutoffs d for the data set of 99 xenon binding proteins. In the top panel box plots are shown representing the distributions of the number of neighbouring protein atoms for xenon atoms (orange) and water molecules (light blue) for distance cutoffs between 4 Å to 20 Å. Data are represented as normalised values expressed in times standard deviations from the mean value of the respective distribution. In the boxes, horizontal lines represent the median values, and whiskers extend to the most extreme data point within $1.5 \times (75\% - 25\%)$ data range. In the bottom panel, for the same distance cutoffs, the performance of a set of binary classifiers discriminating xenon from non-xenon based on the number of neighbouring atoms is described by values of the area under the curve (AUC, red), Matthews Correlation Coefficient (MCC, black), Youden's index (green), and minimum distance to the point of perfect classification (blue), for a receiver operating characteristic analysis (see Methods section 3.4.9 for definitions).

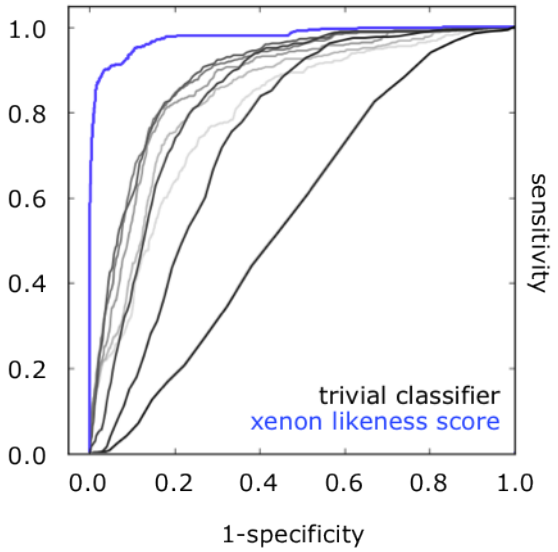


Figure 3.14 – Receiver operating characteristic (ROC) analysis of classifiers based on the number of protein atoms close to xenon in the data set of 99 xenon binding proteins. *Sensitivity* is shown as function of *1-specificity* for protein atoms within different distance cutoffs between 4 Å (dark grey) and 20 Å (light grey) from xenon atoms, yielding *areas under the curve* (AUC) of 0.56 to 0.89. The maximum occurs for the classifier with the 11 Å cutoff. For comparison, the ROC curve for the xenon likeness score is shown (blue; AUC=0.98; see Methods section 3.4.9 for definitions).

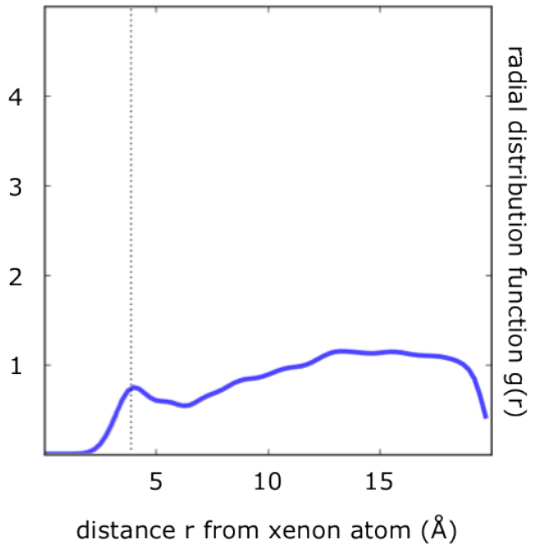


Figure 3.15 – Radial distribution function of xenon and water. Around an internuclear distance of $r_0 = 3.86$ Å (dotted vertical line), $g(r_0)$ assumes a local maximum value of 0.70, suggesting a mild dispreference of xenon ($r_{\mathrm{vdW}} = 2.2$ Å) to directly interact with water ($r_{\mathrm{vdW}} = 1.7$ Å). This radial distribution function has not been used in the calculation of the composite xenon likeness score (Equation 3.7) but can be compared to those for xenon and non-hydrogen protein atoms (see Figure 3.5).

to estimate the xenon likeness score (Equation 3.7), reveals a mild dispreference for xenon to interact directly with water (Figure 3.15): a local optimum of $g(r_0) = 0.70$ occurs at a distance of around $r_0 = 3.9$ Å, coinciding with the sum of the van der Waals radii of xenon (2.2 Å) (Cohen et al., 2006; Liu et al., 2010) and that proposed for water (1.7 Å) (Li and Nussinov, 1998).

3.3.7 The method outperforms a classifier based on an electrostatic contact potential

The results presented in this chapter so far indicate that the interactions xenon entertains with protein moieties are predominantly of dispersive nature. This suggests xenon interaction can be modelled based on its van der Waals interactions, using a Lennard-Jones potential at a molecular mechanics level of theory. Indeed, xenon van der Waals parameters have been determined experimentally (Verlet and Weis, 1972) and were successfully used in simulation studies investigating xenon

interacting with myoglobin (Cohen et al., 2006) and the NMDA receptor (Liu et al., 2010). It is tempting to use a simple Lennard-Jones potential to score xenon binding propensity as in doing so, the xenon scoring becomes compatible with molecular mechanics approaches that can be used to sample protein plasticity and solvation in a rigorous way (see below; section 3.3.9). To investigate the extent to which this is possible, a classifier was constructed based on the Lennard-Jones potential between xenon and the CHARMM22 protein atom types. In the case of protein atoms, van der Waals parameters were takes from the standard CHARMM22 force field (MacKerell et al., 1998) (see Table 3.2) that assumes that assumes the following functional form of the Lennard-Jones potential,

$$V(r_{ij}) = \epsilon_{ij}\left(\left(\frac{R_{\mathrm{min},ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\mathrm{min},ij}}{r_{ij}}\right)^{6}\right)$$

with r_{ij} the interatomic distance of two nuclei i, j, $\epsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ the energy well depth, and $R_{\mathrm{min},ij} = R_{\mathrm{min},i}/2 + R_{\mathrm{min},j}/2$ the equilibrium distance. The overall interaction energy for a given xenon atom i is then obtained by summing over all nuclei within the periodic arrangements of asymmetric unit copies, $V_i = \sum_{j \notin \mathrm{xenon}} V(r_{ij})$. It should be noted that the functional form of $V(r_{ij})$ implies that the function returns to zero rapidly for values of r_{ij} not much larger than the equilibrium distance, and therefore, no distance cutoff needs to be defined other than for speeding up the calculation.

Classification of all xenon and non-xenon atoms contained in the full data set described earlier in this chapter was then performed based on this empirical scoring function, and evaluated using the familiar performance measures. Cross-validation did not have to be performed since the empirical scoring function was not based on part of the test data. Using $\varepsilon_{\mathrm{XE}} = -0.494$ kcal mol^{-1} and $R_{\mathrm{min,XE}}/2 = 2.24$ Å,[27] overall a good discrimination power was obtained. The area under the curve was found to be 0.91, the maximum of Youden's index 0.70, and the minimum distance to the point of perfect classification, 0.23. A maximum of Matthew's Correlation coefficient was found for an energy cutoff of -5.2 kcal mol^{-1} (MCC of 0.52).[28] Alternative cutoffs were -3.3 kcal mol^{-1}, maximising Youden's index, and -2.2

[27] as used in an MD simulation study described previously (Cohen et al., 2006).
[28] predicting xenon if the energy was *lower* than that cutoff.

kcal mol^{-1}, minimising the distance to the point of perfect classification.[29]

The performance measures indicate the empirical scoring function to perform inferior to the knowledge-based scoring function derived throughout this chapter for which an AUC of 0.98 and a maximum MCC of 0.72 was obtained. However, importantly, the empirical scoring is fully compatible with molecular mechanics approaches (see below).

It should be noted that as xenon carries a zero net charge, electrostatic interactions using a Coulomb potential,[30] which assumes a point charge model, do not contribute to force field terms involving xenon. It is tempting to speculate that the potentially important omission of electrostatic interactions of xenon that can result from its and polarisability and the phenomenon of induced dipoles, is at least in part implicitly captured by the knowledge-based approach developed throughout this chapter. The knowledge-based approach is not limited by the simple functional form of the empirical (force field) scoring function, and this fact could explain its better performance in the analysis conducted in this section.

3.3.8 The method is limited by the sampling of xenon positions

Up to this point, the ability of the xenon likeness score to discriminate xenon atoms from non-xenon atoms, given a coordinate in three dimensional space, has been discussed. In order to validate and benchmark the method, known positions of water molecules, solvent ions and xenon atoms have been scored, and it has been demonstrated that the method is able to successfully discriminate between the former and the latter (Results sections 3.2.4 and 3.2.5). However, a prospective application of the method might aim on detecting novel xenon binding sites in a protein. A process therefore needs to be devised to suggest relevant positions in three dimensional space to the scoring method; based on the score of the respective

[29] to evaluate the classification, the same parameters were used as described earlier in this chapter. Using alternative van der Waals parameters for xenon, $\varepsilon_{\mathrm{XE}} = -0.433$ kcal mol^{-1} and $R_{\mathrm{min,XE}}/2 = 2.21$ Å, following (Liu et al., 2010) who cite early experimental work (Verlet and Weis, 1972), very similar performance measures were obtained: AUC of 0.91, maximum MCC 0.53 (threshold -5.0 kcal mol^{-1}), maximum Youden's index 0.71 (threshold -2.9 kcal mol^{-1}), and minimum distance to point of perfect classification 0.22 (threshold -2.2 kcal mol^{-1}).

[30] the Coulomb potential has the form $V = \frac{q_i q_j}{\epsilon_{ij} r_{ij}}$ with q the atom point charges.

position, potential xenon binding sites could then be detected.

This is reminiscent of the methodology of *molecular docking*, which aims to derive the structure of a ligand-protein-complex, starting from the *apo* structure of the free protein (see (Meng et al., 2011) for a review). The underlying problem can be divided into two sub problems, namely generating binding poses by sampling possible conformations of the ligand and its interactions with the protein, and secondly, ranking the resulting poses by a quantifying score to tell likely from unlikely poses, and ultimately good from bad ligands. The first problem, pose generation, is widely regarded as easier than scoring, or correlating molecular docking scores to binding affinities (Kitchen et al., 2004). For a ligand to favourably interact with a protein, a significant amount of its polar surface area must interact with the protein. Therefore, typically a ligand will prefer cavities within the protein surface displaying complementary interaction points to its pharmacophore. It is relatively uncommon that a ligand interacts with non-buried, hydrophobic surface moieties of proteins, and to date, identification of those potential protein-protein interaction modulators by structural computational methods is a challenging field of ongoing research (Arkin and Wells, 2004; Yin and Hamilton, 2005; Domling, 2008; Geppert et al., 2012). Therefore, in a typical application of molecular docking, search space for pose generation is limited by confining it to protein cavities, or even the known catalytic centre of the protein. It is conceivable that xenon atoms do not require large cavities as drug like molecules do, which increases the search space for the prediction of xenon binding sites. However, xenon is isotropic and can thus be represented as a rigid sphere with a defined radius to good approximation, thus it is not necessary to generate a potentially high number of possible internal ligand conformations as required for molecular docking, thereby decreasing search space; the pose generation process is effectively reduced to sampling possible xenon positions.

Given the similar size of xenon atoms and water molecules (Figure 3.16), and the general abundance of water molecules in high-resolution protein structures, experimentally derived water positions could be used as surrogates for potential xenon atom positions. However the assumption that water molecules probe the surface of the protein sufficiently to suggest relevant positions for scoring is problematic, given the apparent differential preferences of xenon and water environ-

ments[31] that is also suggested by the aversion of xenon atoms to directly interact with water molecules (Figure 3.15). On the other hand, it is possible that xenon atoms might replace a water molecule weakly bound to a moiety that would actually favour the binding of xenon.

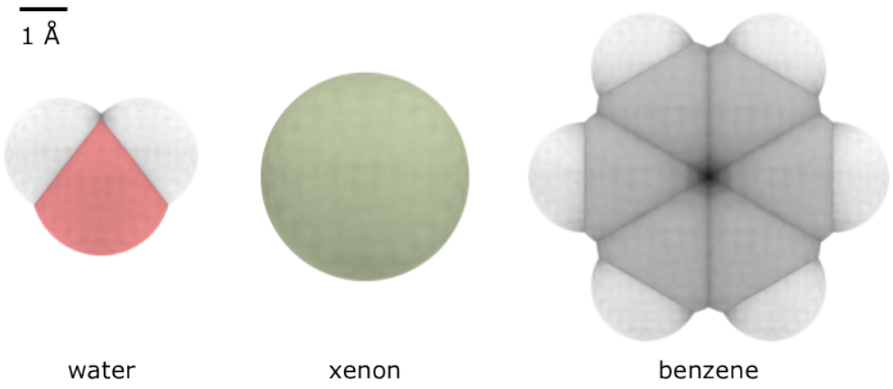


Figure 3.16 – Comparison of the size of xenon atoms to water molecules and benzene. Atom van der Waals radii, bond lengths and angles for water and benzene were taken from the CHARMM22 force field, the TIP3P model of water (Jorgensen et al., 1983), and a xenon radius of 2.2 Å was assumed.

It is conceivable that the more water/ion positions which are used as potential xenon positions, the higher the likelihood of correctly identifying a true xenon position. A rule of thumb in crystallography states that one water molecule per residue is detectable at an experimental resolution of 2.0 Å. This has been demonstrated to approximately hold true experimentally (Carugo and Bordo, 1999), and values of about 0.8 and 1.6 to 1.7 water molecules per residue have been suggested at 2.0 Å and 1.0 Å resolution, respectively. The experimental resolution alone was found to be a good predictor of the number of water molecules detected, other predictors including the fraction of crystal volume occupied by the solvent, the number of residues in the asymmetric unit, and the fraction of apolar protein surface area or secondary structure.

An analysis of 2,771 protein chains sharing their UniProt identifier with one of the 64 xenon binding proteins indicated a clear resolution dependence of the number of water molecules (Pearson correlation coefficient of $R = -0.622$), with

[31] which is a prerequisite for the performance of the classifier devised in this chapter.

median values of about 1.55 and about 0.73 at resolutions of 1.0 and 2.0 Å, respectively, only slightly lower than the values suggested in the study by Carugo and Bordo.

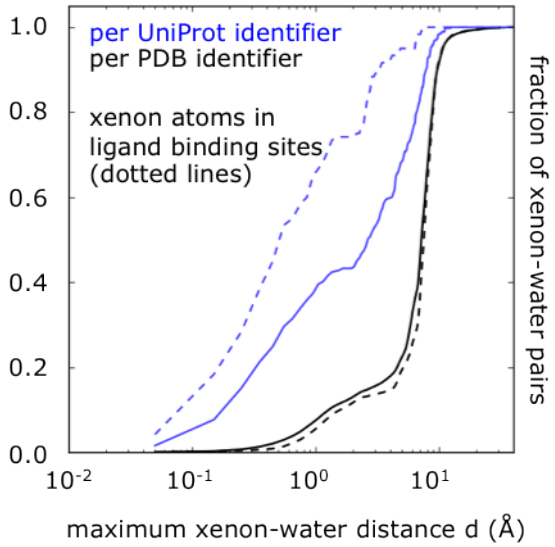


Figure 3.17 – Distances of the closest water molecule to any xenon atom in pairs of super-imposed protein structures. Data were aggregated per UniProt identifier (blue) or Protein Data Bank (PDB) identifier (black). The entire data set consisted of 2,771 protein chains with the same UniProt identifier as one of the 64 xenon binding protein chains. Solid lines indicate the entire set of xenon atoms (458 xenon atoms bound to a protein chain that has a UniProt identifier with multiple known structures in the PDB), while dotted lines correspond to the 120 xenon atoms found in ligand binding sites, according to an analysis carried out earlier in this chapter (see Results section 3.2.2). A bin width of 0.1 Å was used to group data.

It remains to be demonstrated whether water molecules are often enough found close to the true xenon positions, and that the value of the xenon likeness score for a surrogate water molecule is similar to that of the actual xenon to be detected. To address this question, firstly, for the data set of the 2,771 protein chains sharing their UniProt identifier with one of the 64 xenon binding proteins, water coordinates were retrieved for the water molecule closest to each of the xenon atoms, after super-imposing all protein structures with matching UniProt identifiers. Distances between those closest water molecules and the respective xenon atoms were then determined (Figure 3.17). For data aggregated per UniProt identifier[32] it was

[32]that is, distances minimised globally across all PDB structures for a particular UniProt

found that for 27 % of xenon atoms there was a water molecule found within 0.5 Å, and for 44 % of xenon atoms in less than 2.0 Å. If only those xenon atoms which were found to overlap with known ligand molecules were considered, this number increased to 40 % and 74 %, respectively.[33] This number dropped considerably when individual PDB entries were considered, i.e. without, for a given xenon atom, retrieving only the one water molecule closest to that xenon atom across multiple PDB structures with the same UniProt identifier (Figure 3.17). This indicates that if there are multiple structures known for a given system identified by a common UniProt identifier, there is a higher likelihood that water molecules can be found close to the xenon position to be detected. For systems of particular interest, i.e. xenon atoms found in ligand binding sites, this likelihood is even higher.

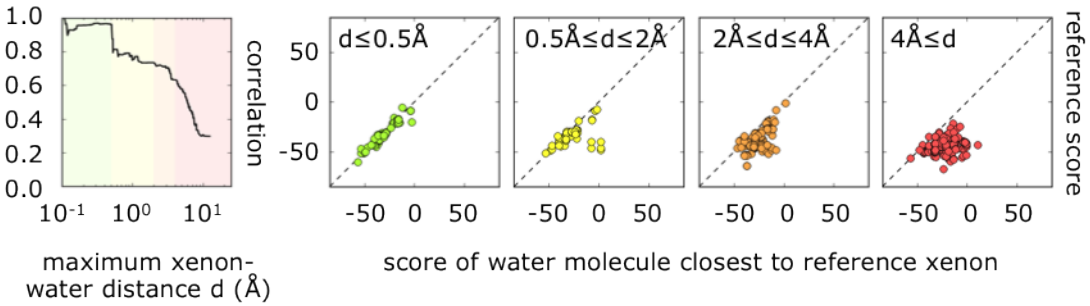


Figure 3.18 – Correlation of the xenon likeness score of xenon atoms and their closest water molecules, per Protein Data Bank record. Leftmost panel, Pearson correlation coefficient of xenon likeness scores as a function of their maximum inter-nuclear distance. Pairs of xenon atoms and water molecules were sorted by increasing inter-nuclear distance, and for N pairs, Pearson correlation coefficients R_i are plotted at distance d_i for all pairs up to that distance, with $i = 1..N$. The distance range up to 0.5 Å is shaded green and corresponds to the second panel where a scatter plot of the xenon likeness is shown, with the score for the water molecule along the x-axes and the score of the xenon atom along the y-axis for reference. Other distance ranges are shaded yellow (0.5 Å to 2 Å, third panel), orange (2 Å to 4 Å, fourth panel), and red (> 4 Å, fifth panel).

Next, it was investigated how the xenon likeness scores correlate between the actual positions of xenon atoms and the positions of the surrogate water molecules.

identifier.

[33] assuming the same criteria to determine whether or not a xenon atom was located in a ligand binding site as in Results section 3.2.2.

To this end, again xenon atoms were assigned their closest water molecule, for each PDB record (Figure 3.18), or across all PDB records with the same UniProt identifier (Figure 3.19). The data suggest that for inter-nuclear distances of up to 0.5 Å or 2.0 Å, the xenon likeness scores for water molecules correspond well to those of the xenon atoms they are meant to serve as a surrogate for (as measured by the value of the Pearson correlation coefficient of > 0.80, and their score being comparable to, or lower than, the discrimination score thresholds suggested in Results section 3.2.5). For water molecules further separated from xenon than 2.0 Å the score correlation quickly declines. In general, scores of surrogate waters are higher (less favourable) than those of xenon atoms. A distance cutoff of 2.0 Å on useful xenon likeness scores is well within the 3.2 Å that where reported for the average distance that probe molecules travel from their initial positions towards local energy minima in a computational 'solvent scanning' study (Dennis et al., 2002), aiming at the computational localisation of 'hot spot' regions on the protein surface that are able to bind small molecule ligands (see also Introduction, sections 1.1 and 1.1.3).

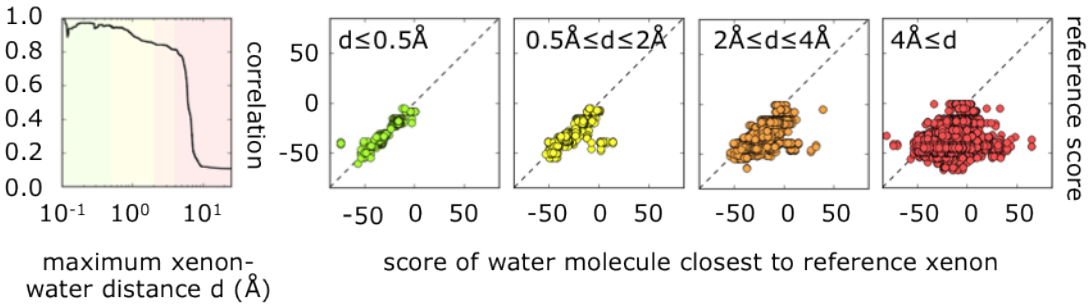


Figure 3.19 – Correlation of the xenon likeness score of xenon atoms and their closest water molecules, per UniProt identifier. Same as Figure 3.18, only that for each xenon atom, closest water molecules are determined globally per UniProt identifier, instead of for each Protein Data Bank record.

3.3.9 A grid based sampling approach might outperform the sampling of water positions

Overall, this distance dependent score correlation analysis suggests that water positions seem to be a reasonable estimate for potential xenon atom positions in most relevant cases, specifically excluding cases where xenon is found to be present buried deep within the protein core where water molecules are unlikely to occur.

It should be noted that in using the positions of water molecules as surrogates for possible xenon atom positions, sensitivity is sacrificed in order to limit the search space, but the specificity of the method is not impaired. The lack of sensitivity however can only be accounted for by investigating more potential xenon binding sites. This could be achieved by employing a grid based approach, where the protein of interest is embedded in an artificial three dimensional grid.[34] Each grid point could be examined and scored by the xenon likeness score, aiming to detect low scoring grid points as potential xenon binding sites. Grid points overlapping with protein atoms[35] and those too far away from the protein[36] could be eliminated in order to limit search space. A regularly spaced grid could be used with a spacing of around 0.5 Å. In the distance dependent score correlation analysis (Figures 3.18 and 3.19), this was found to be a distance range where score for water molecules corresponded well to those of the xenon atoms to be detected.

The time necessary to evaluate the entire grid will be directly proportional to the number of grid points, and inversely proportional to the third power of the grid spacing distance,[37] which could be used as an adjustable parameter to trade sensitivity for speed of evaluation. Regardless the grid topology, a grid based approach is never rotation and translation invariant, so some care would have to be taken for the approach not to suffer from artifacts induced by this, e.g. the unintended omission of grid points because the arbitrary orientation of the grid would make them overlap (clash) with protein atoms. Possible ways of avoiding this would be to conduct the analysis with multiple different grid orientations, or to

[34] In this context, xenon could be considered the experimental equivalent to a hydrophobic probe in the computational GRID approach (Goodford, 1985).

[35] i.e. being separated by less than the sum of the van der Waals radii of xenon and that protein atom type, minus a certain distance to account for some 'fuzziness' of the approach.

[36] xenon atoms do not assume stationary positions without dispersive contact to protein atoms.

[37] if a cubic grid topology is used.

sufficiently decrease grid spacing. If the search space is considered to be a volume slice of fixed width around a roughly elliptical shape (the protein), it approximately scales with the second power of the surface area which is proportional to the the molecular mass of the protein, assuming a constant average protein density. This is a slight improvement over the number of grid points increasing with the third power of protein/unit cell dimensions, but can still lead to a number of grid points to be investigated at the order of magnitude of 10^6 or 10^7 grid points for large systems and fine grid spacing. This limitation could potentially be addressed by first using a coarse-grained approach, i.e. a larger grid spacing with fewer points to be evaluated, and then re-sampling interesting regions in a more fine-grained way. This adaptive sampling, however, might be prone to sensitivity problems if well scoring positions are missed in the first, coarse iteration; yet given the results of the correlation analysis between xenon likeness scores obtained for xenon atoms and that for their closest water molecules found in related complexes presented above (Figures 3.18 and 3.19) initially missing the exact position of a xenon atom by about 1 to 2 Å seems to be affordable.

Furthermore, with a grid based approach discerning *positive* from *negative* samples will be complicated, since, given a fine grid spacing, a single xenon atom would be represented by a group of neighbouring, low scoring grid points. To reduce the number of positive predictions to a realistic number,[38] a post processing of results could be conducted, e.g. clustering them. A similar approach has been used recently in related work by Zheng et al. where discrete protein solvation sites are predicted by a grid based method (Zheng et al., 2012). Alternatively, a flood fill algorithm could be used (Weisel et al., 2007) in order to obtain discrete, connected sets of low scoring grid points. However, this approach might produce unrealistically large clusters, not physically corresponding to individual xenon atoms, and a representative point in the cluster might have to be chosen as prediction of the position of the xenon atom. In any case, validation of the method would need to be carried out defining a success criterion such as, regarding a prediction as a *true positive* if it was not further separated from the true xenon position than a certain distance threshold.

[38] clearly, a number of 64 low scoring grid points arranged in a cube of 2 Å edge length with a grid spacing of 0.5 Å will not correspond to 64 discrete predictions.

Alternatively, a simulation-based approach could be envisaged: given the structure of the protein unit cell, and applying periodic boundary conditions,[39] a molecular mechanics approach could be employed to sample possible xenon binding sites. By using the xenon likeness score as a potential energy function, the trajectory of synthetic xenon atoms could be computed as it moves through the simulated system. This could be either achieved by a Monte Carlo approach, where new positions for xenon atoms would be sampled from a random distribution, and accepted or rejected based on the energy difference between the current and suggested state; or as an alternative, a Molecular Dynamics approach could be employed computing molecular replacements by integrating the equations of motion in order to obtain forces displacing that probe particle; this, however, would require the potential energy function to be differentiable. In the present case, the potential energy function is a sum of empirically determined radial distribution functions (rdfs) (see Equations 3.6 and 3.7) which are not differentiable analytically. This problem could be circumvented by approximating the individual rdfs by differentiable functions parametrised to closely resemble their functional form, containing a defined minimum at fixed inter-nuclear distance, and a tail converging to zero, or numerical integration. A possible choice of function would be a Lennard-Jones potential, which is continuously differentiable. Simulations could be carried out at different levels of solvation, i.e. using a continuous solvent model, or explicit solvent molecules competing with xenon for potential binding sites at the protein, and simulation parameters could be used to represent pressure conditions of the actual xenon soaking experiment. After the experiment, data could be aggregated to determine the probability of finding xenon at a preferred position; which, following a Boltzmann approach would be related to its interaction energy, thereby closing the line of argument of the derivation of the knowledge-based potential.

A similar approach, related to the widely used GRID methodology (Goodford, 1985), has been employed recently for the computational fragment-based binding site identification by the SILCS (site identification by ligand competitive saturation) method (Guvench and MacKerell, 2009), where atomic protein models are computationally immersed in æqueous solutions containing different probe

[39] to avoid simulation boundary effects, and in order to take into account possible interactions with image atoms.

molecules, i.e. benzene, propane, and water, representing the most important interaction types of drug-like molecules (see also Introduction section 1.1.3). From the frequency these probe molecule types are found at specific locations on the protein surface during several short MD simulations using the CHARMM force field (MacKerell et al., 1998), probability maps for probe binding were constructed, rigorously incorporating solvation and protein plasticity at nanosecond timescales. In using xenon instead of benzene or propane molecules, convergence of the approach could likely be achieved more easily since due to the isotropic shape of xenon, internal degrees of freedom need not be sampled.

The challenging aspect of applying these approaches to the prediction of xenon binding sites would likely be the reconciliation of the empirica force field terms (for sampling protein conformations) with the knowledge-based xenon likeness score, i.e. the sampling of xenon positions. In particular, there is a conceptual difference between the knowledge-based *potential of mean force* (Hendlich et al., 1990; Koppensteiner and Sippl, 1998) terms that can be computed from radial distribution functions (Leach, 2001) and are related to the *binding free energy*, and the *enthalpic*, interaction energy computed from a molecular force field.

The performance investigation of an empirical xenon scoring described above (section 3.3.7) has revealed it to be inferior to the knowledge-based scoring, yet still encouragingly powerful. Having been used successfully in MD simulations before (Cohen et al., 2006; Liu et al., 2010), the empirical scoring function might still be sufficiently accurate to be used during the simulation itself, in conjunction with re-scoring performed with the knowledge-based approach.

3.3.10 Outlook

In future work, some of these alternative approaches to detect potential xenon binding sites are to be followed up on. The implementation of a grid based approach conceptually is the more straightforward option, and could be regarded to as an exhaustive search method. The simulation-based approach, however, would allow for a direct inclusion of protein flexibility, and solvation effects (and therefore, entropic contributions to the binding event).

3.4 Methods

3.4.1 Retrieval and processing of protein structures

All protein structure files discussed in this chapter were retrieved from the Protein Data Bank (PDB) (Berman et al., 2000) in April 2012. Amino-acid sequences were extracted from the respective structure itself rather than using the sequence information provided in the FASTA[40] file associated with every protein record using a custom Python script; residues were checked for completeness of their backbone, that is, presence of coordinates for the amide-nitrogen (N), α-carbon (C^{α}), carboxy-carbon (C^{O}) and -oxygen (O) atoms, as well as connectivity with adjacent residues, requiring an inter-nuclear distance of ≤ 1.4 Å between N_i and C^{O}_{i-1} for residues $i, i-1$. Only complete residues were included in the amino acid sequence to be further subjected to sequence alignments; others were treated as gaps in the amino acid sequence. Groups of hetero-atoms (solvent, ions, single atoms, small organic ligands) were assigned to the protein chain they were closest to such that the pairwise distance between one atom in the group of hetero-atoms and one atom in the assigned protein chain was the minimum over all such pairs. For residues with multiple conformations the conformation with the highest population according to the *occupancy* column in the PDB file was chosen to represent the protein structure. In case of ties between multiple conformations, the conformation listed first was chosen arbitrarily.

3.4.2 Mappings between Protein Data Bank and UniProt

Mappings between the Protein Data Bank (PDB) (Berman et al., 2000) and UniProt (UniProt Consortium, 2012) are provided by a RESTful (Fielding and Taylor, 2002) web service based on sequence-alignment derived SIFTS mappings (Velankar et al., 2005), following a DAS (Prlic et al., 2007) protocol. The service is available at `http://www.pdb.org/pdb/rest/das/pdb_uniprot_mapping/alignment?query=[PDBID]`, returning an XML (Extensible Markup Language) file representing a *many-to-many* mapping; each protein chain present in the PDB record is assigned its corresponding UniProt identifier.

[40] (Pearson and Lipman, 1988).

The inverse was achieved by directly querying the PDB with UniProt identifiers, yielding all PDB records where the query corresponds to at least one protein chain. An additional search filter limited the search only to structures solved by X-ray crystallography.

3.4.3 Protein sequence comparison

Pairwise protein sequence alignments were performed using the *needle* program of the EMBOSS suite of programs (Rice et al., 2000), implementing the Needleman-Wunsch algorithm for global pairwise sequence alignment (Needleman and Wunsch, 1970). To score alignments, the BLOSUM62 (Henikoff and Henikoff, 1992) amino acid substitution matrix was used with gap open and -extension penalties of 10.0 and 0.5, respectively. Sequence identities were then calculated as the percentage of identical residues over the length of the resulting alignment, which can be longer than either of the input sequences.

Multiple protein sequence alignments (see next chapter) were approximated using ClustalW2 (Larkin et al., 2007) with parameters consistent to those used for the pairwise alignment above, a mutual gap distance of 5.0, and building a single *Neighbour Joining* (Saitou and Nei, 1987) guide tree.

Hierarchical protein sequence clustering was performed as follows: firstly, sequences were aligned *all-to-all* in a pairwise fashion, as described above. The resulting identity matrix was then converted to a distance matrix (with the percentages similarity and -distance adding up to 100) which served as input for *complete linkage* clustering performed by the *hclust* routine in R (RDe). Dendrograms and heatmaps were drawn using the *Heatplus* package (Plo) in R.

3.4.4 Calculation of xenon solvent accessible surface area

After adding hydrogen atoms to the protein structures using the REDUCE program (Word et al., 1999), the solvent accessible surface areas (SASA) of xenon atoms were calculated using the naccess program (Hubbard and Thornton, 1993), assuming *van der Waals* radii of $r_{vdw,Xe} = 2.2$ Å for xenon (Cohen et al., 2006; Liu et al., 2010) and a probe radius of $r_{probe} = 1.4$ Å for water. Values were reported normalised to a maximum SASA of $A = 4\pi \left(r_{vdw,Xe} + r_{probe}\right)^2 \approx 159.26$ Å^2

corresponding to a fully solvated xenon atom.

3.4.5 Crystal unit cell expansion

Crystal packing of protein crystals in the vicinity of an isolated asymmetric unit was reproduced by a crystal symmetry expansion using the XPAND program (Kle). Unit cell dimensions and crystal space groups as reported in the respective PDB file headers were used to generate crystal images of protein chains around a point of expansion, in this case, xenon atoms, within a user-defined cutoff, which was set to 20 Å, and in a spherical manner.

3.4.6 Protein superimposition

Pairs of individual protein chains were superimposed by minimising the *root mean square deviation* (RMSD) of their C^{α} atoms using the *align*-command of the pymol program (DeL). The option to progressively refine the superimposition by eliminating highly divergent regions of the proteins was not used as it was assumed that since both proteins have the same UniProt identifier they would have almost identical amino acid sequences and hence structures, especially close to the ligand binding site, which is most important for the analysis presented here.

3.4.7 Conversion and atom typing of ligand structures

Ligands were extracted from PDB files and converted into the *mol2* format using the Open Babel suite of programs (O'Boyle et al., 2011). During this conversion, atoms are assigned SYBYL (Clark et al., 1989) atom types according to their connectivity and bond order of the covalent interactions in which they partake. Table 3.5 shows a selection of different SYBL atom type symbols and their definition.

3.4.8 Protein atom types

Non-hydrogen protein atoms were assigned CHARMM22 (MacKerell et al., 1998) atom types depending on the atom name and amino acid name found in the PDB file, provided by the standard CHARMM22 *topology* file. Histidine residues were

Table 3.5 – SYBYL ligand atom types

C.3	sp^3 carbon
C.2	sp^2 carbon
C.1	sp carbon
C.ar	aromatic carbon
N.3	sp^3 nitrogen
N.2	sp^2 nitrogen
N.1	sp nitrogen
N.ar	aromatic nitrogen
N.4	sp^3 positively charged nitrogen
N.pl3	trigonal planar nitrogen
N.am	amide nitrogen
O.3	sp^3 oxygen
O.2	sp^2 oxygen
S.3	sp^3 sulphur
S.2	sp^2 sulphur
F	fluorine
Cl	chlorine
Br	bromine
I	iodine
Se	selenium
B	boron

arbitrarily assigned the residue type HSE, assuming a neutral charge and default tautomer with the proton located at the nitrogen denoted by NE2. The 25 non-hydrogen atom types used in this work are listed in Table 3.6.

3.4.9 ROC analysis and performance measures

The Receiver operating characteristic (ROC) curve is a way of describing the performance of a *binary classifier*. As a score threshold t is varied, for a given threshold value t_0 a pair of values consisting of *1-specificity* and *sensitivity* can be calculated from ratios of *true positive* (TP), *true negative* (TN), *false positive* (FP), and *false negative* (FN) predictions at that threshold value, and plotted graphically. The greater the area under the curve (AUC) described by all these pairs of values, the better the overall performance of the classifier. AUC is thus a common measure to compare the relative performance of different prediction methods.

Sensitivity and specificity can be calculated as $sensitivity = TP/(TP + FN)$ and $specificity = TN/(TN + FP)$, respectively. Values for both measures lie

Table 3.6 – CHARMM non-hydrogen protein atom types. Names of amino acids are abbreviated by standard three letter codes.

C	carboxyl carbon
CT1	carbon with one hydrogen, all C^{α} carbons except Gly, also C^{β} of Ile and Val
CT2	carbon with two hydrogens, majority of aliphatic side chain carbons (C^{α}, C^{β}, C^{γ}, C^{δ}, C^{ϵ})
CT3	carbon with three hydrogens, i.e. methyl carbon
CA	aromatic carbon
CC	acidic carbon of Asn/Gln or Asp/Glu
CPH1	aromatic carbon C^{γ} of the neutral His tautomer
CPH2	aromatic carbon C^{ϵ_1} of the neutral His tautomer
CP1	C^{α} carbon of Pro
CP2	C^{β} or C^{γ} carbon of Pro
CP3	C^{δ} carbon of Pro
CY	aromatic carbon C^{γ} of Trp
CPT	aromatic carbon between the 5- and 6-membered heterocycles of Trp
NH1	amide nitrogen of all amino acid types except Pro
N	backbone nitrogen of Pro
NH2	sp^2 nitrogen of Asn/Gln
NH3	positively charged sp^3 nitrogen of Lys
NC2	guanidine nitrogen of Arg
NR1	protonated aromatic nitrogen N^{ϵ_2} of the neutral His tautomer
NR2	aromatic carbon N^{δ_1} of the neutral His tautomer
NY	aromatic nitrogen in the 5-membered heterocycle of Trp
O	carboxyl oxygen
OH1	alcohol function of Ser, Thr and Tyr
OC	acidic oxygen of Asn/Gln or Asp/Glu
S	sulphur of Cys and Met

between 0 and an optimum value of 1. Sensitivity and specificity are usually inversely correlated, and a ROC analysis investigates their mutual dependence and trade-off; a perfect classification would require both to assume a value of 1. In this work, ROC curves were approximated by 1,000 values of t_0 equally distributed between the minimum and maximum score found for the given system. The AUC has a maximum of 1 and can be found by numerical integration of the ROC curve; in this work, a step size of 10^{-3} was used for the numerical integration.

Matthews Correlation Coefficient (MCC) was calculated as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

and assumes values between -1 and $+1$, with $+1$ perfect prediction, 0 no better than random prediction, and -1 total disagreement between prediction and observation. Youden's index Y is computed as $Y = sensitivity + specificity - 1$ and has an optimum value of 1, describing the vertical distance from the diagonal line described by the equation $1 - specificity = sensitivity$, a classifiction not better than random. Additionally, the Euclidean distance of the pair of *specificity/sensitivity* values closest to the point of perfect classification (where $sensitivity = specificity = 1$ may be reported.
In this work, some additional performance measures are used; *positive coverage* is synonymous with *sensitivity*; *positive accuracy* is synonymous with *precision* and can be calculated as $precision = TP/(TP + FP)$. *Negative coverage* is synonymous to *specificity*, and *negative accuracy* is calculated as *negative accuracy* $= TN/(TN + FN)$.

Positive and negative accuracy are also sometimes referred to as *positive predictive value* (PPV) and *negative predictive value* (NPV), respectively.

3.4.10 Pearson correlation coefficient

The Pearson correlation coefficient between two vectors x, y was calculated as

$$R = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}},$$

with μ the sample average and n the number of samples in each vector, and gives values between -1 and $+1$ (for perfect (anti-)correlation).

3.4.11 Approximation of the error function

The following equation provides a closed form of the error function $erf(z) = \frac{2}{\pi}\int_{t=0}^{z} e^{-t^2}dt$ which is accurate to 7 significant digits for non-negative z,

$$\begin{aligned} erf(z) \approx 1 - t \times \exp(&-z^2 - 1.26551223 + 1.00002368t \\ &+0.37409196t^2 + 0.09678418t^3 \\ &-0.18628806t^4 + 0.27886807t^5 \\ &-1.13520398t^6 + 1.48851587t^7 \\ &-0.82215223t^8 + 0.17087277t^9). \end{aligned}$$

where $t = 1/\left(1 + \frac{1}{2}z\right)$ (Numerical Recipes, Cambridge University Press, 3rd edition, chapter 6.2.2, p. 264f, http://www.nr.com).

3.5 Summary

In this chapter a knowledge-based method to detect xenon binding sites in proteins has been developed and validated. Based on known structures of xenon protein complexes from the PDB, xenon environments have been described using radial distribution functions (rdfs) and CHARMM protein atom types. It has been found that xenon atoms often bind to ligand binding sites and favour a hydrophobic environment, e.g. aliphatic and aromatic carbon rich groups, and sulphurs, and that they preferentially establish direct, dispersive van der Waals contacts with these protein atoms. The rdfs were then combined into a cumulative *xenon likeness score* that can be used to discriminate xenon binding sites from sites that do not bind xenon. This scoring function, together with a technique to generate positions in three dimensional space to be scored, constitutes a method to predict xenon binding sites.

The validation of the scoring function was performed by investigating its ability to discriminate true xenon atom positions (positives) from positions within pro-

tein structures where no xenon atoms but instead water molecules or buffer ions were found (negatives). The method performs remarkably well in a ROC analysis, indicating that true xenon binding sites can be discriminated from false ones with high discrimination power. Possibly owing to the high number of negative over positive samples in the retrospective validation, and a certain number of negative samples that could indeed be positive, scoring *sensitivity* is easier to achieve than scoring *precision*; therefore, three differentially conservative score thresholds were suggested that allow the methodology to be fine-tuned based on the intended application. Taking into account the unbalanced learning data set and employing balanced performance measures, however, the scoring function was found to possess very high discrimination power. A probabilistic interpretation of the scores furthermore allows for their more intuitive and fuzzy, rather than absolute, interpretation. The knowledge-based xenon likeness score was found to outperform a simple classifier based on the count of close protein atoms, and an electrostatic contact potential in a direct performance comparison. It was suggested, however, that the empirical, force field based contact potential might be used in conjunction with the knowledge-based scoring function in a molecular mechanics simulation approach.

For the method to be useful for *prospective* application, i.e. detecting xenon binding sites in unknown proteins, the scoring method has to be supplemented with means to suggest potential xenon positions that are then scored and evaluated. The suitability of using the positions of water molecules in related protein structures as surrogates for potential xenon binding sites has been investigated, and it was found that in cases where many experimental structures are available, water molecules typically sample potential xenon binding sites sufficiently well. However, the true sensitivity of the method in prospective applications is difficult to estimate as it will be greatly influenced by the quality of the test position generator. It is anticipated that the utilisation of grid based approaches, and possibly adaptive sampling techniques, together with a clustering of the prediction results, as well as molecular mechanics based sampling approaches, are likely to outperform the water position based sampling with respect to prediction sensitivity, and in case of the latter, would also allow for the rigorous incorporation of protein plasticity. This will be investigated in future work. More directly, the methodology will be

applied to prospectively predict xenon binding to the N-terminal domain of human Hsp90-α, and the predictions subjected to validation by solving the structures of the protein in presence and absence of xenon by X-ray crystallography.

Chapter 4

Investigation of Xenon Interacting with the Hsp90-NTD Protein

4.1 Introduction

Heat shock protein of 90 kDa (Hsp90) has emerged as a promising target for pharmaceutical cancer intervention in recent years. As a molecular chaperone, it assists folding and maturation of a plethora of 'client' proteins, many of them kinases involved in cancer pathogenesis. A more detailed review of the role of Hsp90 in cancer, and molecular characteristics of the protein, can be found in the general Introduction section 1.3. Because of its pharmaceutical relevance, easy handling and generous structural knowledge about the protein, Hsp90 is an ideal model system for the prospective application of the xenon-binding prediction methodology developed in Chapter 3. In drug discovery campaigns for Hsp90, *fragment based drug design* approaches (see Introduction section 1.1) have played a central role. The previous chapter has highlighted, along with its importance in crystallographic phasing (see section 3.1.1), the possible aptitude of xenon as a fragment-like entity, a notion that will be tested thoroughly in the following. Several other concepts relevant for this chapter have been introduced in previous sections of this thesis and will only be referred to briefly to avoid redundancy. During the structure determination effort conducted in this chapter, a molecular replacement strategy (section 3.1.2) will be used to solve the crystallographic phase

problem (section 3.1.1), and the anomalous scattering properties of xenon (section 3.1.2) will be used as an independent verification of the identity of xenon binding to protein moieties.

4.2 Results

In this chapter, the interaction of xenon with the N-terminal domain (NTD) of the human heat shock protein Hsp90-α (Hsp90, UniProt identifier P07900) is investigated. Firstly, the conformational space spanned by the protein is investigated on the basis of known experimental protein structures in the Protein Data Bank. Secondly, the xenon likeness score, developed in the previous chapter, is used to predict possible xenon binding sites in the protein, and thirdly, these predictions are tested by solving the structure of Hsp90-NTD in the presence of xenon by X-ray crystallography.

4.2.1 Structures of Hsp90-NTD in the Protein Data Bank

Experimental protein strutures of Hsp90-NTD solved by X-ray crystallography were retrieved from the Protein Data Bank (PDB) (Berman et al., 2000). There are 134 structures of the N-terminal domain (NTD) of Hsp90-NTD solved by X-ray crystallography, with 162 individual protein chains, excluding structures that contain missing residues.[1]

The vast majority of crystals of Hsp90-NTD falls into orthorhombic[2] space groups (106 out of 138), with most of them belonging to the I 2 2 2 space group ($N = 87$, with mean unit cell dimensions $\pm$ standard deviation, all of which in Å: $a = 66.29 \pm 0.89$, $b = 89.87 \pm 1.07$, $c = 99.08 \pm 0.72$). Other orthorhombic space groups are P 2_12_12 (N=9), P $2_12_12_1$ (N=7), and C 2 2 2_1 (N=3). The other space groups are P 1 $2_1$1 (N=25), C 1 2 1 (N=2, both triclinic), P 1 (N=4, monoclinic), and P 4_32_12 (N=1, tetragonal).[3]

[1] as of October 2012. Human proteins only. Note that some experimental structures are multimeric.

[2] $\alpha = \beta = \gamma = 90°$.

[3] space group names are given in Herman-Mauguin notation with the first letter describing the centering of the Bravais lattice (for the space groups found in this analysis: 'P', primitive,

The consensus amino acid (AA) sequence of the core alignment (AA 18 to 222) is

```
>UniProt P07900 18-222
  1                    ETF AFQAEIAQLM
 31 SLIINTFYSN KEIFLRELIS NSSDALDKIR
 61 YESLTDPSKL DSGKELHINL IPNKQDRTLT
 91 IVDTGIGMTK ADLINNLGTI AKSGTKAFME
121 ALQAGADISM IGQFGVGFYS AYLVAEKVTV
151 ITKHNDDEQY AWESSAGGSF TVRTDTGEPM
181 GRGTKVILHL KEDQTEYLEE RRIKEIVKKH
211 SQFIGYPITL FV
```

Hsp90-NTD possesses an active enzymatic centre binding adenosine triphosphate (ATP) with ATP-coordinating residues Asn^{51}, Asp^{93}, Lys^{112}, and Phe^{138}, and the ATP binding pocket assuming the so-called *Bergerat* fold (Bergerat et al., 1997; Prodromou et al., 1997).[4]

Next, the conformational space sampled by the structures of the 162 individual protein chains of Hsp90-NTD structures was investigated. To this end, protein structures were represented by internal coordinates, and after utilising a simple secondary structure based approach to coarsely investigate conformational heterogeneity, a principal component analysis (PCA) was employed for primary dimensional reduction and covariance analysis, aiming to visualise the conformational determinants of Hsp90-NTD.

To study the secondary structure of the ensemble of structures of Hsp90-NTD, the sequence logo concept (Schneider and Stephens, 1990) was adapted to use an alphabet of eight letters denoting specific secondary structure elements derived from the experimental protein structure and assigned by DSSP (Kabsch and Sander, 1983), instead of the protein amino acid sequence. This novel approach allows for intuitive and straightforward state-based interpretation of conformational flexibility.

'I', body centred, and 'C', C-centred), and the next three, the most relevant symmetry element.
[4]comprising amino acid residues 44 to 57, 93 to 99, and 130 to 140 in the above sequence.

Figure 4.1 – Secondary structure variability across known structures of Hsp90-NTD. Information content is measured in bits, and the alphabet used was 'H' (α-helix), 'B' (residue in isolated β-bridge), 'E' (extended strand participating in β-ladder), 'G' (3_{10}-helix), 'I' (π-helix), 'T' (hydrogen bonded turn), 'S' (bend), and the blank symbol for all residues where no secondary structure element could be assigned based on the molecular geometry. Numbers below the sequence correspond to the amino acid residue numbering of UniProt identifier P07900. Helical structures are displayed in blue, β-strands in black, and turns and bends in green.

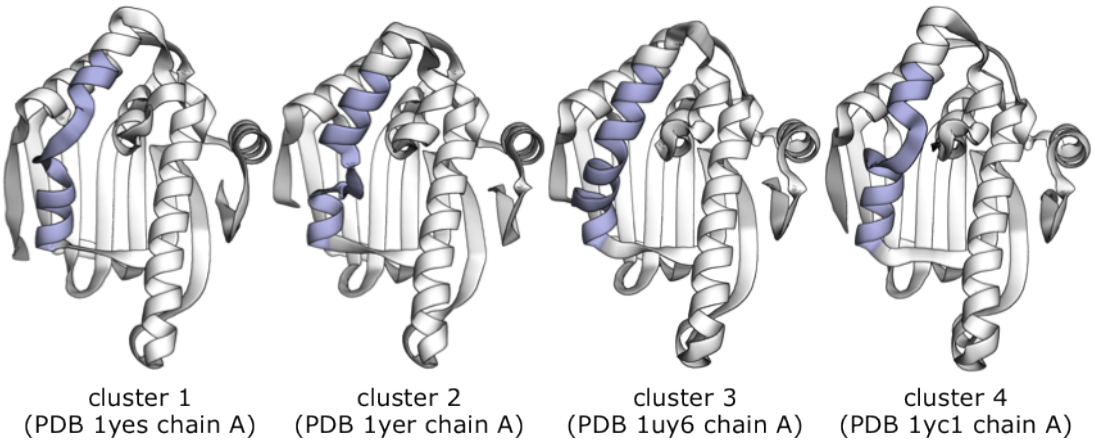


Figure 4.2 – Manual clustering of structures of Hsp90-NTD based on the conformation of residues 100 to 120 (coloured light blue). One representative structure for each of the four largest clusters is shown. The complete cluster assignment can be found in Table 4.1. Protein structures are shown in cartoon representation with round α-helices, flat β-strands and simplified straight loop segments. All *apo* structures of the protein are contained in cluster 2.

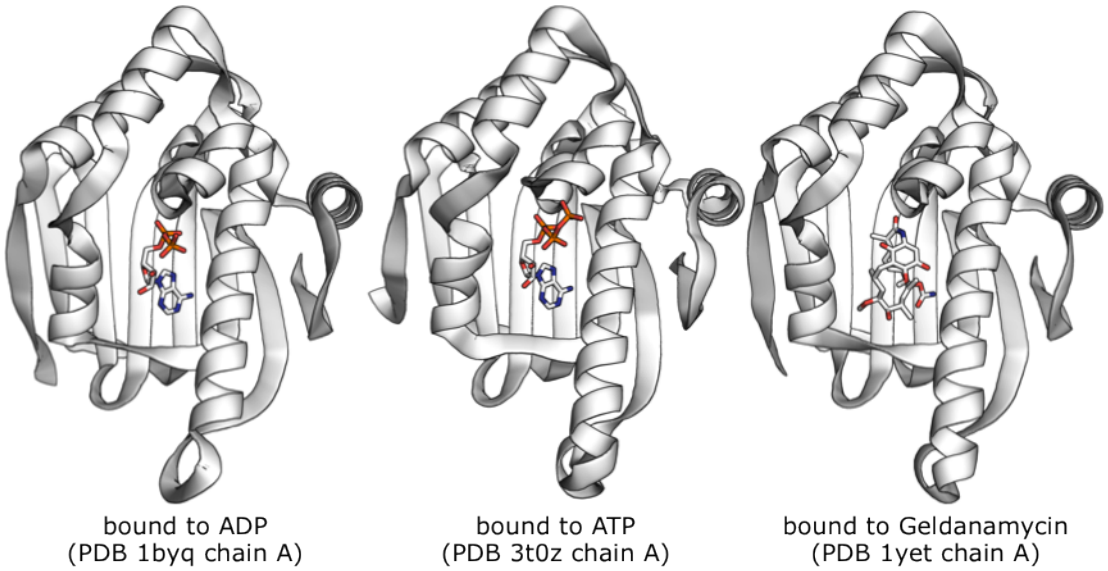


Figure 4.3 – Hsp90-NTD bound to ADP, ATP and geldanamycin. Protein structures are shown in cartoon representation with round α-helices, flat β-strands and simplified straight loop segments, while ligand structures are shown in stick representation with white carbon, blue nitrogen, red oxygen and orange sulphur atoms.

Table 4.1 – Manual cluster assignment of Hsp90-NTD structures, based on manual inspection of 162 superimposed protein structures. *Apo* structures of the protein are contained in cluster 2.

cluster number	PDB identifiers	cluster size
1	1byq.A, 1osf.A, 1yc3.A, 1yes.A, 1yet.A, 2byh.A, 2byi.A, 2xdk.A, 2xdx.A, 2xjg.A, 2xjj.A&B, 2xjx.A, 2xk2.A, 2yef.A, 3b27.A, 3ekr.A, 3hek.A, 3k97.A, 3k99.B&C, 3r4m.A, 3r4n.A, 3r4o.A, 3r4o.B, 3r4p.A, 3rlr.A, 3t0z.A, 3t10.A, 3t1k.A&B, 3t2s.A&B, 3vha.A, 3vhc.A, 3vhd.A, 4egi.A, 4egk.A	38
2	1uyl.A, 1yer.A, 2bsm.A, 2bt0.A, 2ccs.A, 2cct.A, 2ccu.A, 2jjc.A, 2qfo.B, 2uwd.A, 2vci.A, 2vcj.A, 2wi1.A, 2wi2.A, 2wi3.A, 2wi5.A, 2xab.A, 2xdl.A, 2ye2.A, 2ye3.A, 2ye4.A, 2ye5.A, 2ye6.A, 2ye9.A, 2yea.A, 2yeb.A, 2yec.A, 2yed.A, 2yeg.A, 2yeh.A, 2yi0.A, 2yi6.A, 2yi7.A, 2yjw.A, 3b24.A, 3b26.B, 3bm9.A, 3hhu.A, 3owb.A, 3t0h.A, 4eeh.A, 4egh.A	42
3	1uy6.A, 1uy7.A, 1uy8.A, 1uy9.A, 1uyc.A, 1uyd.A, 1uye.A, 1uyf.A, 1uyg.A, 1uyh.A, 1uyi.A, 1uyk.A, 2fwy.A, 2fwz.A, 2h55.A, 2qg2.A, 2wi4.A, 2wi7.A, 2xds.A, 2xdu.A, 2ye7.A, 2ye8.A, 2yee.A, 2yei.A, 2yej.A, 2yjx.A, 2yk2.A, 2yk9.A, 2ykb.A, 2ykc.A, 2yke.A, 2yki.A, 2ykj.A, 3b25.A, 3d0b.A, 3hyy.A, 3hz1.A, 3hz5.A, 3inw.A, 3inx.A, 3mnr.P, 3o0i.A, 3qdd.A, 3qtf.A, 3r91.A, 3r92.A, 3rkz.A, 4eft.A, 4efu.A	49
4	1yc1.A, 1yc4.A, 2bt0.B, 2bz5.A, 2qfo.A, 2qg0.A&B, 2wi2.B, 2xab.B, 2yeg.B, 3b24.B, 3b26.A, 3b28.A, 3hhu.B, 3ow6.A, 3tuh.A&B	17
5	2qf6.A&B&C&D	4
6	2xhr.A, 3b28.B, 3bmy.A	3
7	3ekr.B, 3owd.A, 3r4n.B, 3r4p.B	4
8	2bz5.B	1
9	2wi6.A	1
10	3hek.B	1
11	3rlr.B	1
12	3vhd.B	1

Symbols used were 'H' (α-helix), 'B' (residue in isolated β-bridge), 'E' (extended strand participating in β-ladder), 'G' (3_{10}-helix), 'I' (π-helix), 'T' (hydrogen bonded turn), 'S' (bend), and the blank symbol for all residues where no secondary structure element could be assigned based on the molecular geometry (Kabsch and Sander, 1983). The information content of a given position was calculated as $R_i = \log_2(8) - H_i$ with 8 being the length of the alphabet and H_i the uncertainty of position i, calculated as the Shannon entropy (Shannon, 1948) $H_i = -\sum f_{a,i} \log_2 f_{a,i}$ with $f_{a,i}$ the relative frequency of symbol a at position i. To generate the sequence logo, the height of each letter representing a symbol of the a of the alphabet to be displayed at position i was then calculated as the product $f_{a,i} \cdot R_i$, with the symbols sorted by their frequency from bottom (most frequent) to top (least frequent). It is evident that a symbol with frequency 1.0 has an uncertainty of $H_i = 0$ and therefore a height of $R_i = \log_2(8) = 3$ *bits*, while the uniform distribution of all eight symbols leads to $H_i = -\sum_{a=1}^{8} 1/8 \log_2 1/8 = 3$, leading to the minimum, zero information content $R_i = \log_2(8) - 3 = 0$ bits.

The analysis was conducted with, and figures produced using the WebLogo server (Crooks et al., 2004) without small sample correction. Conformational variability was detected in the protein segment comprising amino acid residues 100 to 120 having predominantly helical nature (Figure 4.1). Manual inspection of super-imposed protein structures suggested the presence of distinct conformational clusters (Figure 4.2 and Table 4.1).

Next, a PCA was carried out to investigate in more detail the conformational variability of Hsp90-NTD (see Methods section 4.4.1). To this end, protein structures were represented by internal coordinates. Two different approaches were used to compute internal coordinates. In the first approach, each amino acid residue i of protein structure a was represented by a vector $\vec{d}_{a,i} = d_{ij}$ for all residues $j \neq i$ where d_{ij} is the distance in Cartesian space between the coordinates of C^{α} atoms of residues i and j. Then, for each protein structure a and for each of its amino acid residues i, the Euclidean distance was calculated between vector $\vec{d}_{a,i}$ and the average vector $\vec{d}_{m,i}$ for that amino acid position i, with the latter obtained by averaging vectors $\vec{d}_i$ over all protein structures, thereby avoiding the arbitrary choice of a single reference structure. This procedure yields a vector of one scalar

value per amino acid for each protein structure. In the second approach, instead of using C^{α} distance vectors, backbone dihedral angles ϕ and ψ were calculated for each residue, and a vector $\vec{d}_{a,i} = (\sin\phi, \cos\phi, \sin\psi, \cos\psi)_{a,i}$ was constructed; then, analogous to the first approach, scalar Euclidean distances were calculated for each residue between vectors $\vec{d}_{a,i}$ and $\vec{d}_{m,i}$, where the latter was an average over all available protein structures. Again, this yields a vector of one scalar value per amino acid for each protein structure. However, as for the definition of backbone dihedral angles for a given amino acid i depends on coordinates of atoms in residues $i-1$, i, and $i+1$, amino acid residues on the N- and C-terminus as well as residues before and after gaps in the multiple alignment of amino acid sequences extracted from the protein structure files had to be omitted; i.e., the first approach yields 162 vectors of length 205, whereas the second approach yields 162 vectors of length 203.

In order to conduct the principal component analysis (PCA), the *sample covariance matrix* $\mathbf{C}$ was computed and decomposed as follows: the N protein descriptor vectors of length K were arranged in the $N \times K$ data matrix $\mathbf{X}$ such that x_{ij} denotes the j-th amino acid position of the i-th protein structure descriptor x_i with $i = 1..N$ and $j = 1..K$. For each amino acid position/matrix column j, the sample mean vector $\bar{x}_j = N^{-1} \sum_{i=1}^{N} x_{ij}$ was calculated, and the $K \times K$ sample covariance matrix $\mathbf{C}$ estimated with individual elements $c_{jk} = (N-1)^{-1} \sum_{i=1}^{N} (x_{ij} - \bar{x})(x_{ik} - \bar{x})^{\top}$.

$\mathbf{C}$ was then digonalised as $\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$ with $\mathbf{D}$ the diagonal matrix of eigenvalues of $\mathbf{C}$ ($d_{ij} = \lambda_m$ the m-th eigenvalue of $\mathbf{C}$ for $i = j = m$ with $m = 1..K$, and 0 otherwise), and $\mathbf{V}$ the $K \times K$ matrix of eigenvectors arranged in columns in corresponding order. Both matrices $\mathbf{D}$ and $\mathbf{V}$ were then re-arranged such that $\lambda_k \geq \lambda_l$, $\forall l > k = 1..K-1$.[5] The relative importance of each principal component can be estimated as the percentage of variance explained by it, $\mathrm{var}_{\mathrm{ex},k} = \lambda_k / \sum_{j=1}^{K} \lambda_j$.

The eigenvector matrix was then used to project the original data matrix onto the new orthogonal basis,[6] generating the $N \times K$ matrix $\mathbf{Y} = \mathbf{V}^{\top}\mathbf{X}^{\top}$. To normalise data, values of $\mathbf{Y}$ were expressed as *z-scores*, that is $z_{ij} = (y_{ij} - \mu_y)/\sigma_y$ with μ_y

[5] as there are pairs of corresponding eigenvectors and -values, the order on $\mathbf{V}$ is used to permutate $\mathbf{D}$.

[6] $v_a \cdot v_b = 0$ for $a \neq b$ and v_a and v_b the a-th and b-th column vector of $\mathbf{V}$, respectively.

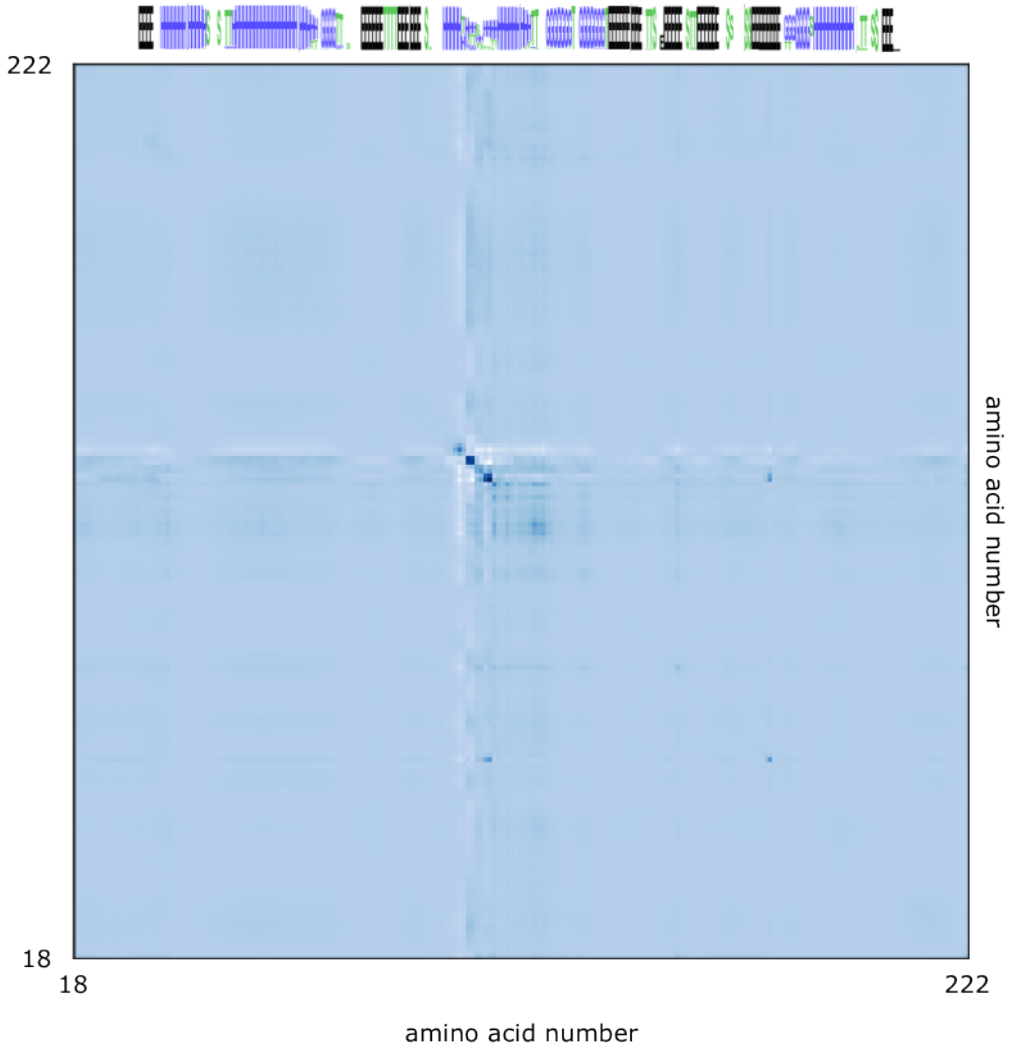


Figure 4.4 – Covariance matrix of protein structure descriptors of Hsp90-NTD, based on 162 protein structures and mutual C^{α} backbone atom distances descriptors (see main text). Data is coloured from minimum (white) to maximum (blue) value (-37.4 to 104.1). On the top of the figure, the secondary structure of the protein structure ensemble is represented in a sequence logo based way (see Figure 4.1 and main text for reference).

and σ_y the mean and standard deviation, respectively, calculated over the column vector y_j with $i = 1..N$ and the j-th (ordered) principal component ($j = 1..K$).

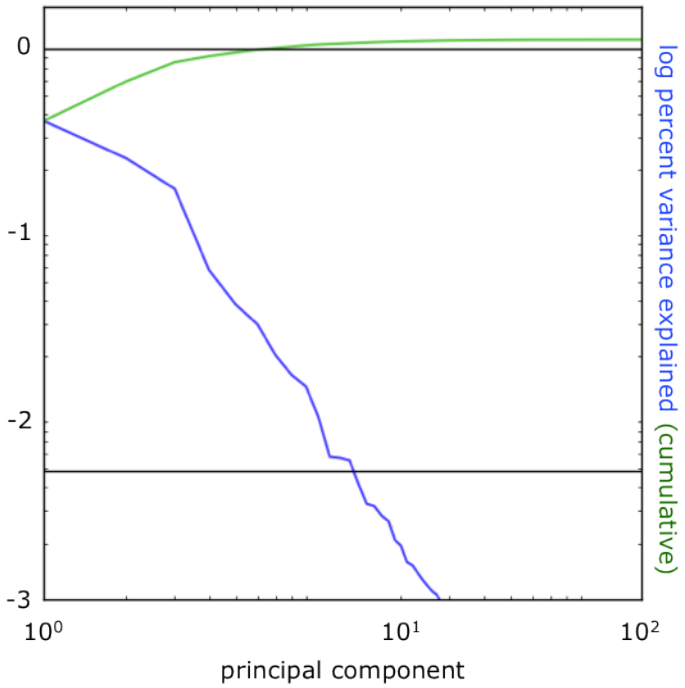


Figure 4.5 – Scree plot of the conformational analysis of Hsp90-NTD, based on the mutual C^{α} backbone atom distances descriptors (see main text). Horizontal black lines indicate values of 0.90 (top) and $1/205$ (bottom) of variance explained.

This approach was applied to investigate the conformational space sampled by Hsp90-NTD in 162 structures deposited in the PDB. Firstly, the sample covariance matrix of the structures represented by the C^{α} backbone distance descriptors (see above) was constructed (Figure 4.4), indicating correlated structural variation being present in the central part of the amino acid sequence, corresponding to the region of amino acid residues 100 to 120 (see also top of Figure 4.8 for a vector of mean values of the columns of the covariance matrix) already suggested to be variable in the DSSP based sequence logo analysis (Figure 4.1).

Next, the covariance matrix was diagonalised, decomposing it into a matrix of orthogonal eigenvectors and a diagonal matrix of 205 corresponding eigenvalues. Ordered by their magnitude, the eigenvalues suggested most of the variance in the data being captured by the first few principal components (Figure 4.5).

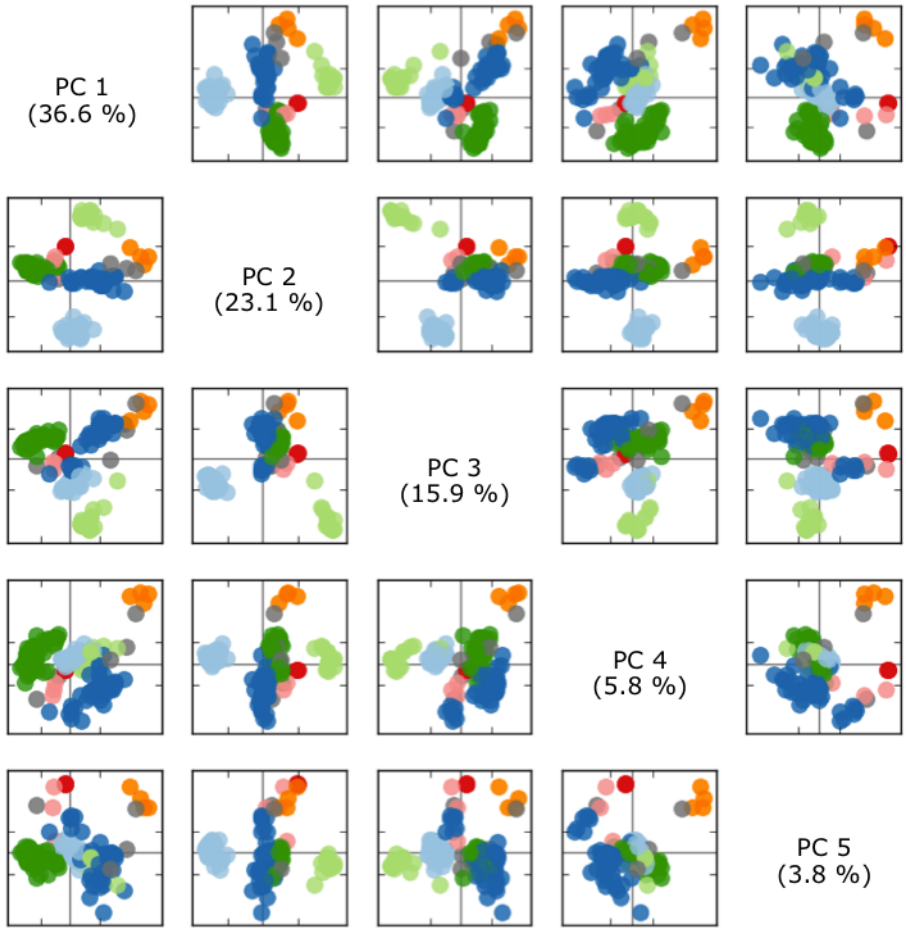


Figure 4.6 – Scatter plot of PC-space transformed structure descriptors of Hsp90-NTD, based on the mutual C^{α} backbone atom distances descriptors (see main text). The fraction of variance explained by each principal component is indicated in parentheses. Each dot corresponds to one structure of Hsp90-NTD, coloured by an assignment to a conformational group based on the conformation of amino acid residues 100 to 120 (see main text, Figure 4.2 and Table 4.1). Members of cluster 1 are coloured blue; cluster 2: light blue; cluster 3: green; cluster 4: light green; cluster 5: red; cluster 6: light red; cluster 7: orange; and all singleton clusters (8-12): grey. Data are expressed as *z-scores*, with 0 indicated by solid grey lines, and ± 1 standard deviations by tick marks.

Transforming the data into principal component space (Figure 4.6), the conformational clustering suggested by the conformation of the amino acid residues 100 to 120 (Figure 4.2 and Table 4.1) was found to be well reflected by the PCA. About 23.1 % of the variance was found to be explained by variation along the second principal component, corresponding to conformational differences in amino acid residues around position 110 (Figure 4.7, and third panel of Figure 4.8), with an additional contribution of residues around position 180 that was found to be more prominent in the third principal component.

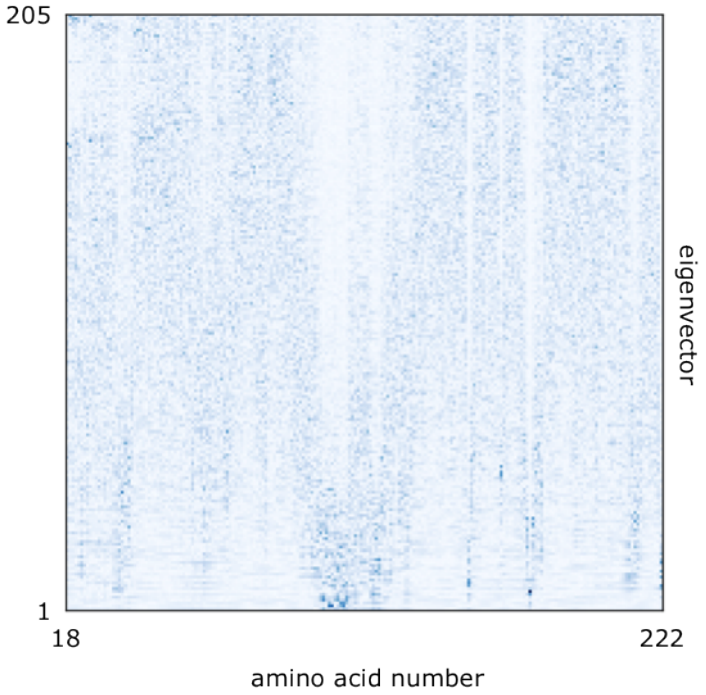


Figure 4.7 – Eigenvector matrix of the conformational analysis of Hsp90-NTD, based on the mutual C^{α} backbone atom distances descriptors (see main text). Eigenvectors are arranged by the magnitude of their corresponding eigenvalue λ_k, with the largest at the bottom of the matrix. Data is coloured from minimum (white) to maximum (blue) value (0 to 0.79) of the absolute values in matrix $\mathbf{V}$ (see main text). Note that the columns of the matrix correspond to the numbering of amino acid residues in Hsp90-NTD.

The PCA was then repeated with the descriptors based on the backbone dihedral angles (see above) which are more sensitive to local conformational differences but do not take into account more global changes as the C^{α} based descriptors do.

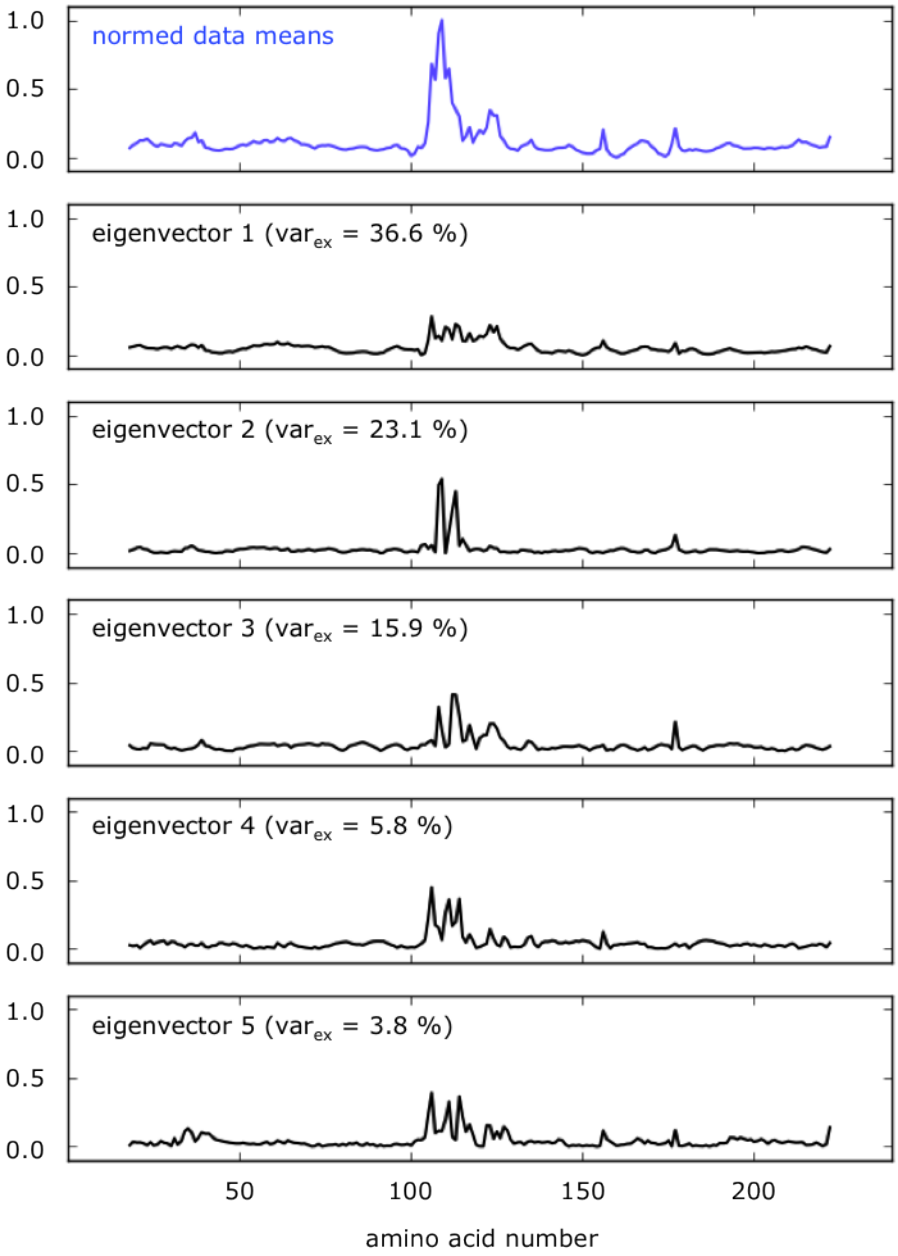


Figure 4.8 – Data mean and eigenvectors of the conformational analysis of Hsp90-NTD, based on the mutual C^{α} backbone atom distances descriptors. Top panel (blue): mean values of the untransformed input data (averaged values of the column vectors of matrix **X** in the main text), scaled to the interval $[0, 1]$. Top to bottom, black: absolute values of elements in the first five eigenvectors (row vectors of the eigenvector matrix **V** in the main text), arranged by the magnitude of their corresponding eigenvalue, with the first eigenvector having the largest eigenvalue (fraction of variance explained). See Figure 4.10 for a mapping of eigenvectors onto the structures of Hsp90-NTD.

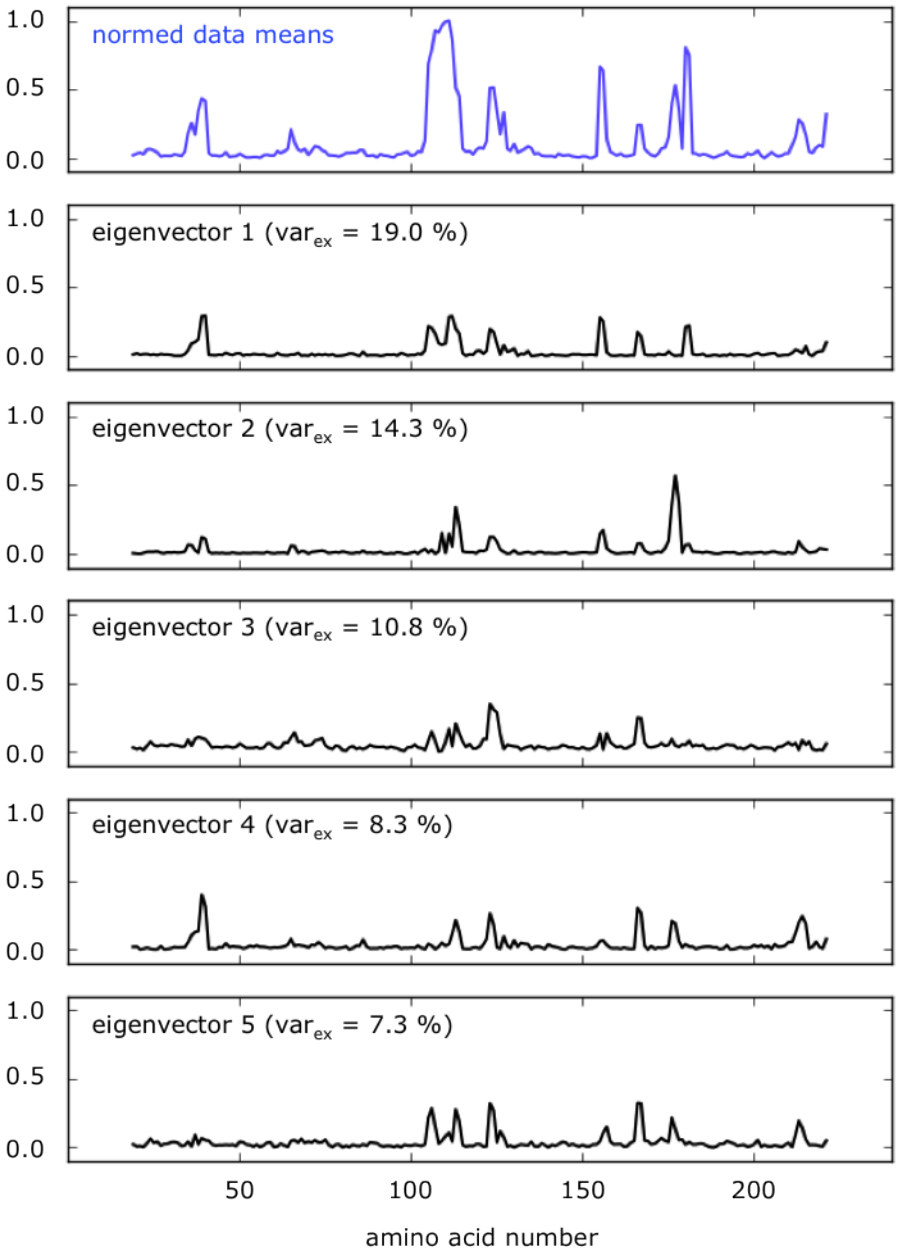


Figure 4.9 – Data mean and eigenvectors of the conformational analysis of Hsp90-NTD, based on the backbone dihedral angles. See Figure 4.8 for reference.

The variance explained by the first five principal components was found to be lower (59.7 %) when compared to the first, global descriptor (85.2 %), and the clustering reflecting the conformations of residues 100 to 120 was found to be less pronounced than in Figure 4.6 (data not shown). However, as in the analysis based on the first decriptor (Figures 4.7 and 4.8), detailed inspection of the absolute values of the first eigenvectors from the analysis based on the second descriptor indicates variability in the region of residues 100 to 120, with a stronger contribution from some regions located C-terminally of this (Figure 4.9). The first five eigenvectors of both PCAs were mapped on the *apo* structure of Hsp90-NTD (Figure 4.10) for a more intuitive structure based visualisation.

4.2.2 Xenon binding does not strongly affect global protein structure in proteins other than Hsp90-NTD

Next, the PCA based conformational analysis was used to investigate whether the binding of xenon affects the overall protein structure of proteins other than Hsp90-NTD when compared to a set of reference structures of the respective protein system *without* xenon. To this end, for each xenon binding protein structure, protein structures with the same UniProt identifier were retrieved from the PDB as described in Methods sections 3.4.1 and 3.4.2, omitting structures with missing residues[7] as well as UniProt identifiers with less than 20 known structures.[8]

Then, for each set of structures with the same UniProt identifier, the fit of the xenon binding protein structures to the distribution of reference structures was quantified, based on the conformational PCA conducted for the set of these protein structures. To this end, the weighted Euclidean distance $d(x, y) = \sqrt{\sum_i w_i (x_i - y_i)^2}$ between two protein conformations[9] $\vec{x}$ and $\vec{y}$ was calculated with the weight $w_i = \mathrm{var}_{\mathrm{ex},i}$ (note that by definition $\sum_i \mathrm{var}_{\mathrm{ex},i} = 1$). Then, histograms of the distributions of pairwise distances were compared: $P(a|x_a \in \mathrm{XE}) = \sum_{y \notin \mathrm{XE}} H\left(d(x_a, y)\right)$ the histogram of weighted distances of the xenon binding protein structure a to any other, non-xenon binding structure in the data sub

[7] as they would hinder calculation of internal coordinate descriptors, see above.
[8] as the distribution of structures in the conformational space would likely be too coarse.
[9] represented by their coordinates in principal component space.

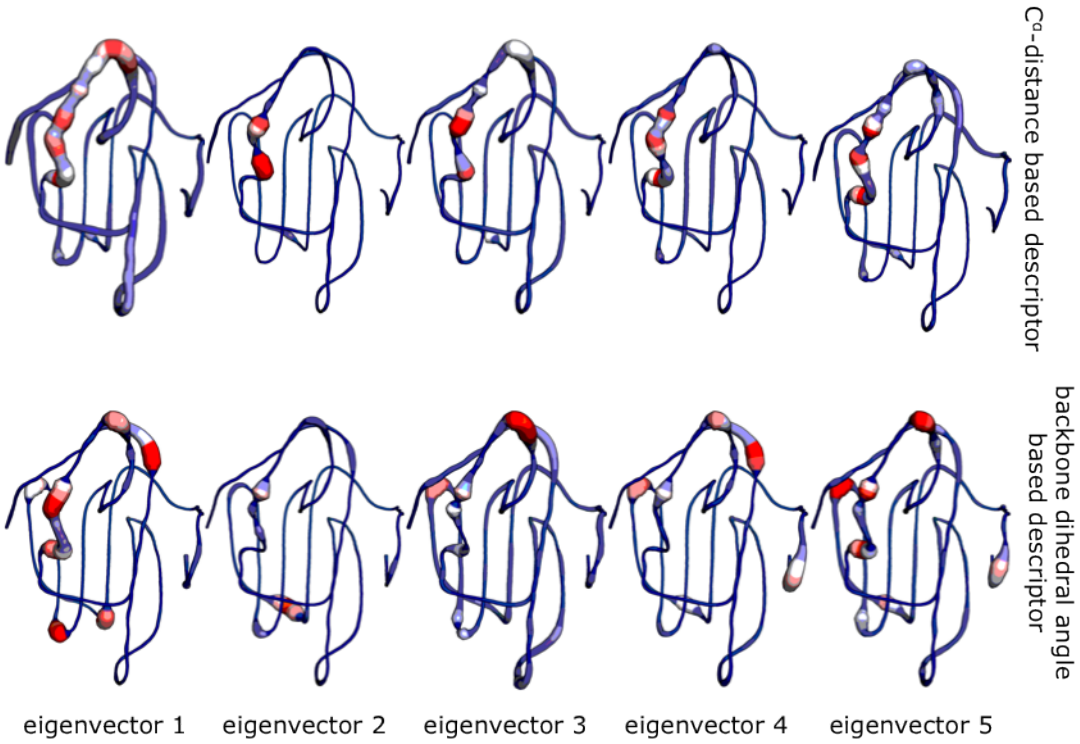


Figure 4.10 – Eigenvectors of the conformational PCA of the ensemble mapped onto a single structure of Hsp90-NTD. The absolute values of the first five eigenvectors (see Figures 4.8 and 4.9) were mapped on the structure of *apo* Hsp90-NTD (PDB identifier 1uyl, chain A). Top row of panels, PCA conducted with C^{α} distance based internal coordinate descriptors, representing 36.6, 23.1, 15.9, 5.8, and 3.8 % of variance, respectively (from left to right). Bottom row of panels, PCA conducted with backbone dihedral angle based descriptors, 19.0, 14.3, 10.8, 8.3, and 7.3 %. Structures are shown as simplified backbone traces, with the diameter of the backbone trace and its color chosen according to the magnitude of the value of the respective eigenvector at a given amino acid position (expressed as its absolute value; see Figures 4.8 and 4.9) on a relative scale from blue to white to red, with blue and red the minimum and maximum value, respectively, for that eigenvector. The protein region displaying most variability corresponds to a helical segment comprising amino acids 100 to 120 close to the ligand binding site (see main text).

set; and $Q = \sum_{x \notin \mathrm{XE}} \sum_{y \notin \mathrm{XE}} H\left(d(x,y)\right)$ the background distribution of distances between any two non-xenon binding protein structures in the same data sub set. Both histograms were scaled such that $\sum P(a) = \sum Q = 1$ in order to allow for treating them as empirical probability density functions.

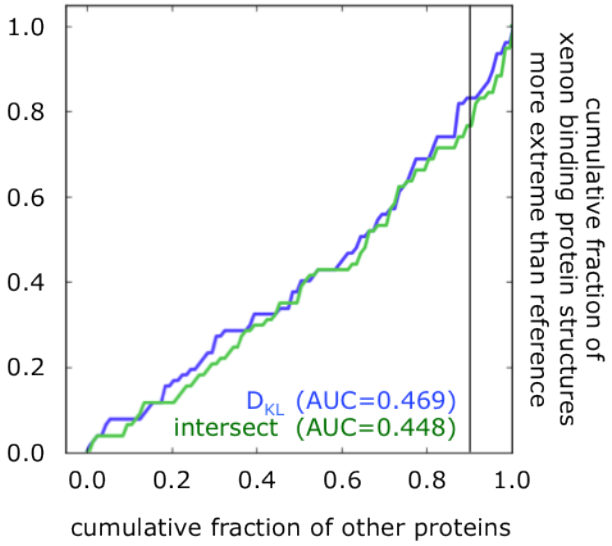


Figure 4.11 – General influence of xenon binding on protein structures. Fractions of values $s'_1 \leq s_1$ (green) and $s'_2 \leq s_2$ (blue) were computed for each xenon binding protein (see main text) and subjected to histogram analysis. The graph assumes a value of y_0 at threshold x_0, indicating that a fraction of y_0 xenon binding protein structures have a *more extreme* s'-value than a fraction of x_0 protein structures which do not bind xenon. The vertical black line indicates that there is a fraction of about 16.9 to 23.4 % of xenon binding protein structures where each one of them is *more extreme* than 90 % of the non-xenon binding protein structures with the same UniProt identifier. The area under the curve (AUC) was calculated by numerical integration of the histogram bins.

Differences between the distributions were quantified as $s_1(a) = \mathit{intersect}(P(a), Q)$, based on the histogram intersection: $\mathit{intersect}(P, Q) = \sum_i |p_i - q_i|$ with p_i and q_i the contents of the i-th bin of the respective histogram; and $s_2(a) = D_{\mathrm{KL}}(P(a) \| Q) + D_{\mathrm{KL}}(Q \| P(a))$, with $D_{\mathrm{KL}}(P \| Q) = \sum_m \ln(p_m/q_m) p_m$ the *Kullback-Leibler divergence*.[10] The optimal value of $s_1(a)$ and $s_2(a)$ is 0 for perfect agreement of both distributions; the maximum value of $s_1(a) = 2$ for non-overlapping histograms,

[10] as can be seen from its definition, D_{KL} is not symmetric; thus, a symmetrised version was used in order to compute $s_2(a)$.

while the maximum of $s_2(a)$ depends on the distributions under consideration. Values of $s_1(a)$ and $s_2(a)$ were computed for all xenon binding protein structures $x_a \in \text{XE}$ and compared to values $s'_1(b)$ and $s'_2(b)$ ($x_b \notin \text{XE}$) computed for non-xenon binding proteins as before, but this time using $P(b|x_b \notin \text{XE})$ instead of $P(a|x_a \in \text{XE})$. Results were then reported as fractions of the number of values $s' \leq s$, with the fraction assuming a value of 0 if all non-xenon binding proteins have a *more extreme* distance distribution than the xenon binding protein under consideration, and a value of 1 if the xenon binding protein is *more extreme* than any non-xenon binding protein in the data set.[11]

These fractions were computed for each xenon binding protein structure and represented as cumulative histograms in Figure 4.11. It can be appreciated that most xenon binding structures do not behave *more extremely* than most non-xenon binding structures when it comes to the distribution of distances in PC-space. About 16.9 to 23.4 % of xenon binding protein structures can be regarded as outliers in the conformational analysis if an outlier is defined as being *more extreme* than 90 % of non-xenon binding protein structures. Values of the area under the curve of 0.47 (with the Kullback-Leibler divergence as a histogram distance measure) and 0.45 (using the histogram intersection, see above) do not differ vastly from a value of 0.5, which would be obtained if s was randomly distributed on the interval $[s'_{\text{min}}; s'_{\text{max}}]$ (Figure 4.11).

4.2.3 Prediction of xenon binding to Hsp90-NTD

Next, the xenon likeness score developed in the previous chapter was applied to predict the possible xenon binding to Hsp90-NTD. To this end, experimental structures of Hsp90-NTD falling into the same conformational cluster as *apo* Hsp90-NTD were selected from the available structures in the PDB (Table 4.1) and processed as described previously (see Results section 3.2.3) by normalising the structures (see Methods section 3.4.1) and expanding the crystal unit cell images to account for crystal contacts (Methods section 3.4.5). Then, water and ion

[11] Histograms P and Q were estimated using 10 bins for distances d between 0 and the maximum overall distance detected. For the calculation of D_{KL} and the histogram intersection, as well as numerical integration (see below), each adjacent pair of histogram bins was replaced by ten new bins, interpolating linearly between them.

positions were used as surrogates for potential xenon binding positions and used to score xenon likeness in their respective structural context. Since cluster 2 in Table 4.1 contains all Hsp90-NTD *apo* crystal structures that form a discrete cluster in conformational space (Figure 4.6), and the unliganded form of the protein was to be investigated, only those 42 crystal structures were processed and included in the analysis.

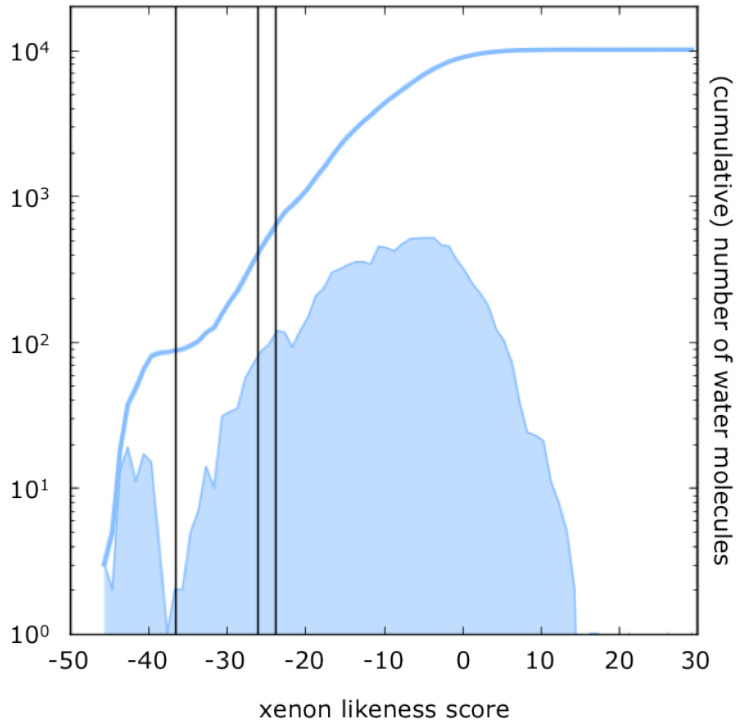


Figure 4.12 – Prediction of xenon binding to Hsp90-NTD, xenon likeness score distribution. For water molecules and ion positions within 42 Hsp90-NTD structures from the PDB (second cluster in 4.1), xenon likeness scores were computed and represented as a regular (shaded area) and cumulative histogram (blue line). A bin width of 1 score unit was used (with scores s_0 falling into the i-th bin if $i \leq s_0 < i+1$). Possible score threshold values from a threshold scanning analysis (-36.6, -26.1, and -23.8, see Results section 3.2.5 in the previous chapter) are represented as vertical black lines.

The distribution of xenon likeness scores of water molecules and ions from the structures that were obtained are depicted in Figure 4.12. A more detailed spatial analysis of well scoring positions can be found in Figures 4.13 to 4.16. While the majority of positions possess a score indicated unfavourable for xenon binding,

there are some protein moieties with favourable scores of less than −25.0 (indicated by cyan to blue colour and labelled with numbers one to four in Figures 4.13 to 4.16). For the subsequent score analysis, only positions with scores of less than −25.0 were taken into account.

The first group of well scoring positions (denoted '1' in Figures 4.13 to 4.16) was identified in the active site of the enzyme. Water molecules partially overlapping with ligands of the protein (compare also Figure 4.3) were found to exhibit xenon likeness scores of −25.1 to −34.1 ($N = 120$, mean $\mu = -28.8$, standard deviation $\sigma = 2.59$, median −29.0), while in 4 out of 42 structures there were water molecules with a xenon likeness score of −38.2 to −45.3.

A second group of well scoring positions, denoted '2a' and '2b', was found buried in the protein core proximal to the active site of Hsp90-NTD with water molecules present in 41 of 42 structures and xenon likeness scores between −39.3 and −46.1 ($\mu = -42.4$, $\sigma = 1.58$, median −42.5) / −36.4 and −43.8 ($\mu = -40.6$, $\sigma = 1.55$, median −40.5), respectively.

Finally, two surface-exposed groups of water molecules close to the entrance of the active site (labelled '3') or distal to it ('4') were found to have a moderately favourable xenon likeness score of around $\mu = -28.6$ ($\sigma = 2.18$, median −28.4, minimum score −25.1, maximum score −34.3) / $\mu = -26.2$ ($\sigma = 0.91$, median −25.9, min. −28.1, max. −25.3), respectively, with 40 or 13 water molecules observed experimentally in 42 protein structures, respectively.

It should be noted that the water position groups 2a, 2b and 4 are spatially more well-defined than the other groups where multiple water molecules were observed in single protein structures. This can be seen from the *radius of gyration* R_{gyr} which is defined as $R^2_{\text{gyr}} = N^{-1} \sum_i^N (r_i - r_m)^2$ for a group of N points r in Cartesian space and the centre of gravity[12] r_m: values of R_{gyr} for water position clusters 1, 2a, 2b, 3, and 4 were 3.64, 0.21, 0.20, 1.46, and 0.35 Å, respectively.

The results of this section are summarised in Table 4.2. Next, the prediction is validated by solving the structure of Hsp90-NTD in the presence and absence of xenon experimentally by X-ray crystallography.

[12] mean coordinate.

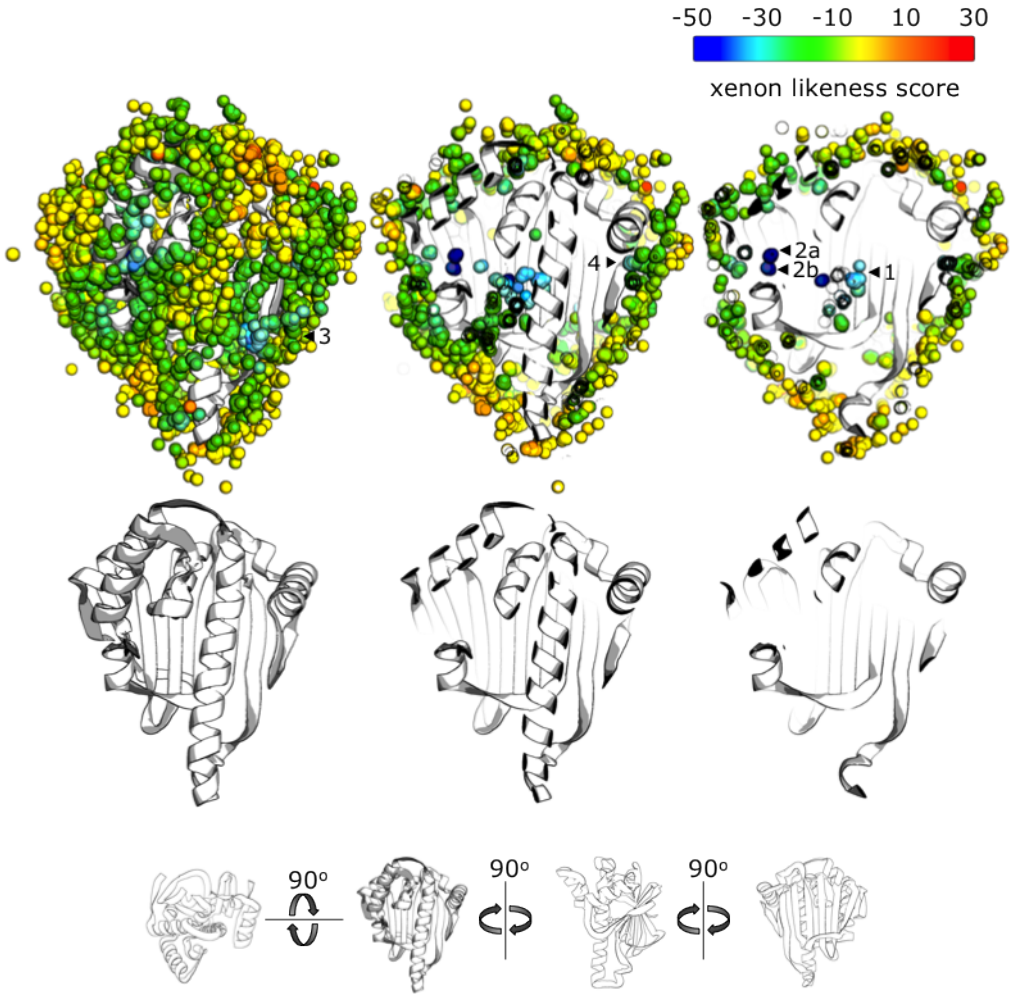


Figure 4.13 – Prediction of xenon binding to Hsp90-NTD (front view). In the topmost row of panels, xenon likeness scores of individual water molecules and ion positions (circles; positions taken from 42 experimental structures of Hsp90-NTD from the PDB) are represented coloured on a rainbow scale from a score of −50 (most favourable, blue) over −30 (generally favourable, cyan) to unfavourable (green, yellow, orange, and red, the latter corresponding to a score of +30). Three different sections through the structures are shown from left to right, with the plane of vision parallel to the paper plane. Clusters of potentially favourable xenon interaction (with xenon likeness scores of less than −25.0) are indicated by black arrows and numbers as discussed in the main text. In the second row of panels from top, the conformation of a single representative protein structure of Hsp90-NTD is shown (white, cartoon representation of PDB identifier 1uyl, chain A) in the same sections and perspective as in the panel immediately above, for the sake of clarity. In Figures 4.14 to 4.16, alternative views of the structure can be seen which are related to 90 or 180 degrees of rotation along different orthogonal axes, as indicated at the bottom.

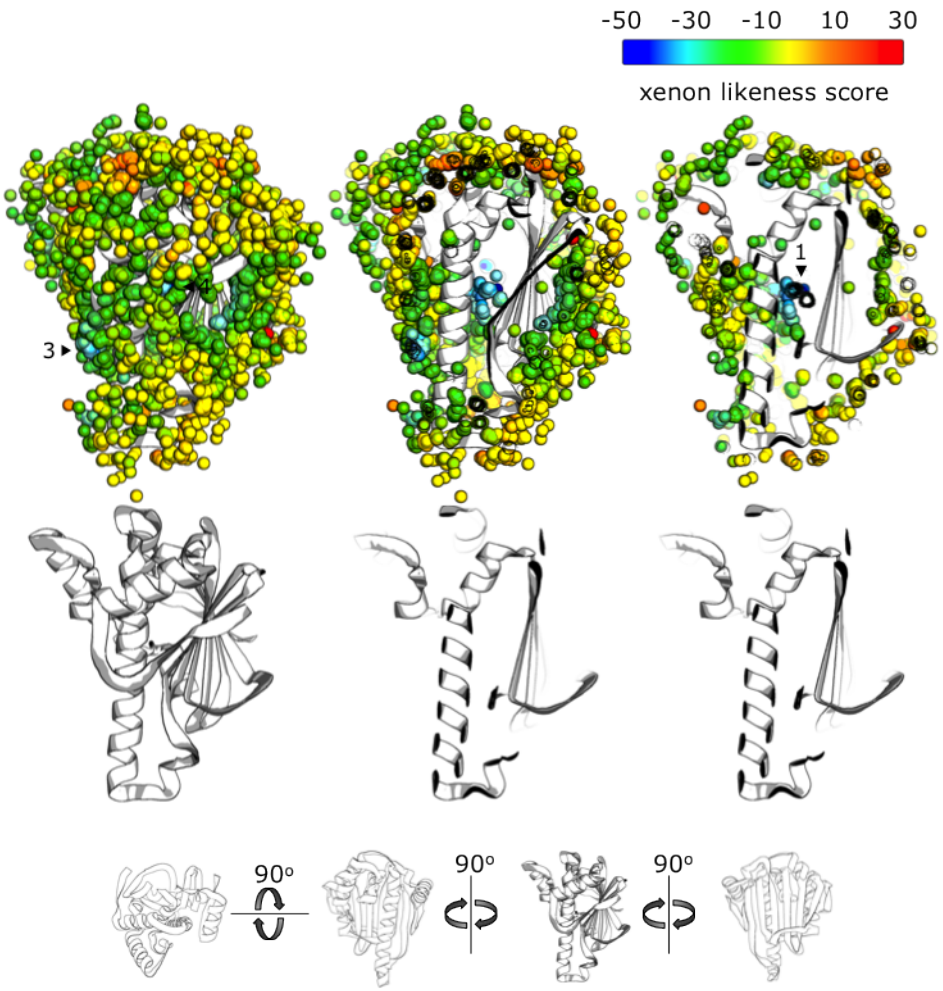


Figure 4.14 – Prediction of xenon binding to Hsp90-NTD (side view). See legend of Figure 4.13 for details.

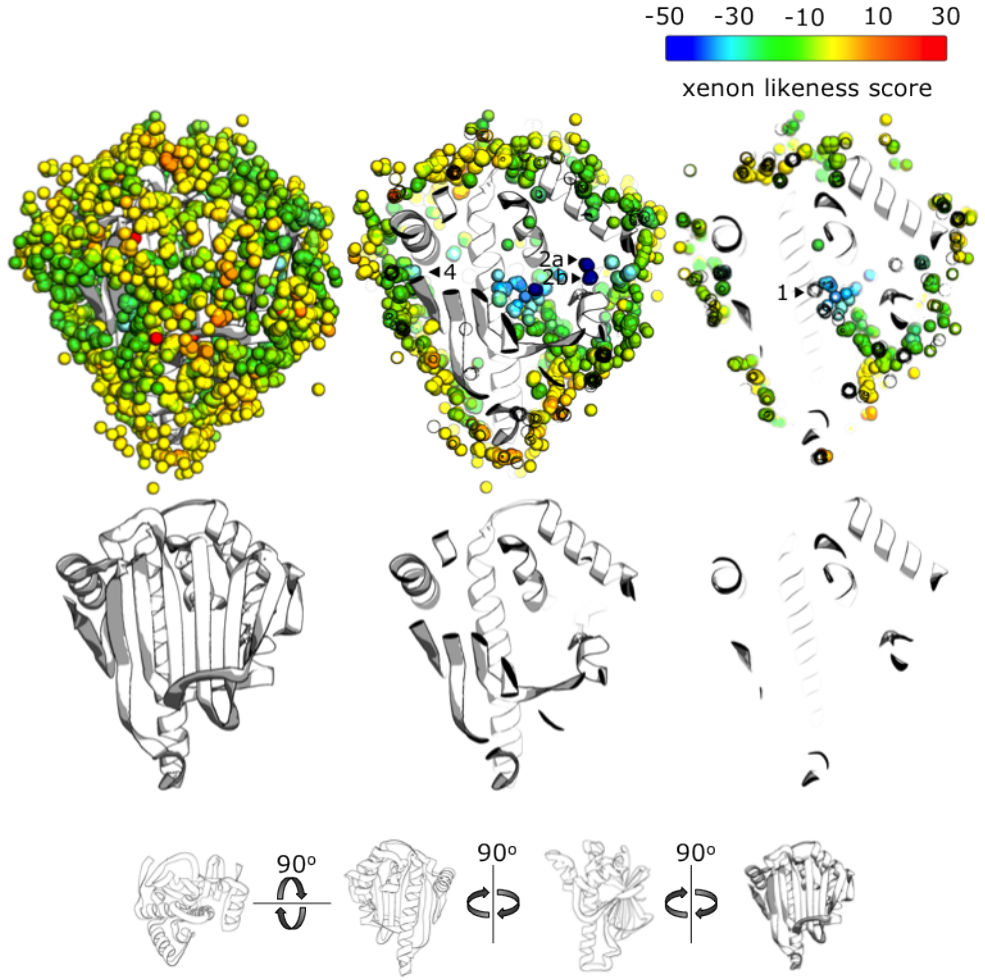


Figure 4.15 – Prediction of xenon binding to Hsp90-NTD (back view). See legend of Figure 4.13 for details.

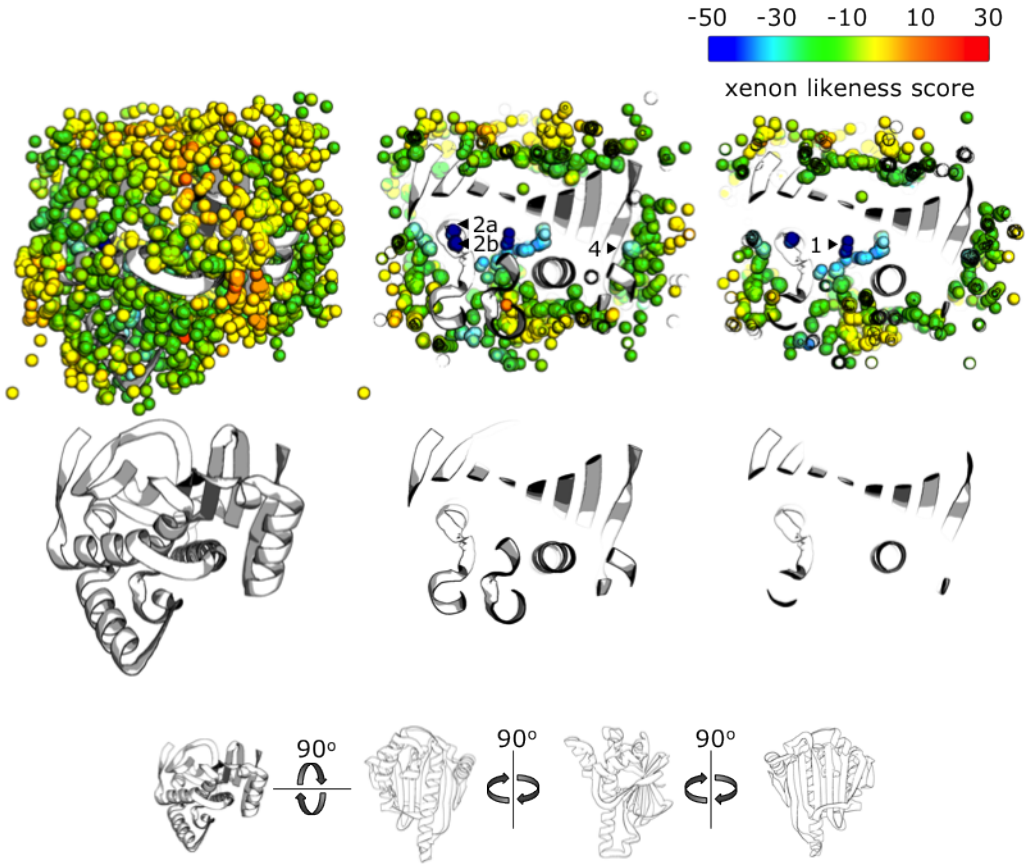


Figure 4.16 – Prediction of xenon binding to Hsp90-NTD (top view). See legend of Figure 4.13 for details.

Table 4.2 – Xenon likeness score predictions of water molecule positions within 42 structures of Hsp90-NTD. Groups of water molecules scoring more favourably than a xenon likeness score of −25.0 were retrieved from a graphical analysis (see Figures 4.13 to 4.16) and the number of water molecules in this group (N), the xenon likeness score average (μ) and standard deviation (σ), median, minimum and maximum xenon likeness score, and radius of gyration (R_{gyr}, see main text) are reported for each group of positions.

cluster	1	2a	2b	3	4
N	124	41	41	40	13
μ	−29.3	−42.4	−40.6	−28.6	−26.1
σ	3.53	1.58	1.55	2.18	0.91
median	−29.1	−42.5	−40.5	−28.4	−25.9
minimum	−45.3	−46.1	−43.8	−34.3	−28.1
maximum	−25.1	−39.3	−36.4	−25.1	−25.3
R_{gyr} (Å)	3.64	0.21	0.20	1.46	0.35

4.2.4 Structure determination of Hsp90-NTD binding to xenon

To investigate xenon binding to Hsp90-NTD experimentally, X-ray crystallography was used to determine the structure of Hsp90-NTD in the presence and absence of xenon by X-ray crystallography. Pure Hsp90-NTD was a kind gift of Prof. Roderick Hubbard and Drs. Ben Davis and Allan Surgenor of Vernalis (R&D) Ltd.[13] and crystallised as previously described (Wright et al., 2004; Barril et al., 2005) (see Methods section 4.4.2).

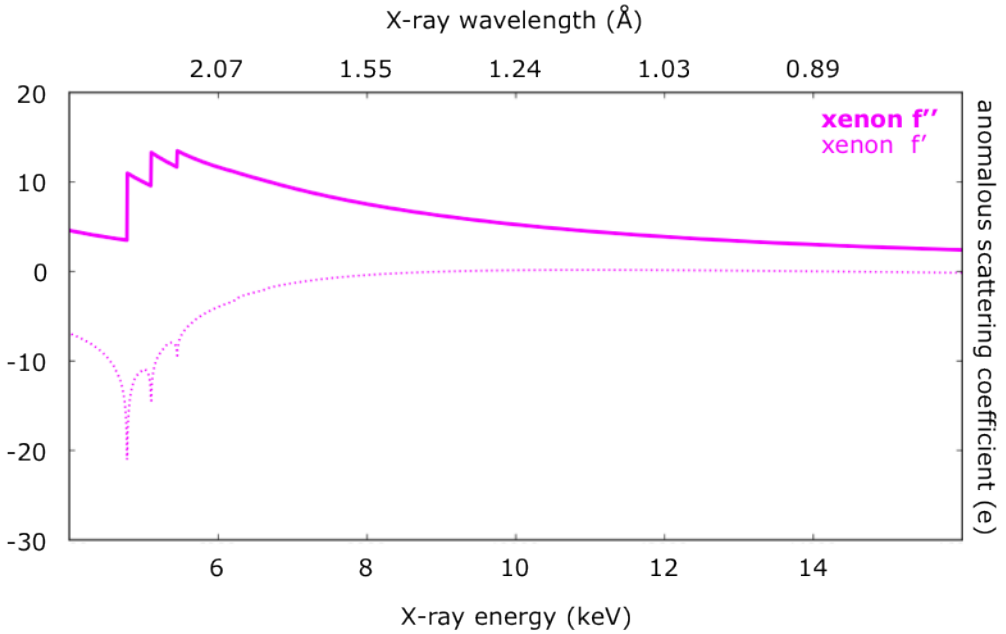


Figure 4.17 – Xenon anomalous scattering coefficients as a function of X-ray energy. The observable scattering coefficient f" (bold magenta line) is shown in dependence of the X-ray energy (x-axis, in keV) and the corresponding X-ray wavelength (in Å). X-ray energy (E, in eV) and wavelength λ (in metres) are related by $\lambda = hc/E$ with h Planck's constant ($4.13566733 \times 10^{-15}$ eV s) and c the speed of light in vacuum (299,792,458 m/s). Theoretical scattering coefficients were taken from http://skuld.bmsc.washington.edu/scatter/AS_periodic.html.

Diffraction data were collected using synchrotron radiation from the X12 beam line at the DORIS storage ring at the EMBL Outstation Hamburg (Germany), in

[13]Vernalis (R&D) Ltd., Granta Park, Abington, Cambridge (United Kingdom).

collaboration with Dr. Michele Cianci, who also processed the data and assisted during structure determination and refinement (see Methods section 4.4.3).

A xenon chamber was used to soak protein crystals with xenon gas (Figure 4.18 and Methods section 4.4.3). Figures 4.19 and 4.20 show representative diffraction patterns of Hsp90-NTD in native state and after incubation with 10 atm xenon pressure. A wavelength of 1.54 Å was chosen for data collection as a good compromise between xenon anomalous signal strength (Figure 4.17) and resolution. A summary of data collection statistics can be found in Table 4.3. Crystals were found to diffract to a maximum resolution of 1.65 Å in case of the native data set in absence of xenon and 1.68 Å resolution in the case of the protein crystal exposed to a xenon pressure of 10 atm.

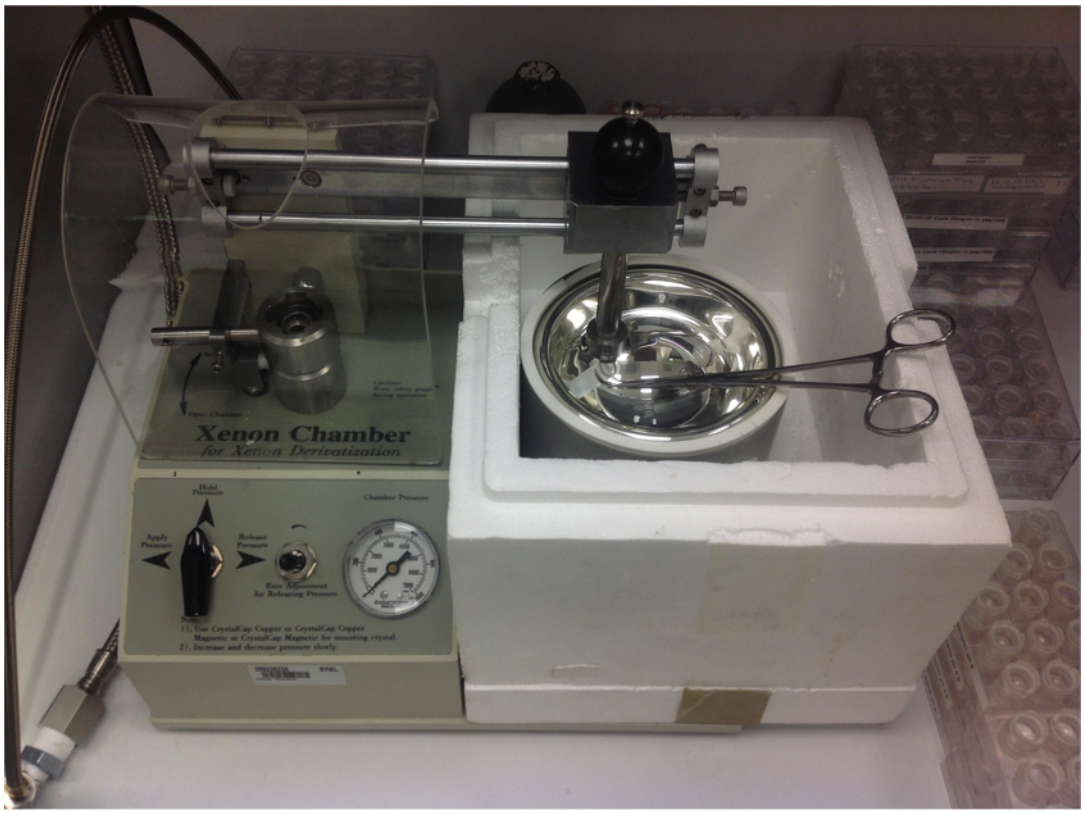


Figure 4.18 – Xenon pressure chamber used to expose protein crystals to xenon gas. Technical details and a schematic representation of the device can be found in Methods section 4.4.3 and Figure 4.29. Photo courtesy of Dr. Michele Cianci, EMBL Hamburg (Germany).

After data processing (see Methods section 4.4.3), the protein structures were solved using a molecular replacement method (see section 3.1.2 in the previous chapter) with the apo structure of Hsp90-NTD (PDF identifier 1uyl) (Wright et al., 2004) as the initial model template. The atomic structure model was refined iteratively by alternating automatic structure refinement using the Refmac5 (Murshudov et al., 1997) program, and manual rebuilding and inspection using COOT (Emsley et al., 2010) (see Methods section 4.4.4). Refinement statistics are summarised in Table 4.4. The final R-factor/R_{free} of the two data sets were 16.12/20.58 % (native) and 15.38/19.25 % (10 atm xenon data set), respectively. The overall stereochemical quality was excellent as documented by the Ramachandran plot analysis conducted (Table 4.4). Figure 4.22 shows a comparison between the initial template model and the final structures of the native and 10 atm xenon data set. The native and the 10 atm xenon structure show a mutual C^{α}-RMSD of 0.27 Å, while they have an RMSD of 0.33 Å (native) and 0.21 Å (10 atm xenon) to the template structure. When residues 176 to 178 (see Figure 4.22 and 4.23 (E) and (F) for reference) are omitted from the RMSD analysis, the mutual RMSD is 0.11 Å, while the RMSD to the template structure is 0.19 Å and 0.20 Å, respectively. Employing the PCA based conformational analysis described above, the novel structures cluster well with structures located in the conformational space occupied by 'cluster 2' (Table 4.1) containing the *apo* structures of Hsp90-NTD (Figure 4.21).

In the 10 atm xenon data set, several xenon atoms were detected, and in both the native and the 10 atm xenon data set, several chloride ions were found (Figure 4.24). A more detailed view of selected structural features of the atomic model and the electron density map can be found in Figure 4.23. Anomalous signal intensities were determined for all sulphur atoms in methionine residues[14] (B-factors can be found in Table 4.6), chloride ions, and xenon atoms (Table 4.5). For numbering reference of the chloride and xenon moieties see Figure 4.24. A detailed discussion of the agreement between prediction and experimental verification follows.

[14]there are no cystein residues present in Hsp90-NTD.

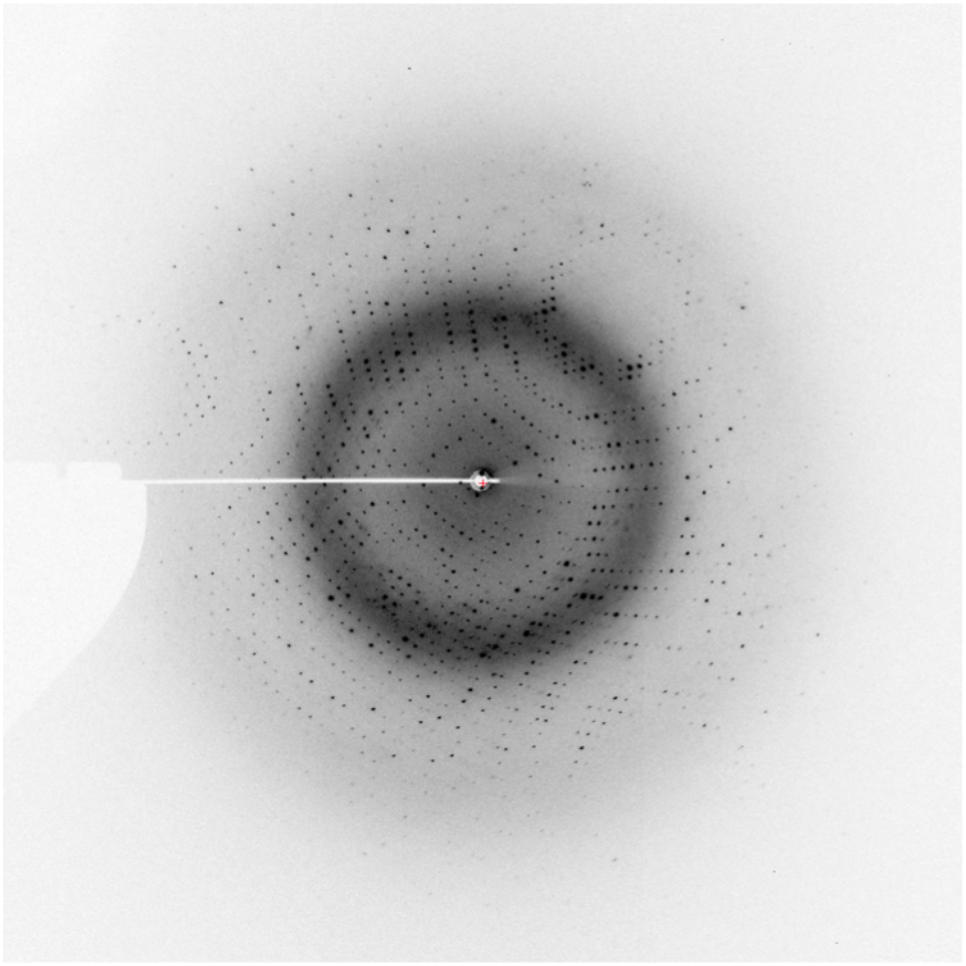


Figure 4.19 – Representative diffraction pattern of Hsp90-NTD in absence of xenon.

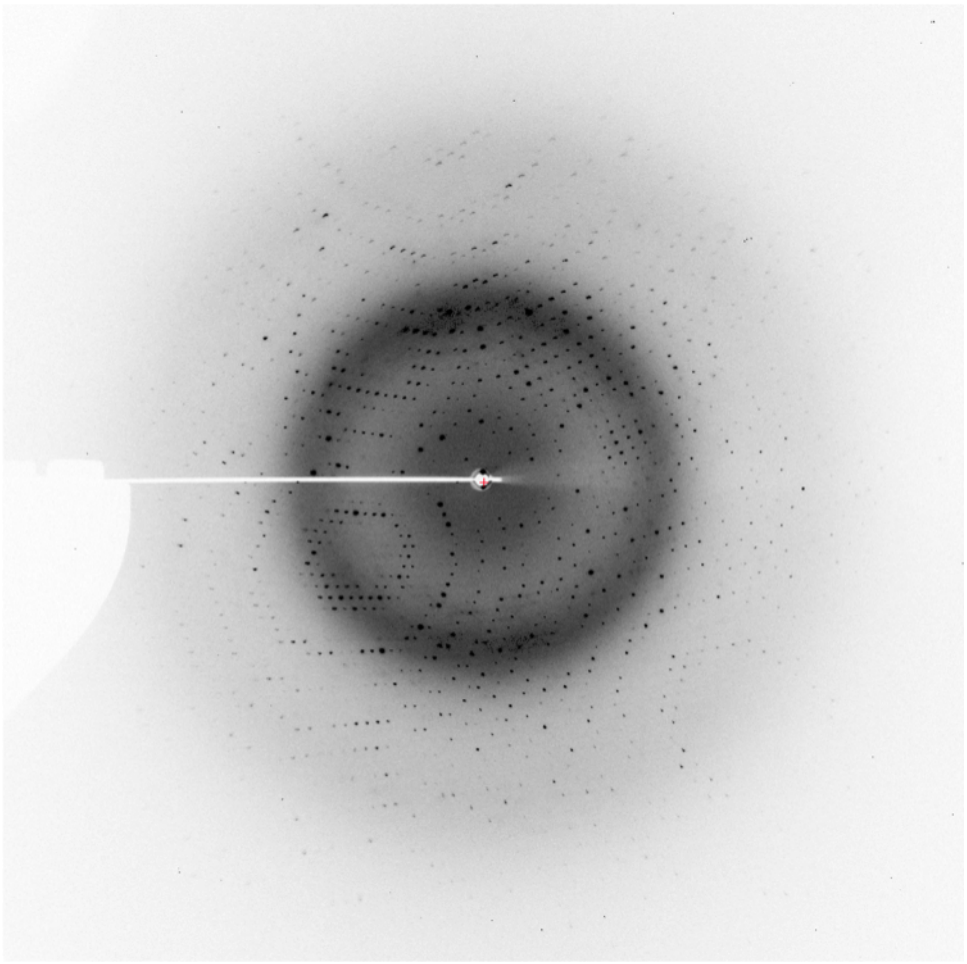


Figure 4.20 – Representative diffraction pattern of Hsp90-NTD in the presence of 10 atm pressure of xenon.

Table 4.3 – Data collection statistics.

	native data set	xenon derivative
Data collection		
Wavelength (Å)	1.54	1.54
Space group	I 2 2 2	I 2 2 2
Unit cell dimensions (a, b, c, Å)	64.8, 88.22, 100.0	65.12, 88.74, 99.79
Mosaicity (degrees)	0.26	0.18
Resolution range (Å)	66.16-1.65	66.31-1.68
Resolution range outer shell (Å)	1.68-1.65	1.71-1.68
Redundancy[a]	6.82 (4.5)	5.88 (4.08)
Total reflections[a]	236,915 (7,252)	192,091 (5,802)
Unique reflections[a]	34,725 (1,613)	32,689 (1,422)
Completeness[a](%)	99.8 (95.8)	98.7 (84.7)
R_{sym}[a,b]	3.7 (62.4)	4.3 (22.4)
$\langle\langle I\rangle/\langle\sigma(I)\rangle\rangle$[a]	28.3 (1.9)	24.9 (5.3)
Anomalous completeness[a]	99.2 (92.7)	96.6 (78.7)
Anomalous multiplicity[a]	3.5 (2.2)	3.0 (2.0)
Mid-Slope of anomalous normal probability	1.04	0.952

all data taken from *aimless* (CCP4) (Winn et al., 2011).

[a] highest resolution bin in parentheses.

[b] $R_{sym} = \sum_{hkl}\sum_i |I_i - \langle I\rangle| / \sum_{hkl}\sum_i I_i$ where I is the intensity of a reflection, and $\langle I\rangle$ is the mean intensity of all symmetry related reflections j.

Table 4.4 – Model refinement statistics.

	native structure	xenon derivative
Refinement statistics		
Molecules per asymmetric unit	1	1
Number of residues[c]	208	208
R-factor[c](%)	16.12	15.38
R_{free}[c](%)	20.58	19.25
Cruickshank's DPI [c,d](Å)	0.081	0.081
Average B-factors[e]		
Main chain atoms	19.38	18.67
Side chain atoms and waters	27.99	27.98
Average RMS B-factors[e]		
Main chain atoms	1.33	1.43
Side chain atoms and waters	2.23	2.77
Total number of atoms[f]	2,164	2,133
Total number of water molecules[f]	440	411
Solvent content[f](%)	47.45	48.65
Ramachandran plot[g]		
Core	91.4 %	89.8 %
Allowed	8.0 %	9.6 %
Generously allowed	0.5 %	0.5 %
Disallowed	0.0 %	0.0 %

[c] taken from *Refmac* version 5.6.0117 (CCP4) (Winn et al., 2011).
[d] after (Cruickshank, 1999).
[e] taken from *B-Average* (CCP4).
[f] taken from *Sfcheck* version 7.03.16 (Vaguine et al., 1999) (CCP4).
[g] taken from *Procheck* (Morris et al., 1992; Evans, 2006) (CCP4).

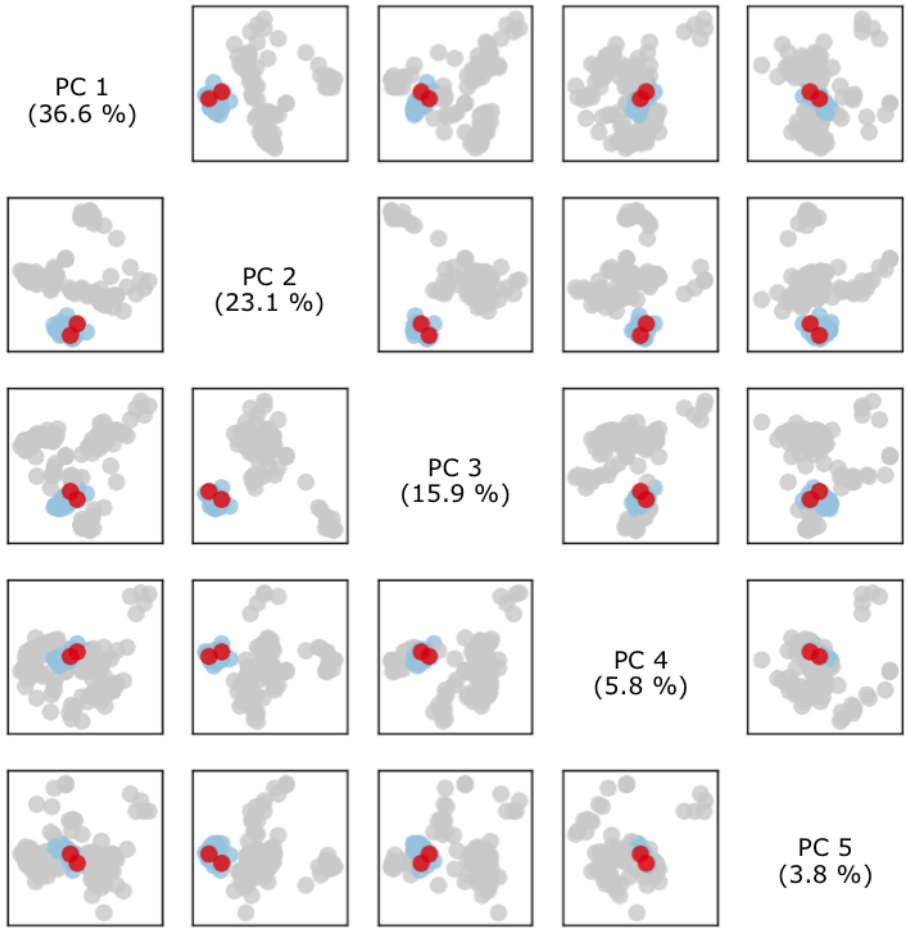


Figure 4.21 – Location of novel structures of Hsp90-NTD within the conformational space spanned by previously known protein structures. The novel structures of Hsp90-NTD in the presence and absence of xenon (red) are shown in comparison to other structures of the same conformational cluster (light blue), and other protein conformations (grey) after a PCA based conformational analysis based on the C^{α} distance profiles, as conducted above (Figure 4.6). The novel structures have been excluded from the computation of loading vectors and eigenvalues. The relative importance of the components, as per cent variance explained, is indicated in parentheses.

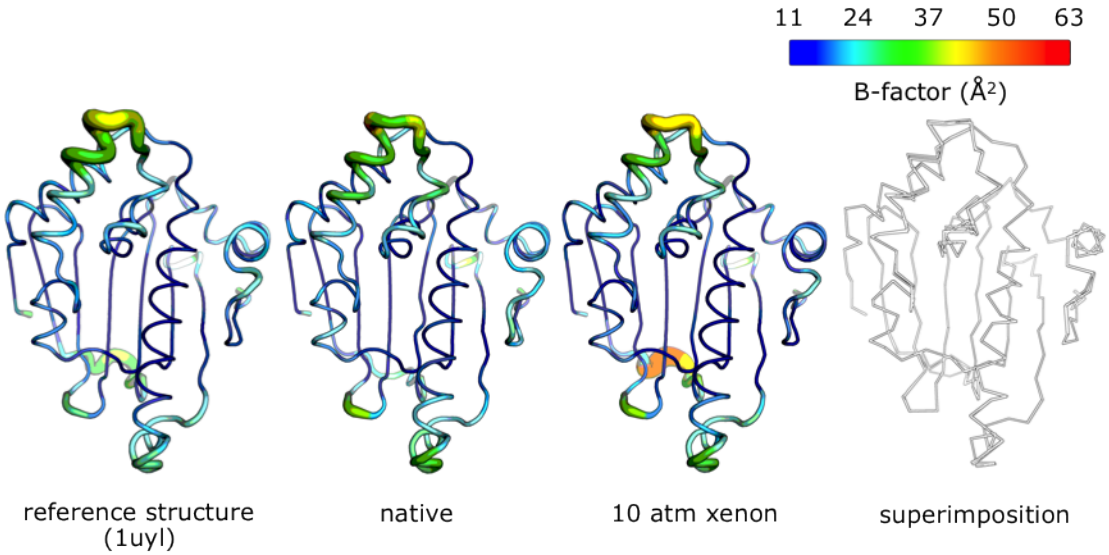


Figure 4.22 – Comparison of template and final atomic models of the native and 10 atm xenon data sets. Protein structures are depicted as simplified backbone traces, with backbone trace diameter and -colour chosen according to the experimental isotropic B-factor (see colour scale on top, in $Å^2$). Water molecules, ions and other hetero atoms are omitted from the representation for the sake of simplicity. The region with the highest B-factor (coloured red and yellow in the 10 atm xenon structure) comprises amino acids 176 to 178.

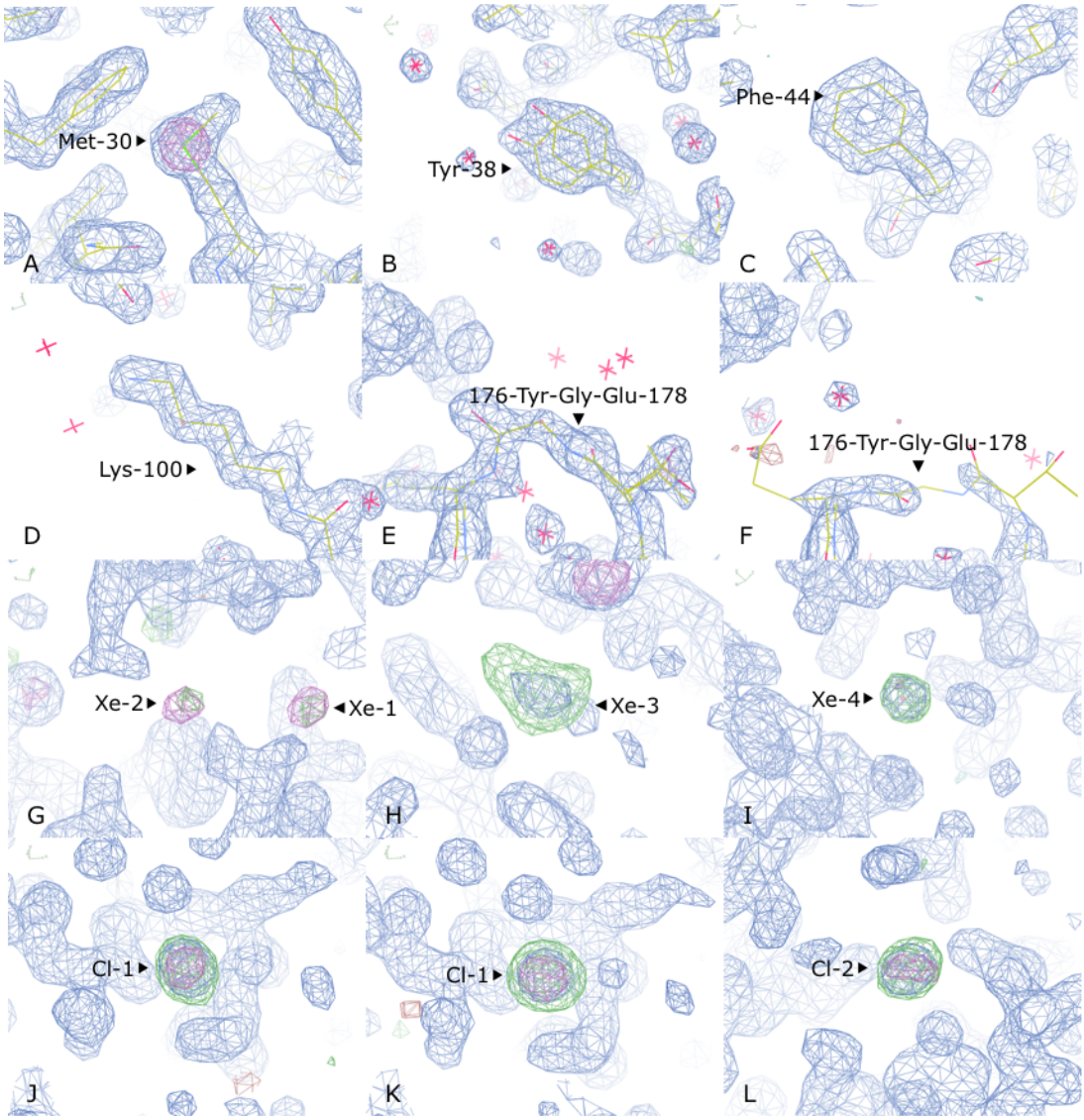


Figure 4.23 – Detailed views of solved structures of Hsp90-NTD in the presence and absence of xenon. (A) Residue 30, methionine (Met) of the xenon complexed structure. (B) Static disorder identified for residue 38, tyrosine (Tyr) of the native structure. (C) Residue 44, phenylalanine (Phe) of the xenon complexed structure. (D) Residue 100, lysine (Lys) of the native structure. (E) and (F), region comprising residues 176 to 178 (tyrosine (Tyr), glycine (Gly), and glutamic acid (Glu)) in the native and xenon complexed structures, respectively. (G) to (H), unbiased (see main text; unbiased maps were produced by omitting chloride ions and xenon atoms before a final step of automatic refinement using the Refmac5 program) electron density maps showing the positions of xenon (Xe) atoms 1 to 4. (J) and (K), unbiased electron density maps showing the position of chloride (Cl) ion 1 in the native

and xenon complexed structure, respectively; and (L), position of chloride ion 2 in the native structure. Blue maps *2Fo-Fc* are shown at a level of 1.5σ, green/red (balanced) difference maps *Fo-Fc* at level 3.5σ, and purple maps show anomalous signal at level 4σ. In panels (A) to (F), non-hydrogen protein atoms and water oxygens are shown in stick representation, coloured yellow (carbon), blue (nitrogen), or red (oxygen).

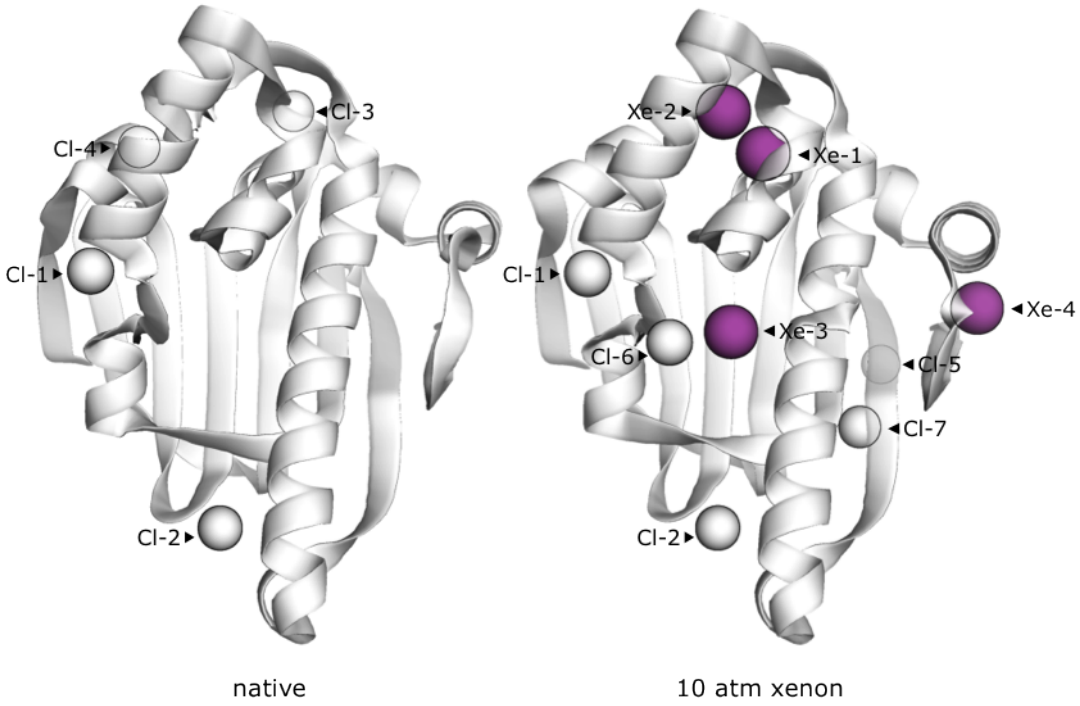


Figure 4.24 – Positions of chloride ions and xenon atoms in native and xenon derivatised Hsp90-NTD. Proteins are shown in cartoon representation with round helices, flat sheets and simplified, straight loop segments, and chloride ions (white) and xenon atoms (magenta) as spheres and are numbered. The protein is shown semi-transparent in order to allow for identification of completely or partially concealed chloride or xenon moieties.

4.3 Discussion

In this chapter, the interaction of xenon with Hsp90-NTD has been investigated. Known experimental structures spanning the conformational space available to

Table 4.5 – Anomalous signal intensity in Hsp90-NTD. Methionine (Met) residues are numbered in accordance to the UniProt amino acid reference sequence. For the numbering reference of chloride (Cl) and xenon (Xe) residues see main text and Figure 4.24.

	native structure	xenon derivative
Met^{30}[a]	12.65 (12.90)	13.18 (12.97)
Met^{98}[a]	8.34 (8.47)	11.83 (11.54)
Met^{180}[a]	6.14 (6.04)	5.56 (5.88)
Met^{119}[a]	5.42 (5.29)	5.17 (5.74)
Met^{130}[a]	4.58 (4.81)	4.37 (4.32)
Cl-1[a]	8.25 (8.23)	8.67 (8.49)
Cl-2[a]	5.71 (5.81)	5.79 (5.91))
Cl-3[a]	4.27 (4.33)	
Cl-4[a]	4.24 (4.08)	
Cl-5[a]		8.51 (8.06)
Cl-6[a]		7.07 (6.91)
Cl-7[a]		6.71 (6.53)
Xe-1[a]		7.03 (7.23)
Xe-2[a]		6.44 (6.65)
Xe-3[a]		4.08 (4.04)
Xe-4[a]		4.13 (4.44)

[a] anomalous peak intensities from the *unbiased* map are given in parentheses; both in units of electrons per cubic Å.

Table 4.6 – Debye-Waller factors of methionine sulphur atoms in Hsp90-NTD.

	native structure	xenon derivative
Met^{30} (Å^2)	15.80	14.78
Met^{98} (Å^2)	17.39	16.68
Met^{180} (Å^2)	25.10	31.03
Met^{119} (Å^2)	28.90	37.22
Met^{130} (Å^2)	40.84	39.18

Hsp90-NTD have revealed a ligand dependent conformational change of an α-helix of Hsp90-NTD. The protein conformation of the unliganded structure was utilised to predict binding sites of xenon by the xenon likeness score developed in the previous chapter, revealing several likely binding sites. Finally, the prediction has been corroborated by solving the structure of Hsp90-NTD in the presence and absence of xenon by using X-ray crystallography.

The conformational space sampled by Hsp90-NTD has been investigated at different levels of detail. Firstly, a coarse analysis has been conducted based on the variability of the secondary structure of the protein in form of a *sequence logo* analysis (Figure 4.1), indicating variance to be present in a helical region spanning residues 100 to 120. A more detailed analysis of this region and a visual analysis of superimposed protein structures (Figure 4.2) suggested the presence of discrete conformational clusters (Table 4.1). A *principal component analysis* (PCA) of the conformational variability based on internal coordinates was found to coincide well with said clustering (Figure 4.6), and highlighted the helical region mentioned above (amino acid residues 100 to 120) as the main source of conformational variance (Figures 4.8, 4.9, and Figure 4.10). The same PCA-based technique was applied to demonstrate that the process of soaking protein crystals with xenon, and the presence of xenon atoms in the protein structure does not strongly affect the global protein structure (Results section 4.2.2) of proteins other than Hsp90-NTD. The unliganded conformation of Hsp90-NTD was then used to predict where xenon was most likely to interact with the protein.

Interestingly, the helical region (amino acid residues 100 to 120) contains a glycine residue (Gly^{108}) that correspond to one of the 'hinge' residues of a 'lid' segment of the protein in the yeast orthologue (UniProt identifier P02829) that spans amino acid residues 94 to 125 (Chadli et al., 2000; Prodromou et al., 2000) (see Introduction Chapter 1 and Figure 1.4). This lid segment has been observed to undergo a 180 degrees swing from an open (free NTD) to a closed conformation in the holo structure of the full length protein containing C-terminal domains and co-chaperones (PDB 2cg9) (Ali et al., 2006). The lid closure is thought to be a functional consequence of complex formation, and to enhance the ATPase

activity carried out by the NTD during the catalytic cycle of the protein (Ali et al., 2006). A corresponding closed conformation of the human Hsp90 protein (UniProt P07900) has not been observed to date as in constrast to structures of Hsp90-NTD atomic structures of the full protein are not available. However, the amino acid sequence of the lid segment is conserved between yeast and human,[15]

```
 94-GTIAKSGTKAFMEALsAGADvSMIGQFGVGFY-125 yeast, UniProt P02829
108-GTIAKSGTKAFMEALqAGADiSMIGQFGVGFY-139 human, UniProt P07900
```

and the human protein segment (amino acid residues 108 to 139) corresponds precisely to the protein region where most conformational variability was found to be present in the PCA analysis (Figures 4.8 and 4.9). It is therefore tempting to speculate that human Hsp90-NTD undergoes a similar conformational change to the lid closure of the yeast orthologue in the full length enzyme, and that the crystal structures of human Hsp90-NTD capture some of the conformational variability the enzyme possesses in solution. Ligand-induced conformational changes of the helical segment have been discussed in the literature before (Wright et al., 2004; Roughley et al., 2012). They result in line broadening of resonances for residues 105 to 121 in NMR experiments (Huth et al., 2007).

Using the positions of water molecules and ions from 42 crystal structures of Hsp90-NTD, the xenon likeness score developed in the previous chapter was used to score and rank potential xenon positions. A number of protein moieties displayed a favourable xenon interaction potential (Figure 4.12), their xenon likeness score comparing well with three score thresholds determined previously. Inspected in a structural context, four clusters of well scoring positions were emphasised (Figures 4.13, 4.14, 4.15, and 4.16): one overlapping with the protein ligand binding site (termed '1'), a second one buried in the hydrophobic core of the protein where there are conserved water molecules (termed '2a' and '2b'), a third one close to the entrance of the ligand binding cavity (termed '3'), and a last one distant from that (termed '4'). Sites 2a and 2b were scored more favourably than the other

[15] with identical residues in capitals.

binding sites, and were found to be sterically defined more tightly (Table 4.2). Two conserved water molecules, one at each site, were found to be present in 41 of the 42 structures, their coordinates showing little variance across all structures. A water network within the active site of the enzyme, consisting of 107 individual water molecules across the 42 structures, obtained generally favourable xenon likeness scores.

4.3.1 Crystal structures of native and derivatised Hsp90-NTD

Crystals of Hsp90-NTD were found to belong to the space group most commonly observed for the protein before, I 2 2 2, and to diffract to high resolution both in the presence and absence of xenon (1.68 Å and 1.65 Å, respectively; see Table 4.3). The structure of the protein in the presence and absence of xenon was solved, using the previously known structure with PDB identifier 1uyl as a template for molecular replacement. The novel structures were found to be consistent with the template, most importantly in the divergent, helical region mentioned above. Furthermore, the novel structures fall into the same conformational space occupied by the protein structure cluster containing the unliganded forms of Hsp90-NTD (Figure 4.21). Excluding a region with missing electron density in one of the novel structures (Figure 4.23, panels (E) and (F)) and high experimental B-factor (Figure 4.22) comprising of amino acid residues 176 to 178, the mutual pairwise backbone C^{α} RMSD between the targets and the template is between 0.11 and 0.20 Å. The resolution of the novel structures is sufficiently high to identify structural details such as static disorder of individual residues and the π-electron cloud shape of phenylalanine residues (Figure 4.23). The stereochemical quality of the structure is excellent, with 89.8 to 91.4 % of residues in the most favourable regions of a Ramachandran plot (Table 4.4). The atomic structures could be refined to good quality without indication of over-fitting of the data, as indicated by R-factor and R_{free} of 16.12/20.58 and 15.38/19.25 % for the native and xenon binding structure, respectively. The coordinate error estimate (Cruickshank's DPI) (Cruickshank, 1999) was found to be very low (0.08 Å in both cases), allowing for the determination of the positions of individual atoms with high precision and

accuracy.

The mid-slope of anomalous normal probability (1.04 for the native data set and 0.95 for the xenon soaked protein data set, see Table 4.3) indicates the presence of anomalous signal in the two data sets. The sulphur atoms of all five methionine residues present in the protein were found to possess strong anomalous signal (up to $> 13\sigma$, see Table 4.5), and the signal intensity was found to be well correlated with the B-factor of that moiety (Table 4.6), in a negative fashion. Furthermore, the anomalous signal allowed for identification of individual chloride ions (see Figure 4.24 and Table 4.5), as well as four xenon atoms in case of the protein soaked with 10 atm of xenon pressure. For detection of xenon and chloride moieties, an anomalous signal cutoff of 4σ was applied.

4.3.2 Xenon concentrations during the derivatisation process

For further discussion, it is useful to estimate the local molar xenon concentration for the structure having been solved in the presence of 10 atm of xenon gas, assuming xenon to be a perfect gas, and to partition freely within the crystal. The *ideal gas law* states $PV = nRT$, where P is the absolute pressure, V is the volume of gas, n is the chemical amount of gas (in moles), and T is the thermodynamic temperature (in Kelvin), and R the gas constant. The size of the protein crystal unit cell determined experimentally gives the volume under consideration, 66.4 Å $\times$ 88.22 Å $\times$ 100.0 Å $= (66.4 \times 88.22 \times 100.0) \times 10^{-30}$ m^3 $= 5.857808 \times 10^{-25}$ m^3. The gas constant is $R = 8.205746 \times 10^{-5}$ m^3 atm K^{-1} mol^{-1}, which gives $N = \frac{PV}{RT}$ with $N = n/(6.0221412927 \times 10^{23}$ mol$^{-1})$ the number of atoms, and thus

$$N = P \times \frac{5.857808 \times 10^{-25} \text{ m}^3}{8.205746 \times 10^{-5} \text{ m}^3 \text{ atm K}^{-1} \text{ mol}^{-1} \times T} \times 6.0221412927 \times 10^{23} \text{ mol}^{-1}, \text{or}$$

$$N = P \times \text{ atm}^{-1} \times \frac{5.857808 \times 6.0221412927}{8.205746} \times 10^3 \times T^{-1} \text{ K}$$

$$= 4.299 \times 10^3 \times P \times \text{ atm}^{-1} \times T^{-1} \times \text{K}.$$

Therefore, at various temperatures of biological interest (273 K, 293 K, and 298 K) an average number of

$$N_{273\mathrm{K}} = 15.75 \times P \times \ \mathrm{atm}^{-1},$$
$$N_{293\mathrm{K}} = 14.67 \times P \times \ \mathrm{atm}^{-1}, \text{ and}$$
$$N_{298\mathrm{K}} = 14.43 \times P \times \ \mathrm{atm}^{-1}$$

xenon atoms per protein crystal unit cell volume can be expected, respectively. Using $V = 5.857808 \times 10^{-25}$ m^3 $= 5.857808 \times 10^{-22}$ L it can be seen that $n_0 = 5.857808 \times 10^{-22} \times 6.0221412927 \times 10^{23} = 352.76$ xenon atoms per unit cell volume are required to achieve a concentration of 1 M = 1 mol L^{-1}, hence xenon molar concentrations at the different temperatures are[16]

$$\mathrm{conc.}_{273\mathrm{K}} = 0.04464 \times P \times \ \mathrm{atm}^{-1} \times \ \mathrm{mol\ L}^{-1},$$
$$\mathrm{conc.}_{293\mathrm{K}} = 0.04159 \times P \times \ \mathrm{atm}^{-1} \times \ \mathrm{mol\ L}^{-1}, \text{ and}$$
$$\mathrm{conc.}_{298\mathrm{K}} = 0.04089 \times P \times \ \mathrm{atm}^{-1} \times \ \mathrm{mol\ L}^{-1}.$$

In other words, xenon pressures of 22.4 atm (273 K), 24.04 atm (293 K), or 24.45 atm (298 K) are required for a xenon concentration of 1 M, and at the xenon pressure used for this experiment (10 atm) about 409 to 446 mM of xenon were present, a number well within concentration ranges routinely used in ligand soaking experiments in fragment based drug design (see Introduction, section 1.1).

4.3.3 Experimental support for binding site predictions

Comparing the predictions to the experimental verification of xenon binding sites, a good general agreement was found, with false positive and false negative results to be discussed in further detail.

Firstly, at the predicted xenon binding sites 2a and 2b, which had been scoring most favourably in the prediction (Table 4.2) at scores of −42.4 and −40.6 (average values over 41 protein structures, corresponding to *a priori* xenon binding probabilities of about 90.6 and 88.3 %, respectively), no anomalous signal indicating

[16] the derivation of concentrations is obviously *independent* of the unit cell dimensions.

the presence of xenon was observed, but rather two water molecules corresponding to those present in 41 out of 42 structures of Hsp90-NTD having low B-factors of 17.11 and 14.38 Å^2, respectively. These water molecules possess xenon likeness scores of about -42.7 and -40.9 in the structural context of the newly solved structure of Hsp90-NTD in the presence of xenon and are thus fully consistent with the ensemble of structures used for prediction (Table 4.8), yet they seem not be replaced by xenon atoms.

A possible explanation of this finding is the tight confinement imposed onto the water molecules found exerted by the densely packed surrounding protein atoms: close to the putative xenon positions, non-hydrogen protein atoms are found, partially overlapping with xenon (Table 4.7). In a data set of 470 known xenon positions within the PDB, less than 1 % of xenon atoms were found with the same amount of non-hydrogen protein atoms at an equal or closer distance. Thus, the potential binding site seems to be too tight for the binding of xenon without structural rearrangements within the protein. This finding is supported by the small radius of gyration of the ensemble of water molecules found at these positions (0.20 to 0.21 Å, see Table 4.2).

Table 4.7, however, shows that for two of the actual xenon atoms found to bind to Hsp90-NTD (denoted Xe-1 and Xe-2 in Figure 4.24), a small number of overlapping non-hydrogen protein atoms seem to be tolerated. In agreement with this finding, in the data set of 470 xenon positions in the PDB there are a number of xenon atoms which seemingly overlap with a few non-hydrogen protein atoms (Figure 4.25, left panel), and the radial distribution function (rdf) of xenon to *any* non-hydrogen protein atom, corrected for the difference across the protein atom types by subtracting their van der Waals radius (see Table 3.2 and Figure 4.25, right panel), shows variance, and importantly, non-zero function values, to be present in the area of the rdf at less than xenon's apparent van der Waals raidus (2.16 Å). This indicates that at some positions within the protein core, xenon atoms can indeed be tolerated, even if they partially overlap with non-hydrogen protein atoms. This can most likely be attributed to the fact that xenon possesses a large electron cloud consisting of 54 electrons and is thus strongly polarisable by surrounding electronegative groups such as protein aromatic residues, effectively changing its apparent molecular shape.

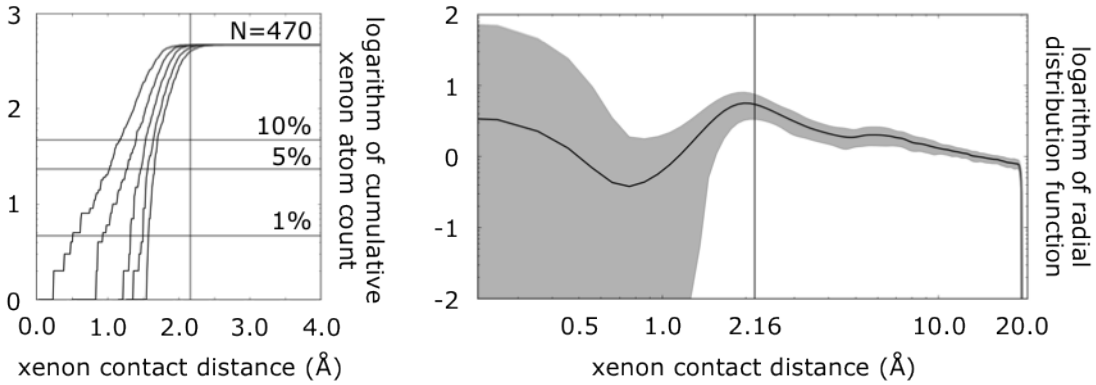


Figure 4.25 – Fuzziness of xenon contact parameters. Left panel, cumulative histogram of xenon contact distances to their five nearest non-hydrogen protein atoms (black lines from left to right) in a data set of 470 xenon atoms from the PDB. The contact distance is defined as the distance between the centres of mass of the two atoms minus the van der Waals distance of the respective protein atom, as indicated in Table 3.2. Horizontal lines indicate the location of 1, 5, and 10 % of the data, the vertical line the suggested xenon van der Waals radius of 2.16 Å. Right panel, radial distribution function of xenon to any non-hydrogen protein atom, calculated as described in the previous chapter; similarly to the left panel, inter-nuclear distances have been corrected for the suggested protein atom radius by subtracting it. Individual rdfs have been calculated for all 470 xenon atoms up to a distance of 20 Å with distance bins of width 0.25 Å, smoothed by convoluting them twice with a Gaussian kernel (see averaging procedure of rdfs during the hierarchical clustering in the previous chapter, Results section 3.2.3), and averaged. The average rdf is shown as a black line, and the area between the average rdf plus/minus the corresponding standard deviation is shaded grey.

Table 4.7 – Tight packing around putative xenon binding sites in Hsp90-NTD. The distance of the moiety to the i-th closest non-hydrogen protein atom (in Å), minus the van der Waals radius of the respective protein atom as suggested by the CHARMM22 force field (see Table 3.2), is indicated for the ten closest protein atoms. For comparison, the fraction of non-hydrogen protein atoms closer than that distance in any known xenon binding protein are given in parentheses (in per cent).

i	cluster 2a	cluster 2b	Xe-1	Xe-2
1	1.05 (6.0)	0.88 (3.0)	0.16 (0.2)	−0.13 (0.0)
2	1.16 (3.0)	0.90 (0.9)	1.36 (8.1)	0.54 (0.2)
3	1.31 (0.6)	1.15 (0.2)	1.66 (23.2)	0.79 (0.0)
4	1.43 (0.4)	1.22 (0.0)	1.72 (18.6)	1.13 (0.0)
5	1.50 (0.2)	1.42 (0.2)	1.83 (23.7)	1.69 (8.7)
6	1.52 (0.2)	1.60 (0.2)	1.85 (17.7)	1.88 (21.1)
7	1.68 (1.1)	1.62 (0.2)	1.88 (14.1)	1.99 (26.9)
8	1.77 (1.7)	1.70 (0.9)	1.90 (9.8)	1.99 (20.0)
9	1.80 (1.5)	1.74 (0.6)	1.92 (5.5)	2.00 (14.7)
10	1.82 (0.9)	1.81 (0.9)	1.94 (3.4)	2.01 (10.0)

An alternative explanation for xenon atoms being found in this seemingly prohibitively tightly packed environment is the presence of a second, minor protein conformation which would allow for easier accommodation of the xenon atoms, but which would not have been detected by the structure determination. Comparing the strength of the anomalous signal for the xenon atoms (around 6.44 to 7.03 electrons per cubic Å, see Table 4.5) to those of the most prominent peaks of anomalous density for methionine sulphur atoms found in the data set (around 11.83 to 13.18 electrons per cubic Å), the occupancy of the two xenon atoms can roughly be estimated to be below 10 %, as xenon possesses 54 electrons and sulphur, 16, a factor of 4.5, and about half the anomalous signal, giving an overall occupancy ratio of about 1:9, assuming full occupancy of the methionine sulphur atoms. Thus, if a minor protein conformation was indeed present in the crystal in about 10 % of instances, each binding xenon at full occupancy, it could evade detection in the *2Fo-Fc* electron density map, a problem potentially aggravated by the build up of the model bias. Furthermore, the experimental resolution contributes to the ability to detect minor conformations. While a minor conformation

with < 10 % occupancy might be detected at atomic resolution, it will most likely evade detection at the resolution range of the data presented (1.65 to 1.68 Å), translating to a DPI of 0.08 Å (see above). For reference, the smallest occupancy values fitted for any residue in the data set are minor conformations present at 20 % of residues Tyr^{38}, Tyr^{160}, Met^{98}, and Met^{180} of the xenon derivative of Hsp90-NTD.

In both cases, as no water molecule was found at the corresponding position in the set of 42 Hsp90-NTD structures used for prediction, the position of these two xenon atoms could not have been predicted with a method scanning experimental water molecule positions, but would have required an alternative approach (see below, and Discussion section of the previous chapter).

Next, the detection of xenon atom Xe-3 in the binding site of the protein corresponds well with the predicted cluster 1, where a somewhat diffuse xenon binding site ($R_{gyr} = 3.64$ Å) with medium affinity (mean xenon likeness score of -29.3 across 124 water positions in 42 protein structures) was predicted. This coincides well with the relatively low anomalous signal for atom Xe-3 (about 4.08 electrons per cubic Å, or, following the argument above, about 5 % occupancy). This can most likely be attributed to the presence of a tight water network within the binding site[17] that would be energetically unfavourable to displace.
Out of the two remaining predicted xenon binding sites (clusters 3 and 4 in Table 4.2), displaying medium xenon likeness scores of -28.6 and -26.1, respectively, a xenon atom was found to be present only for the latter, with anomalous signal of similar strength to the xenon atom detected in the binding site. It should be noted that according to the probabilistic interpretation of the xenon likeness score described in the previous chapter (Results section 3.2.6), a xenon likeness score between -25.0 and -30.0 translates to an *a priori* probability of xenon binding between 26.5 and 47.5 %, thus a score in this region does not guarantee xenon binding but only suggests that possibility.

Table 4.8 summarises the predictions and validated occurrences of xenon binding to Hsp90-NTD, underlining the usefulness of the approach as well as the robustness with respect to small perturbations of the structural environment of potential xenon binding sites, by pointing out the consistency of the score values between

[17] indeed, some of the binding site water molecules display B-factors below 15 $Å^2$.

protein structures slightly differing from each others.

Table 4.8 – Comparison of xenon binding predictions with the experimental validation. For each of the predicted xenon binding sites in Hsp90-NTD and all xenon atoms detected in the novel structure of the protein solved in the presence of 10 atm xenon, xenon likeness scores previously predicted are compared to scores finally observed in the novel structure, and it is indicated whether xenon atoms were or were not observed at that position.

	score prediction[a]	score validation	xenon found?
cluster 2a	−42.4	−47.2	no
cluster 2b	−40.6	−40.9	no
cluster 3	−28.6	−29.5	no
Xe-1	-[b]	−44.8	yes
Xe-2	-[b]	−36.3	yes
Xe-3 (= cluster 1)	−29.3	−31.1	yes
Xe-4 (= cluster 4)	−26.1	−26.0	yes

[a] mean value of predictions (compare to Table 4.2).
[b] no water molecules were found at corresponding positions in any of the structures, thus no prediction had been made before the experiment.

In summary, the validation of the xenon likeness score on Hsp90-NTD as a model system has shown the method to be a useful tool in order to identify xenon binding sites in proteins. The method has successfully identified the protein ligand binding site as a potential binding site where a low occupancy xenon atom has been identified. A further xenon atom has been suggested to bind to a site distal to the ligand binding site and could be detected experimentally as well. A small number of false positive and false negative predictions have been discussed in the context of tight atom packing within the protein core, or inadequate sampling by the method in using only water molecule and ion positions as xenon binding site suggestions. These shortcomings can be overcome by addressing the representation of xenon polarisability and protein plasticity in an adequate manner as well as employing a more exhaustive sampling method for the suggestion of putative binding sites that then can be scored by the knowledge-based method, e.g. grid-based or molecular mechanics-based approaches.

4.3.4 Outlook

In future work, the method will be validated by additional model systems. Predictions of xenon binding have been made for two additional protein systems available for rapid testing at EMBL Outstation Hamburg (Germany): bovine trypsin (UniProt identifier P00760), and lipase B from *Candida antarctica* (UniProt P41365) (Figure 4.26), and predictions for other protein systems are in preparation.

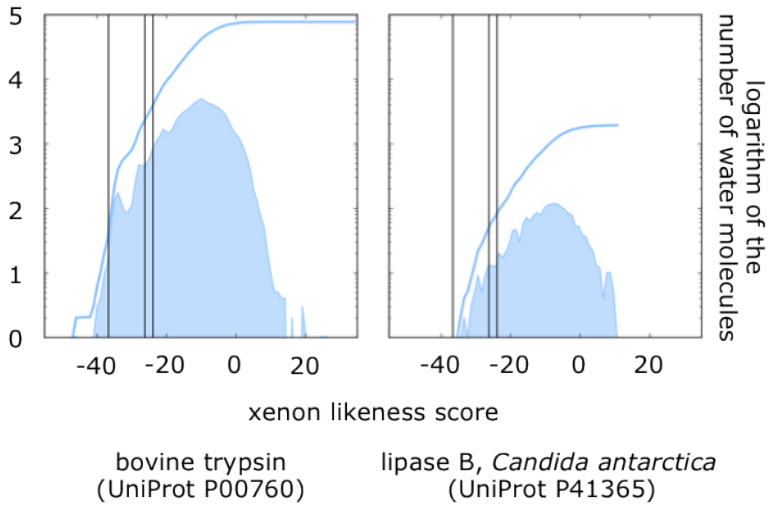


Figure 4.26 – Prediction of xenon binding to two additional model systems, xenon likeness score distributions. Water and ion positions within structures of two model systems were scored and scores represented as (cumulative) histograms. A bin width of 1 score unit was used (with scores s_0 falling into the i-th bin if $i \leq s_0 < i + 1$). Possible score threshold values from a threshold scanning analysis (-36.6, -26.1, and -23.8, see Results section 3.2.5 in the previous chapter) are represented as vertical black lines. The two test systems from left to right are bovine trypsin (76,533 water/ion positions in 337 structure records, minimum score -47.3, maximum score 34.0, mean score -10.6, standard deviation 7.22, median score -10.0), and lipase B from *Candida antarctica* (1,913 positions, 7 structures, minimum -35.7, maximum 10.3, mean -9.0, standard deviation 7.50, median -8.4).

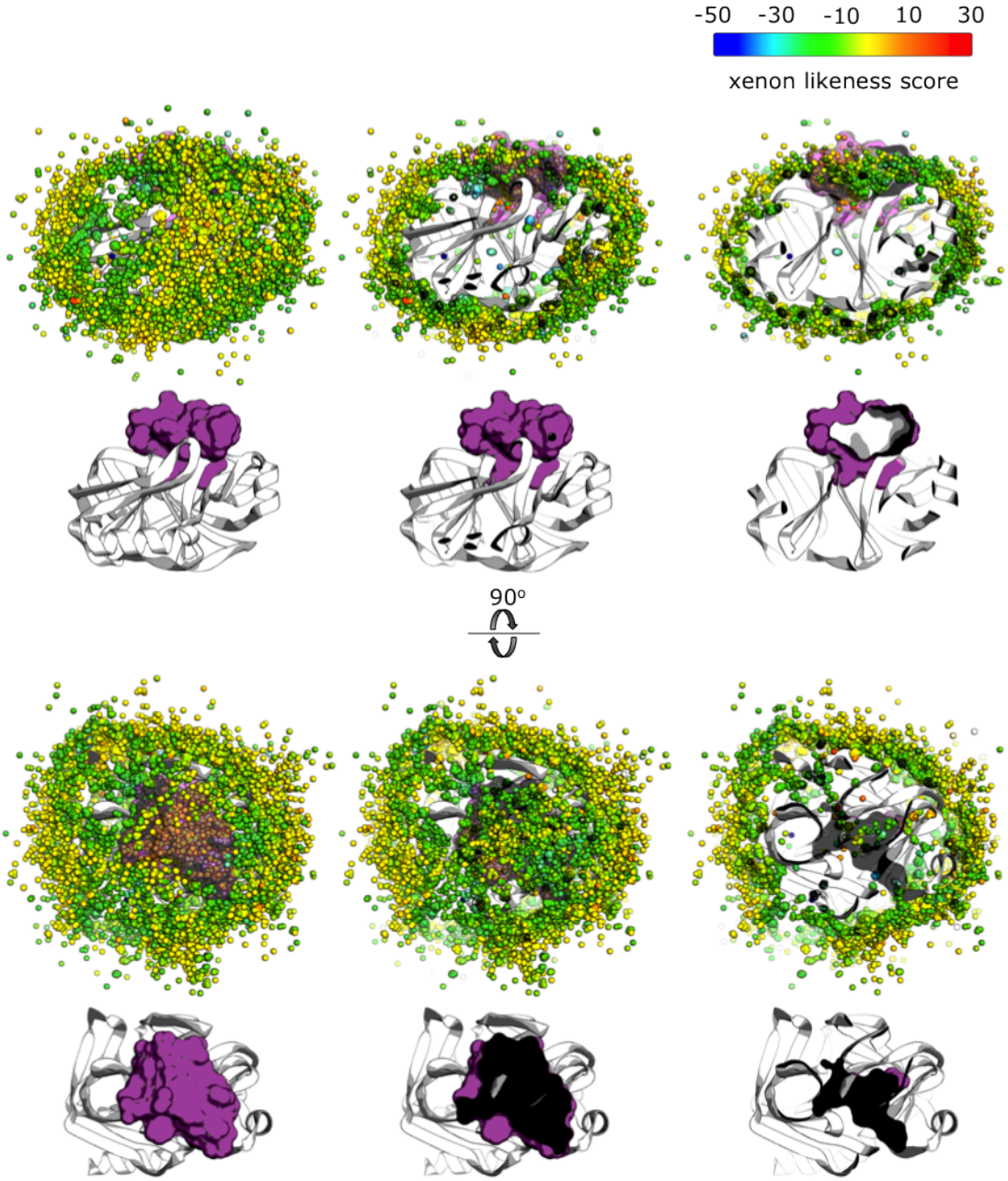


Figure 4.27 – Prediction of xenon binding to bovine trypsin (UniProt P00760). In the first and third row, xenon likeness scores of individual water molecules and ion positions (circles) are coloured on a rainbow scale from a xenon likeness score from −50 to −30 (favourable to unfavourable, blue to red, see top of figure). The ligand binding site of the protein is represented as a semi-transparent purple surface containing all organic molecule ligands present in the experimental structures. A

protein structure representative (PDB identifier 1auj, chain A) (Lee et al., 1997) of the conformational ensemble of structures is depicted in cartoon representation with simplified loop regions. In the second and fourth row, the protein cartoon representation (white) and location of the ligand binding site (purple surface) is given for orientation. Three different sections through the structures are shown from left to right. The bottom of the ligand binding pocket is defined by the residues of the *catalytic triad*, His^{57}, Asp^{102}, and Ser^{195} (amino acid side chains not shown) (Polgar, 2005).

A first analysis of the binding predictions reveals that they reflect the substrate specificities of the diverse model system proteins. Trypsin being a serine protease and exerting its biochemical activity, protein cleavage, through a *catalytic triad* of three residues His^{57}, Asp^{102}, and Ser^{195} at the bottom of its active site (Figure 4.27) (Polgar, 2005), targeting substrate carboxyl groups, receives medium-favourable xenon likeness scores for probe positions located in the substrate binding site of the protein while moieties close to the charged catalytic triad receive unfavourable xenon likeness scores. The positions scored most favourably are located distal to the catalytic triad at the entrance of the S1 catalytic pocket, and buried within the hydrophobic core of the protein (Figure 4.27). Although based on a small number of protein structures (7), the prediction of xenon binding to lipase B (Figure 4.28) reveals differential xenon likeness scores for the main ligand binding site (moderately favourable, located in the centre of the protein representation in Figure 4.28) interacting predominantly with hydrophobic, lipid ligands, and unfavourable scores for the smaller glycosylation site (located left in Figure 4.28). The finding that the scoring method is capable of discriminating between these differential sites is noteworthy even in the absence of experimental validation.

4.4 Methods

4.4.1 Principal component analysis

Principal Component Analysis (PCA) was carried out by decomposing the covariance matrix of the sample data using the *linalg.eig* routine of the *NumPy* (Asc) (version 1.5.1rc1) scientific computing package for Python (version 2.6.6), which provides an interface to the LAPACK (Anderson et al., 1999) routine *dgeev* for

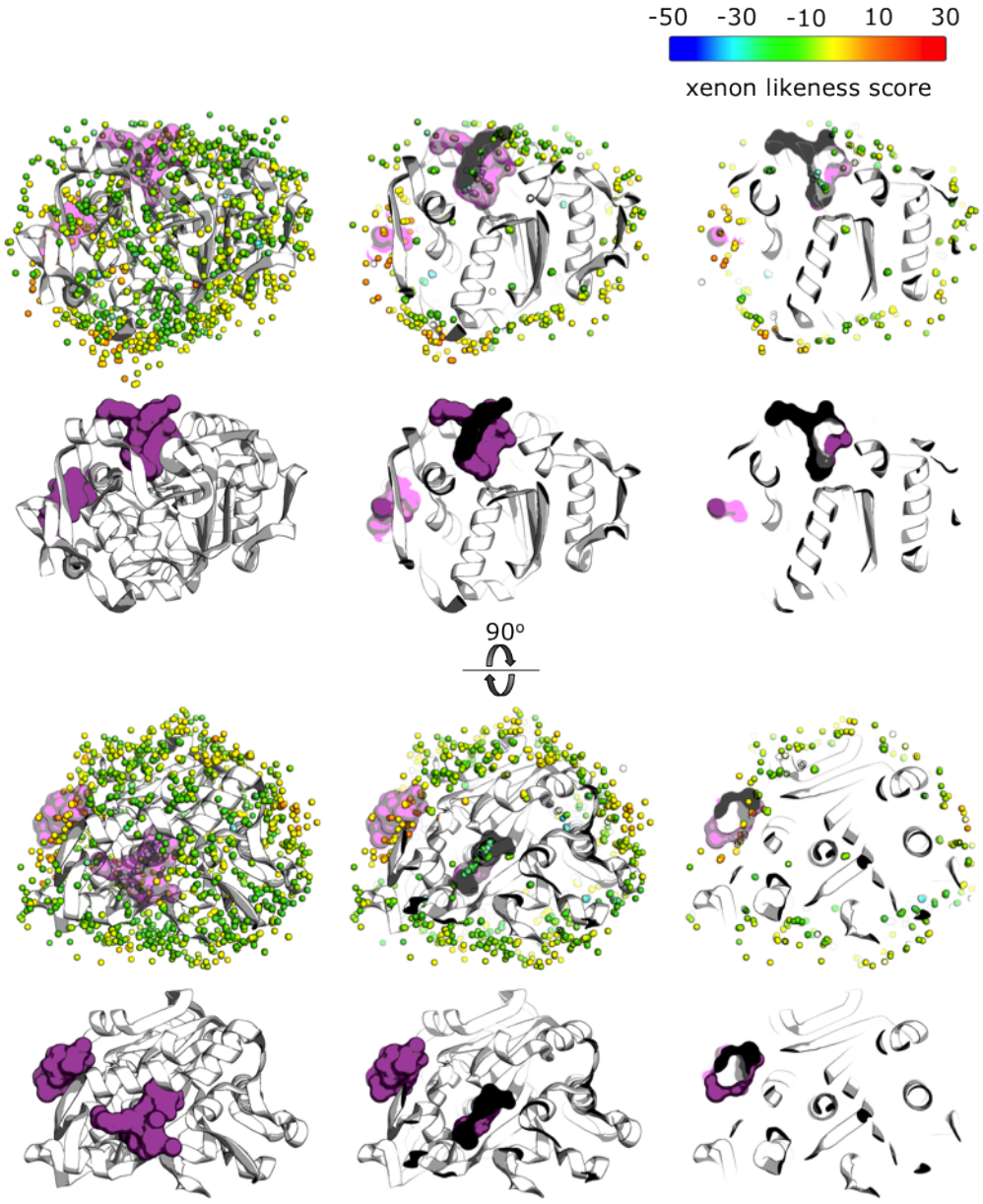


Figure 4.28 – Prediction of xenon binding to lipase B from *Candida antarctica* (UniProt P41365). See legend of Figure 4.27 for details. The representative protein structure depicted in cartoon representation (white) has the PDB identifier 1lbs, chain A (Uppenberg et al., 1995).

eigenvalue and -vector computation of real-valued square arrays.

4.4.2 Crystallisation of Hsp90-NTD

Crystals of Hsp90-NTD were a kind gift of Prof. Roderick Hubbard and Drs. Ben Davis and Allan Surgenor of Vernalis (R&D) Ltd.[18] and obtained using a hanging-drop vapor diffusion technique at 4 degrees Celsius overnight and a buffer solution containing 0.1 M Sodium Cacodylate (pH 6.5), 0.2 M $MgCl_2$, and 25 % PEG 2K MME, as previously described (Wright et al., 2004; Barril et al., 2005), and stored at 4 degrees Celsius. For the cryo buffer used for data collection, the PEG concentration was increased to 35 %. For transport and handling, crystals were transferred to sitting drop plates.

4.4.3 Data collection

Diffraction data were collected on single crystals of Hsp90-NTD at a temperature of 100 K using synchrotron radiation from the X12 beam line, equipped with a MAR-CCD 225 detector, at the DORIS storage ring at the EMBL Outstation Hamburg (Germany), c/o DESY, in collaboration with Dr. Michele Cianci. The beam size typically was about (200H $\times$ 200V) μm^2. A wavelength of 1.54 Å was chosen in order to enhance the anomalous signal for the expected xenon atoms and to allow the necessary experimental resolution. At an X-ray energy of 8 keV the observable value of the observable scattering coefficient f'' of xenon is 7.427370 e^- (see Figure 4.17).

The crystals were scooped up in a cryo loop and either directly plunged into liquid nitrogen for cooling,[19] or rapidly exposed to the target xenon pressure using a Hampton Research Xenon Chamber (Figure 4.29), incubated for around 15 to 20 minutes, and then immediately plunged into liquid nitrogen. The frozen crystals were then transferred to the beam line and maintained at 100 K with a cold nitrogen stream during data collection.

The data collection strategy was optimised using the program BEST (Popov and Bourenkov, 2003). Data were processed using the program XDS (Kabsch,

[18] Vernalis (R&D) Ltd., Granta Park, Abington, Cambridge (United Kingdom).
[19] in case of collection of the native data set.

2010) and scaled with AIMLESS (Evans, 2006). The crystals were found to belong to the space group I 2 2 2. The exact unit cell dimensions and diffraction limits can be found in Table 4.3. The asymmetric unit consists of one single Hsp90-NTD molecule and a solvent content of around 50 % (for details see Table 4.4). The data showed an anomalous signal with Δ_{anom} correlation between half sets of 0.067 (0.013), −0.093 (−0.041) for the native data set and data set in the presence of 10 atm of xenon, respectively, with the values for the highest resolution bin in parentheses. The mid-slope of anomalous normal probability was found to be 1.04, 0.952 respectively. The data collection parameters and data processing statistics are reported in Tables 4.3 and 4.9.

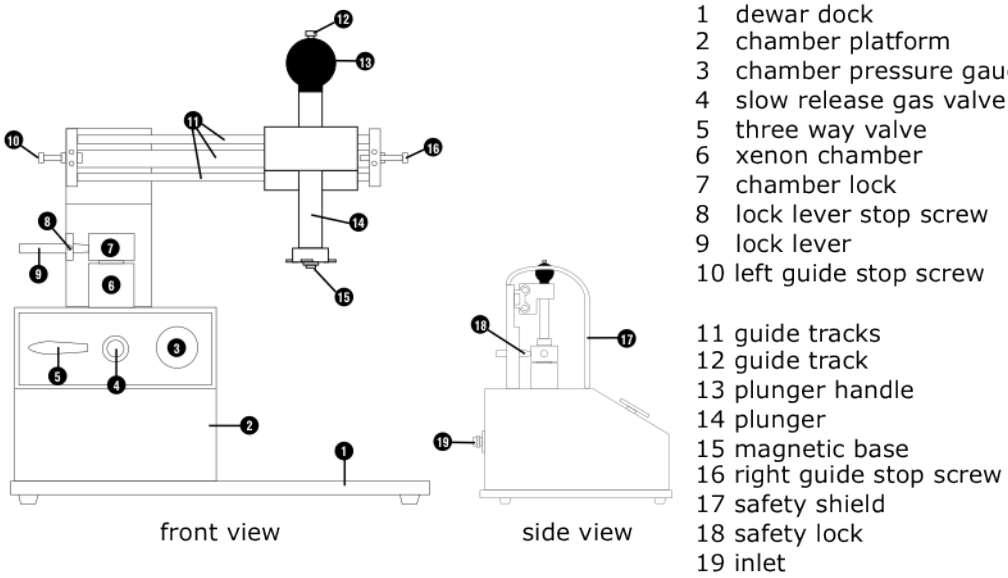


Figure 4.29 – Schematic representation of a xenon pressure chamber.

4.4.4 Structure refinement

The structure of Hsp90-NTD was refined using Refmac5 version 5.6.0117 (Murshudov et al., 1997) in the CCP4 suit of programs, version 6.2.0 (Winn et al., 2011). Protein regions displaying conformations not matching the experimental

Table 4.9 – Data collection parameters.

	apo	10 atm
segments	1	2
exposure time (seconds)	15	20 / 42
oscillation range (degrees)	0.5	0.6 / 0.45
number of images	360	150 / 300
detector distance (mm)	75	84

electron density map were rebuilt manually utilising the program COOT, version 0.6.2 (Emsley et al., 2010). Randomly selected reflections (5 % of the total) were used as an R_{free} set for cross validation. No restraints were imposed on the xenon atom distances. Using isotropic temperature factors the refinement converged to a final R-factor/R_{free} value of 16.12/20.58, and 15.38/19.25, respectively, for data sets of the native Hsp90-NTD and xenon derivative at 10 atm. The stereochemistry of the final model was routinely checked using COOT (Emsley et al., 2010) and PROCHECK (Morris et al., 1992; Evans, 2006). The refinement statistics are reported in Table 4.4. The refined crystallographic coordinates and structure factor amplitudes will be deposited in the Protein Data Bank (Berman et al., 2000).

A molecular replacement method was used to obtain an initial model and phases starting from the published apo structure of Hsp90-NTD (PDB identifier 1uyl) (Wright et al., 2004) and conducting a *rigid body refinement* using the Refmac program. The structure refinement was then carried out in cycles of alternating (1) automated refinement using the Refmac program, (2) combining the refined phases and original experimental reflection data using the CAD program in the CCP4 suit of programs (Winn et al., 2011) in order to project anomalous maps, and (3) manual inspection and modification of the atomic structural model against the combined reflection data using COOT. Each cycle would result in an improved set of phases and amplitudes as well as an improved structural protein model as measured by the R-factor and R_{free}. In step (1), Refmac was used to simultaneously improve the structural model and phases back-calculated from the current structural model and the original experimental reflection data.

A restrained refinement using no prior phase information was carried out with ten iterations of maximum likelihood refinement on the full resolution range, using automatic weighting and isotropic temperature factors as well as the simple scaling option determined from the *working* set of reflections, experimental sigma values and calculating the contribution of the solvent region. Since both the native data set and the data set collected from the xenon derivative possess relatively high resolution (1.65 Å and 1.68 Å, respectively), hydrogen coordinates were generated during the automated refinement step, but not transferred between refinement steps or -cycles, or into structural models used for further analysis. In initial cycles of refinement, the *findwaters* procedure of the COOT plugin to Refmac was used to place water molecules at electron density peaks above 2σ, and deleting those water molecules at levels $< 1\sigma$. Later on in the refinement, this step was omitted, and water positions were exclusively checked manually and modified using COOT. In step (2), the calculated phases obtained by Refmac were combined with the original anomalous maps, and the resulting combined Fourier maps were used in step (3) together with the structural atomic model resulting from step (1) in COOT for manual modification. The atomic model was loaded together with the native electron density map *2Fo-Fc*, showing both modeled and unmodeled parts at full density, which was routinely inspected at a level of 1.5σ. It was aimed to reconcile the structural model with the balanced difference map *Fo-Fc* between the electron density observed and that expected from the structural model, evaluated at 3.5σ, as well as stereochemical and geometric parameters probed with COOT and PROCHECK (Evans, 2006). Anomalous electron density using phases PHWT and amplitudes DANO were displayed and evaluated at 4σ. The modified structural model, together with the original experimental reflection data, would enter a new cycle of refinement (1), data combination (2), and manual adjustment (3).

4.5 Summary

In this chapter, the method for prediction of xenon binding sites in proteins developed in the previous chapter has been validated in a prospective manner by studying the interaction of xenon and the N-terminal domain of human Hsp90-α

(Hsp90-NTD). The conformational space sampled by known structures of Hsp90-NTD has been visualised by a novel auxiliary method based on the concept of *sequence logo* extended to process experimental secondary structure assigned by DSSP, and a method based on performing a *principal component analysis* (PCA) on different types of protein internal coordinates. The same PCA based analysis has been used on other protein families to show that xenon binding and derivatisation does not strongly affect overall protein structure. The conformational analyses of Hsp90-NTD revealed the conformational flexibility of a central helical protein segment close to the ligand binding site, and protein structures clustering with the unliganded form of the protein were used to predict xenon binding. Xenon was predicted to bind with a strongly favourable score to two closely proximal positions within the protein core where two well defined conserved water molecules were found in all but one of the structures, and with a medium-favourable score (corresponding to a binding likelihood of about 30 to 50 % each) to three other protein moieties, namely the ligand binding site, a site close to the ligand binding site, and a small protein surface indentation distal to the ligand binding site.

By solving the structures of Hsp90-NTD in the presence and absence of xenon to high resolution (1.65 and 1.68 Å), the predictions were assessed, detecting no xenon binding at the site of the buried waters in the protein core, but at two out of three of the other sites, as well as at another two sites close to each other in a hydrophobic core region of the protein that had not been sampled by water molecules. The absence of the former binding sites was discussed in light of the finding of close protein amino acid side chain packing that might be prohibitive for xenon binding. The two xenon positions not predicted but found present at an alternative similarly tightly packed position in the protein hydrophobic core on the other hand were illustrating the potentially sub optimal sampling technique employed, and at the same time, the limitation of representing xenon as a rigid sphere, given its strong polarisability, or, potentially, the failure to detect a marginally populated, minor protein conformation permissive of accommodating said xenon atoms. Overall, the performance of the scoring function was found to be very satisfactory, with the false positive and false negative cases enforcing the caveat of the importance of position sampling for the overall success of the method, as well as raising the additional point of the over simplified xenon representation

as a rigid sphere possibly not sufficiently reflecting its polarisability. In order to investigate these points independently of Hsp90-NTD as sole model system, additional model systems were taken into account, and xenon binding predictions have been made for bovine trypsin and lipase B from *Candida antarctica*. Furthermore, due to the strong dependence of the predictive method on a single set of three dimensional coordinates, a molecular mechanics based approach is intended to be developed, rigorously incorporating pico- to nanosecond protein motion as well as explicit solvation (see section 3.3.9 in the previous chapter).

Chapter 5

Conclusions

In recent years, fragment based drug design (FBDD) has emerged as a paradigm in drug discovery, facilitated by advances in both the sensitivity and throughput of biophysical techniques for the characterisation of protein-ligand interactions. As compared to traditional methods, in FBDD, much smaller molecules are screened against biological targets. The small size of these molecules poses unique challenges to experimental and theoretical investigations, but opens up opportunities to address common shortcomings of traditional drug design approaches.

In FBDD, relatively small libraries of smaller compounds are screened against biological targets, as they cover chemical space more efficiently than collections of larger compounds do. The main goal of this approach is to detect lead-like compounds that can then be developed into specific and highly potent drug-like molecules by rational design, using medicinal chemistry. Ideally, this design process is supplemented by structural information on the complex formed by the protein and the ligand, thereby guiding rational subsitutions and additions to the molecular framework of the compound. X-ray crystallography remains the gold standard for the determination of such complexes at atomic resolution. However, not all systems are 'well-behaved' under crystallographic conditions, and crystallographic structure determination is a time-consuming process. To address this shortcoming, the NMR based INPHARMA methodology has been described that determines the relative binding mode of two competitive ligands, allowing to establish pharmacophores of ligand series. In favourable cases also the absolute binding mode can

be established. The approach is especially well suited for the investigation of series of compounds such as those encountered during a structure based drug design campaign. The INPHARMA methodology works in absence of receptor isotope labelling and requires ligands to be in the fast exchange regime of NMR spectroscopy, which is usually fulfilled in the context of FBDD. NMR based methodologies for the investigation of protein-ligand complexes traditionally face substantial challenges to obtain information at atomic resolution. INPHARMA addresses this shortcoming, and bridges the gap between ligand- and target-detected NMR methods. The method is especially efficient in extending lead series, i.e. determining the orientation of the $i + 1$-st ligand within the binding pocket of the protein when the orientations of i other ligands are already known. A previously described application of the method to the Protein kinase A test system has indicated, however, that the unphysical representation of proteins as rigid molecules can lead to sub-optimal discrimination between multiple binding poses of one ligand that were related by a 180 degrees rotation about one of its main axes. Additional experimental information had to be taken into account to resolve this ambiguity. Similar problems are prevalent in FBDD, since fragment sized molecules are small, can possess a certain degree of internal symmetry, and possibly display multiple ambiguous binding modes. INPHARMA in principle has the potential to achieve higher throughput than X-ray crystallography, and at the same time, comparable high-resolution information, if persisting issues can be addressed and the discrimination power of the method can be improved. In this thesis, it was found that the discrimination power of INPHARMA can be improved by rigorously representing protein plasticity by incorporating protein internal motion. This not only makes the underlying physical model of the method more realistic, but it also increases its discrimination powerin conjunction with the generic order parameter concept that was developed and validated throughout this thesis. This concept allows to increase the throughput of the INPHARMA method by getting independent of MD simulations of the protein-ligand systems under investigation for the estimation of NMR order parameters. These time-consuming simulations are usually needed in order to obtain useful information about protein internal motion, but their reliability is hampered by the limited suitability of current molecular mechanics force fields for the simulation of protein-ligand complexes.

Additionally, initial results presented in this thesis suggest a certain amount of information about the receptor to be present in the ligand-detected INPHARMA spectra, making use of the magnetisation imprint onto the protein that was found to be present in the ligand spectrum. Taken together, the substantial improvement of the INPHARMA methodology, itself addressing an important gap in the FBDD process, brings the methodology closer to routine application and high(er) throughput.

In an orthogonal approach presented in the second part of this thesis, the proficiency of the noble gas xenon was investigated to act as a molecular probe for the detection of protein binding sites. Xenon is known to interact with proteins and form relatively stable complexes that can be observed in flash-cooled protein crystals without further preparation. Furthermore, xenon is traditionally used as heavy atom to derivatise protein crystals for initial phasing. Therefore, a plethora of experimental protein-xenon complex structures exist in the Protein Data Bank that were made use of in this thesis to develop a knowledge-based xenon-protein interaction potential. This potential was used to verify the empirical observation that xenon often binds to the ligand binding sites of proteins, and to characterise them. Indeed, it was found that for the systems investigated, in about one third of the cases, a xenon atom occupied locations within the ligand binding pocket. This potentially renders xenon useful to experimentally probe protein druggability. Furthermore, in the context of FBDD, xenon can be regarded to as 'the perfect fragment': it is isomorphous, not toxic, and has similar size and binding characteristics to benzene groups which are arguably among the most important functional groups in drug-like molecules. Furthermore, the technical means to handle it are present and routinely used at many X-ray facilities. Using xenon to derivatise crystals has the added benefit of being able to obtain phase information. Therefore, for novel target proteins where phasing by molecular replacement is not a viable option, xenon can be used to solve the phase problem, and at the same time, information about the potential druggability of the protein can be obtained 'for free'. Xenon binding to proteins has been described to be much more specific than the binding of organic solvent molecules such as benzene. Due to its large polarisable electron cloud, it possesses an inducable dipole moment, and can en-

tertain a limited number of interactions, mainly of hydrophobic nature, but it can also tolerate the presence of polar protein moieties that are often found within binding sites. Taken together, this renders xenon an excellent *universal* fragment, and the addition of xenon to the standard library of fragments is advocated.

In the last part of this thesis, using the Hsp90 protein system as a test case, a computational approach to predict xenon binding sites of proteins was validated. This approach has the intention to predict protein druggability from a set of three-dimensional coordinates of the protein structure, allowing the application of the approach to novel targets, thereby prioritising them for experimental studies. This computational method was found to be highly sensitive, and to outperform molecular mechanics based approaches that rely on empirical force field terms. This might in part be attributed to the fact that the functional form of a knowledge based potential is independent of the limitations and restrictions of the empirical scoring functions used in molecular mechanics. This is especially relevant in the case of xenon with its large polarisable electron cloud, influencing its interaction properties. This effect is potentially poorly represented by a point charge model at a molecular mechanics level of theory. By combining a molecular mechanics based sampling of possible xenon positions, and importantly, protein plasticity, with the knowledge-based approach developed in this thesis, in future work, likely, the sampling of the method will be improved. In agreement with the results reported in the first part of this thesis, the importance of representing protein plasticity and its impact on ligand binding has been demonstrated.

Taken together, the methods described in this thesis represent novel approaches in the field of drug discovery and FBDD, hopefully opening up a host of new targets and model systems, and provide complementary information to established methods.

List of Publications

Prior to this Ph.D.

M Perković, S Schmidt, D Marino, R A Russell, **B Stauch**, H Hofmann, F Kopietz, B P Kloke, J Zielonka, H Ströver, J Hermle, D Lindemann, V K Pathak, G Schneider, M Löchelt, K Cichutek, and C Münk. Species-specifiv inhibition of APOBEC3C by the prototype foamy virus protein bet. *Journal of biological chemistry*, 2009.

B Stauch, H Hofmann, M Perković, M Weisel, F Kopietz, K Cichutek, C Münk, and G Schneider. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proceedings of the National Academy of Sciences*, 2009.

During this Ph.D.

B Stauch, B Simon, T Basile, G Schneider, N P Malek, M Kalesse, and T Carlomagno. Elucidation of the structure and intermolecular interactions of a reversible cyclic-peptide inhibitor of the proteasome by NMR spectroscopy and molecular modeling. *Angewandte Chemie*, 2010.

B Stauch, J Orts, and T Carlomagno. The description of protein internal motions aids selection of ligand binding poses by the INPHARMA method. *Journal of Biomolecular NMR*, 2012.

Appendix

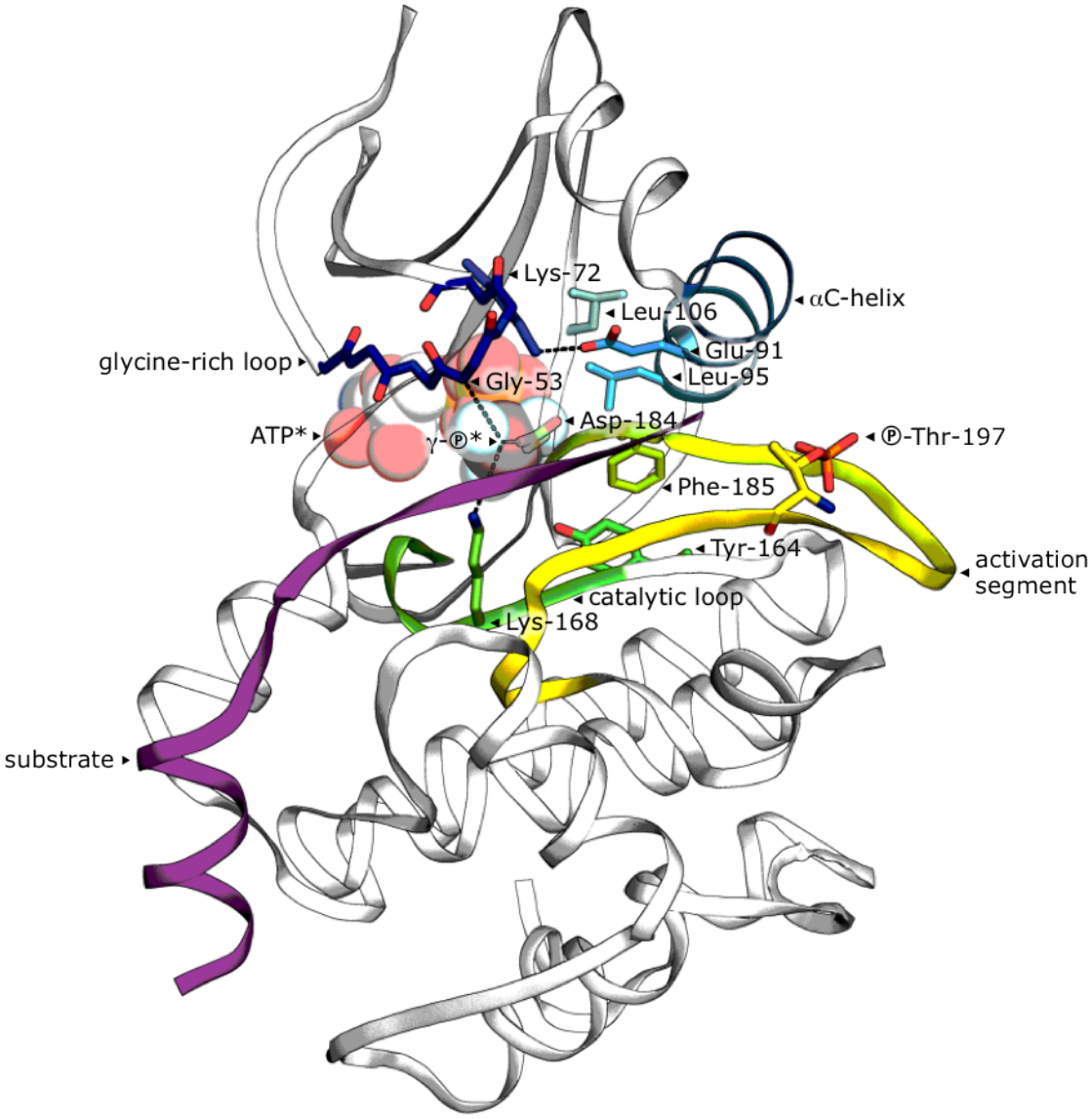


Figure S1 – Detailed view of the core domain (residues 40 to 300) of the catalytic subunit of cAMP-dependent protein kinase, in complex with a stable ATP analogue (ATP*, semi-transparent space filling representation) and substrate peptide (magenta) (PDB identifier 1l3r) (Madhusudan et al., 2002), shown in cartoon representation. Protein residues of particular interest are shown as sticks (blue nitrogen, red oxygen, orange phosphorus). Carbon atoms and relevant consecutive backbone segments (glycine-rich loop, residues 50 to 55; αC-helix, 84 to 96; catalytic loop, 166 to 171; activation segment, 184 to 204) are coloured on a rainbow scale depending on their location within the primary sequence (blue, cyan, green, and yellow, increasing with distance to N-terminus). A salt bridge formed by Lys^{72} and Glu^{91}

is indicated by a dashed black line as well as three protein residues (Gly^{53}, Lys^{168}, Asp^{184}) coordinating the γ-phosphate (Ⓟ) of the nucleotide ligand. The relevance of residues 184 to 186 forming the DFG motif, the regulatory spine (Leu^{95}, Leu^{106}, Tyr^{164}, Phe^{185}), and the protein segments mentioned above are discussed in the main text. The picture was generated using PyMOL (DeL).

Table S1 – List of ATP-competitive kinase inhibitors approved by the FDA, based on a list of inhibitors retrieved from http://kinase.sakura.ne.jp/approved-inhibitors (July 2013). Inhibitor types (type I or type II) were assigned based on their relative affinity for the phosphorylated form of human tyrosine-protein kinase ABL1 (UniProt P00519), corresponding to the active enzyme (type I inhibitor) versus the non-phosphorylated form (inactive enzyme, type II inhibitor) (Davis et al., 2011), where available. ChEMBL (Gaulton et al., 2012; Bento et al., 2013) identifiers are indicated for each compound.

	ChEMBL ID	FDA approval	ATC code	inhibitor type	X-ray structure (PDB ID)[a]	molecular target
Imatinib	941	2001	L01XE01	type II	DFG-out (3gvu)	ABL, PDGFR, KIT
Gefitinib	939	2003	L01XE02	type I	DFG-in (2ity)	EGFR
Erlotinib	553	2004	L01XE03	type I	DFG-in (1m17)	EGFR
Sorafenib	1336	2005	L01XE05	type II	DFG-out (4asd)	VEGFR2, PDGFR, RAF, etc.
Dasatinib	1421	2006	L01XE06	type I	DFG-in (2gqg)	ABL, SRC
Sunitinib	535	2006	L01XE04	type I	DFG-out (4agd)	VEGFR2, PDGFR, KIT
Lapatinib	554	2007	L01XE07	undefined[b]	DFG-in (1xkk)	EGFR, HER2
Nilotinib	255863	2007	L01XE08	type II	DFG-out (3cs9)	ABL
Pazopanib	477772	2009	L01XE11	type I	DFG-in (3cjg)	VEGFR2, PDGFR, KIT
Bosutinib	288441	2010	L01XE14	type I	DFG-out (3ue4)	ABL
Crizotinib	601719	2011	L01XE16	type I	DFG-in (2xp2)	ALK, MET
Ruxolitinib	1789941	2011	L01XE18		JAK1/2	
Vandetanib	24828	2011	L01XE12	type I	DFG-in (2ivu)	VEGFR2, EGFR, RET
Vemurafenib	1229517	2011	L01XE15	type I[c]		BRAF
Axitinib	1289926	2012	L01XE17	type I	DFG-out (4ag8)	VEGFR1-3, PDGFR, KIT
Cabozantinib	2105717	2012				VEGFR2, RET, MET
Ponatinib	1171837	2012		type II[d]	DFG-out (3oxz)	ABL, SRC
Regorafenib	1946170	2012	L01XE21			VEGFR2, TIE2, etc.
Tofacitinib	221959	2012	L04AA29	type I		JAK3
Trametinib	2103875	2013				MEK
Dabrafenib	2028663	2013				BRAF
Afatinib	1173655	2013	L01XE13	type I	DFG-in (4g5j)	EGFR, HER2

[a] cyrstal structure references: 2ity (Yun et al., 2007); 1m17 (Stamos et al., 2002); 4asd, 4ag8, 4agd (McTigue et al., 2012); 2gqg (Tokarski et al., 2006); 1xkk (Wood et al., 2004); 3cs9 (Weisberg et al., 2005); 3cjg (Harris et al., 2008); 3ue4 (Levinson and Boxer, 2012); 2xp2 (Cui et al., 2011); 2ivu (Knowles et al., 2006); 3oxz (Zhou et al., 2011); 4g5j (Solca et al., 2012).

[b] for discussion see main text and (Zuccotto et al., 2010; Liu and Gray, 2006; Zhang et al., 2009).

[c] (Atefi et al., 2011).

[d] (Smith et al., 2013).

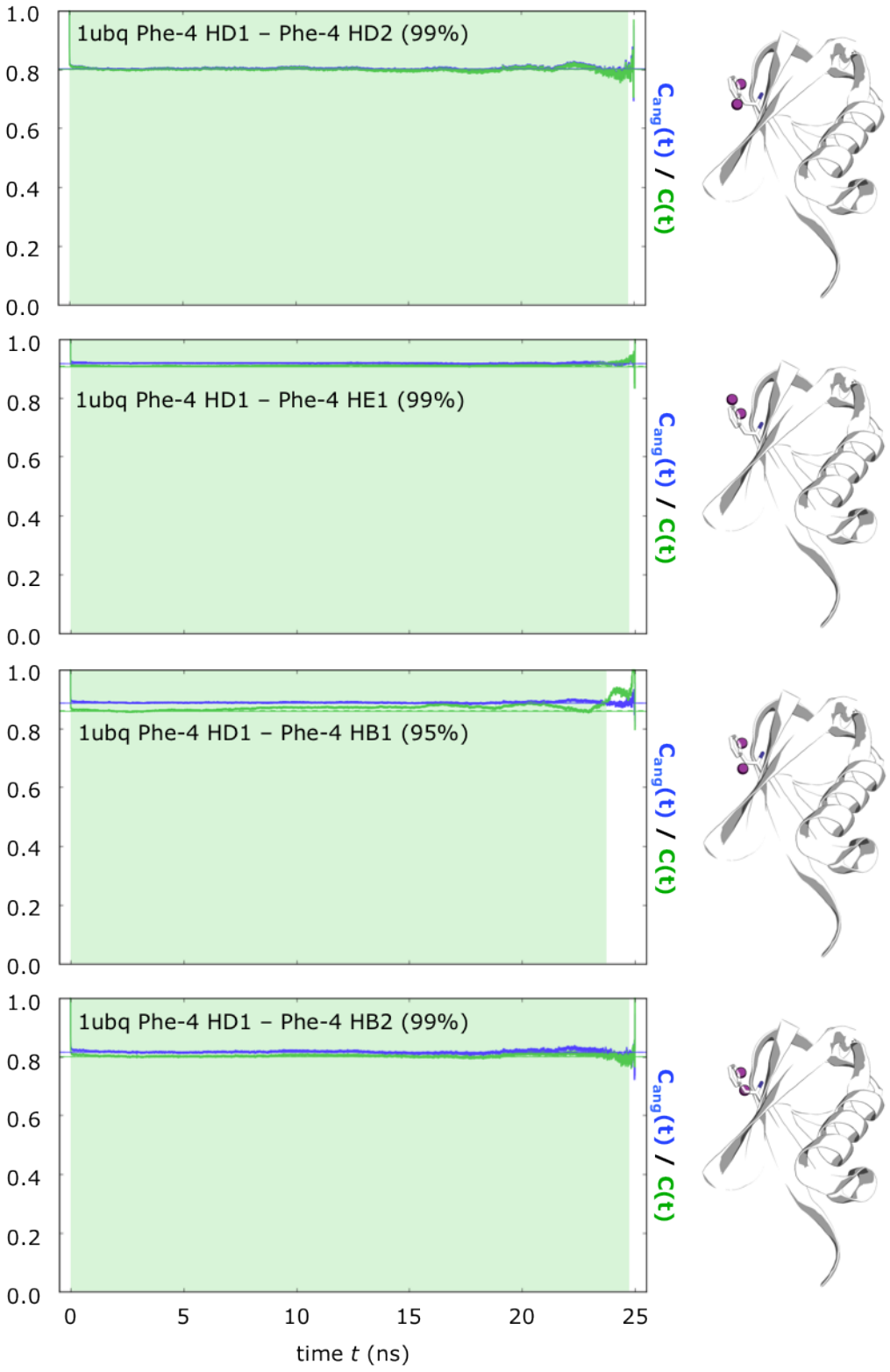

1ubq Phe-4 HD1 – Phe-4 HD2 (99%)
1ubq Phe-4 HD1 – Phe-4 HE1 (99%)
1ubq Phe-4 HD1 – Phe-4 HB1 (95%)
1ubq Phe-4 HD1 – Phe-4 HB2 (99%)
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
time t (ns)
C_ang(t) / C(t)

Figure S2 – Internal correlation functions of representative proton pairs, part 3. See Figure 2.13 in the main text for reference, and Supplementary Figure S4 for an analysis of ring flips.

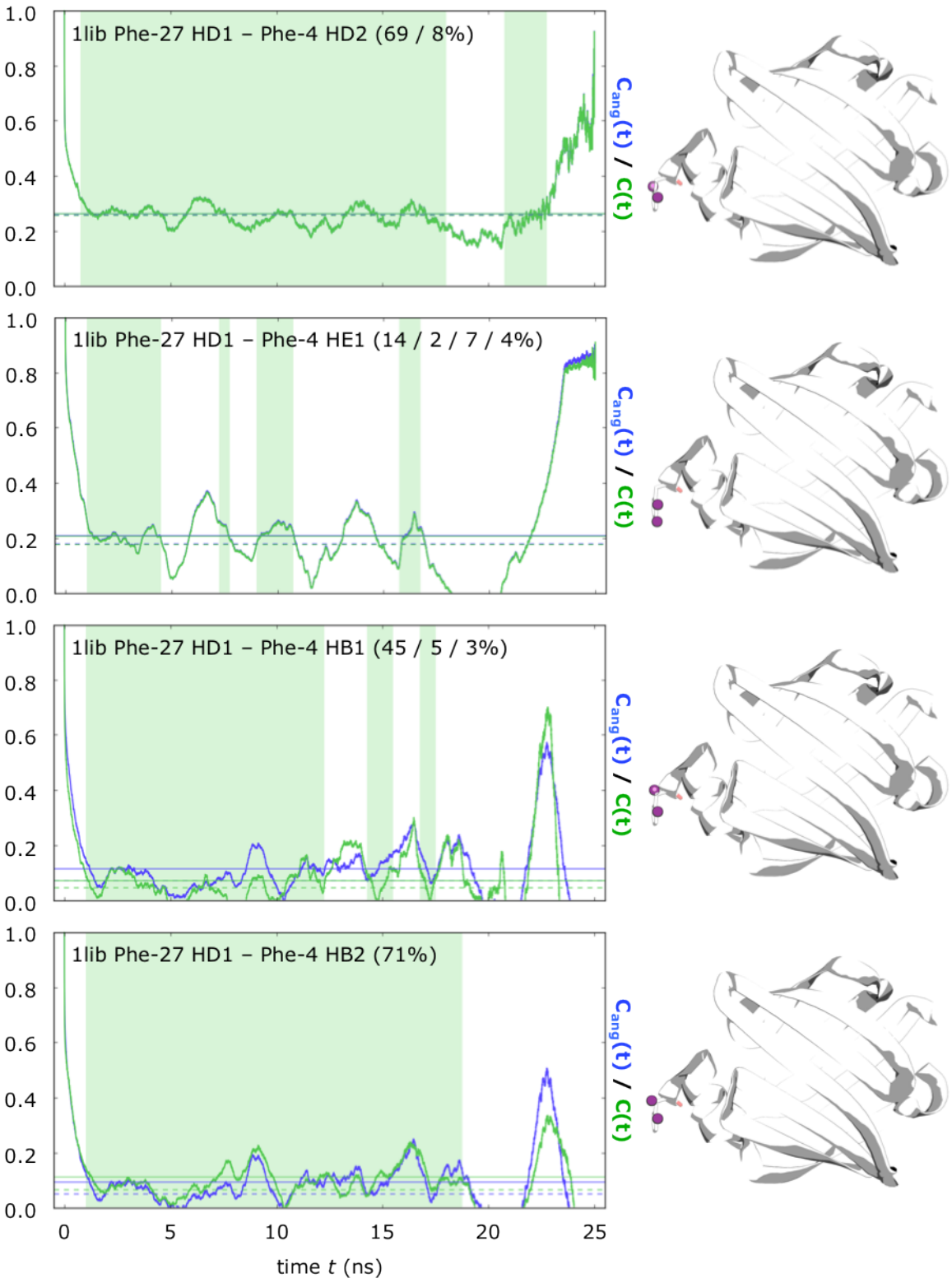

1lib Phe-27 HD1 – Phe-4 HD2 (69 / 8%)
1lib Phe-27 HD1 – Phe-4 HE1 (14 / 2 / 7 / 4%)
1lib Phe-27 HD1 – Phe-4 HB1 (45 / 5 / 3%)
1lib Phe-27 HD1 – Phe-4 HB2 (71%)
C_ang(t) / C(t)
time t (ns)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25

Figure S3 – Internal correlation functions of representative proton pairs, part 4. See Figure 2.13 in the main text for reference, and Supplementary Figure S4 for an analysis of ring flips.

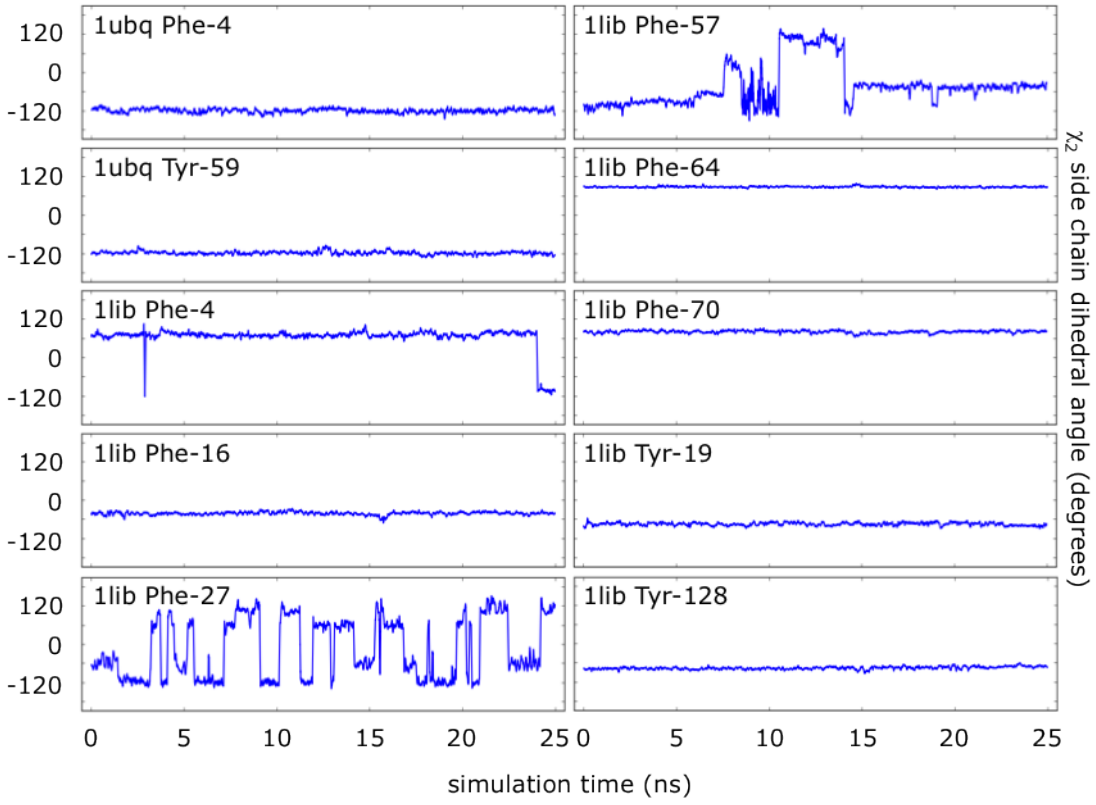


Figure S4 – Aromatic side chain dihedral angle values extracted from molecular dynamics (MD) simulations. Values of the χ_2 dihedral angle of aromatic side chains (phenylalanine and tyrosine residues), involving the atoms CA (C^{α}), CB (C^{β}), CG (C^{γ}), and CD1 (C^{δ_1}), were extracted from 25 ns MD simulations of human ubiquitin (1ubq) and murine adipocyte lipid binding protein (1lib), and are shown over time. Conformational snapshots were saved every 1 ps and are presented as rolling window averages of 50 ps, for clarity. The trajectory for Phe^{45} of human ubiquitin was found to be practically identical to that of Tyr^{19} of the murine adipocyte lipid binding protein and was therefore omitted. Internal correlation functions for Phe^{4} of 1ubq and Phe^{27} of 1lib are shown in further detail in Supplementary Figures S2 and S3.

Table S2 – Protein structures used to derive the xenon likeness score. Protein Data Bank (PDB) (Berman et al., 2000), UniProt (UniProt Consortium, 2012), and Pfam (Punta et al., 2012) identifiers are indicated. Xenon atoms were assigned to the protein chains they were closest to. Only one UniProt or Pfam identifier per PDB identifier is given.

PDB	UniProt	Pfam	PDB	UniProt	Pfam	PDB	UniProt	Pfam
1c10	P00698	PF00062	1uvx	Q08753	PF01152	3cyu	P00918	PF00194
1c1m	P00772	PF00089	1uvy	P15160	PF01152	3fbx	Q3TCN2	PF04916
1c3l	P00780	PF00082	1ux9	Q8WWM9	PF00042	3fgr	Q3TCN2	PF04916
1c62	P00720	PF00959	1uyu	P00183	PF00067	3fz6	P69910	PF00282
1c65	P00720	PF00959	1vau	P00698	PF00062	3g46	P02213	PF00042
1c68	P00720	PF00959	1vgi	P06762	PF01126	3gk9	Q9ER97	PF00042
1c6b	P00720	PF00959	1w2z	Q43077	PF01179	3gln	Q9ER97	PF00042
1c6e	P00720	PF00959	1w53	P40399	PF08673	3ls7	P00800	PF01447
1c6h	P00720	PF00959	1zdm	P0AE67	PF00072	3lzy	P11838	PF00026
1c6k	P00720	PF00959	2a7a	P81461	PF00139	3m3d	P04058	PF00135
1c6n	P00720	PF00959	2a7b	P22892	PF02283	3mou	Q9NAV8	PF00042
1c6t	P00720	PF00959	2a7c	P00772	PF00089	3ord	Q9NAV8	PF00042
1e9v	P31101	PF03063	2a7d	P00698	PF00062	3pjk	Q00511	PF01014
1fo6	Q44184	PF00795	2a9r	Q9S1K0	PF00072	3pk4	Q00511	PF01014
1fzh	P22869	PF02332	2b2j	O29285	PF00909	3pk5	Q00511	PF01014
1fzi	P22869	PF02332	2dki	Q05KQ5	PF01494	3pk6	Q00511	PF01014
1gkz	Q00972	PF02518	2fic	O00499	PF03114	3pkf	Q00511	PF01014
1i4w	P14908	PF00398	2ic0	Q00511	PF01014	3pkg	Q00511	PF01014
1j52	P02185	PF00042	2ie6	P14668	PF00191	3pkh	Q00511	PF01014
1k4k	P0A752	PF01467	2oqe	P12807	PF01179	3pkk	Q00511	PF01014
1kqn	Q9HAN9	PF01467	2oqu	P00772	PF00089	3pkl	Q00511	PF01014
1l0z	P00772	PF00089	2w0q	P46883	PF01179	3ple	Q00511	PF01014
1l1g	P00772	PF00089	2w6v	P69905	PF00042	3plg	Q00511	PF01014
1lls	P0AEX9	PF01547	2w6w	P02185	PF00042	3plh	Q00511	PF01014
1nxt	Q9S1K0	PF00072	2w6x	P02185	PF00042	3pli	Q00511	PF01014
1o75	P29723	PF14888	2w6y	P02185	PF00042	3plj	Q00511	PF01014
1rjo	P46881	PF01179	2w72	P69905	PF00042	3plm	Q00511	PF01014
1rky	Q96X16	PF01179	2xkh	Q76242	PF00042	3qpk	Q70KY3	PF00394
1s56	P0A592	PF01152	2z8a	P02213	PF00042	3tf1	Q8RBX6	PF07700
1u0x	Q94734	PF02087	2z8y	P27989	PF03063	3tf9	Q8YUQ7	PF07700
1uo6	P00772	PF00089	2zfe	P02945	PF01036	3tfa	Q8YUQ7	PF07700
1uoc	P39008	PF04857	3bvd	Q5SJ79	PF00115	3tfe	Q8YUQ7	PF07700
1ury	Q8WWM9	PF00042	3c9i	P35837	n/a	3u52	Q84AQ2	PF02332

Table S3 – Values of the xenon radial distribution functions. A key to the CHARMM22 protein atom types can be found in Table 3.6.

bin centre (Å)	C	CA	CC	CP1	CP2	CP3	CPH1	CPH2	CPT	CT1
0.125	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.375	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.625	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.875	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.125	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.375	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.625	0.000	0.016	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
1.875	0.000	0.033	0.001	0.000	0.002	0.000	0.001	0.001	0.004	0.004
2.125	0.002	0.073	0.006	0.000	0.006	0.001	0.004	0.007	0.019	0.009
2.375	0.008	0.163	0.018	0.002	0.020	0.006	0.015	0.030	0.066	0.022
2.625	0.027	0.346	0.045	0.010	0.054	0.018	0.043	0.092	0.178	0.049
2.875	0.072	0.669	0.097	0.036	0.122	0.044	0.095	0.228	0.389	0.107
3.125	0.156	1.150	0.181	0.098	0.237	0.089	0.177	0.453	0.697	0.220
3.375	0.288	1.750	0.298	0.215	0.399	0.157	0.275	0.746	1.045	0.411
3.625	0.460	2.360	0.439	0.392	0.584	0.243	0.374	1.017	1.330	0.686
3.875	0.646	2.848	0.585	0.604	0.750	0.343	0.460	1.172	1.472	1.029
4.125	0.813	3.122	0.709	0.810	0.857	0.450	0.544	1.156	1.471	1.398
4.375	0.940	3.179	0.782	0.953	0.893	0.558	0.647	1.015	1.415	1.739
4.625	1.022	3.084	0.786	1.012	0.868	0.652	0.777	0.843	1.415	1.996
4.875	1.070	2.917	0.729	0.980	0.805	0.719	0.910	0.732	1.531	2.130
5.125	1.098	2.728	0.638	0.889	0.719	0.747	1.003	0.719	1.737	2.124
5.375	1.115	2.536	0.546	0.773	0.618	0.743	1.021	0.798	1.956	1.997
5.625	1.133	2.344	0.475	0.661	0.516	0.716	0.968	0.927	2.093	1.797
5.875	1.162	2.163	0.428	0.560	0.431	0.680	0.884	1.048	2.110	1.583
6.125	1.211	2.003	0.401	0.463	0.374	0.645	0.818	1.107	2.010	1.405
6.375	1.283	1.872	0.392	0.370	0.349	0.621	0.811	1.085	1.852	1.287
6.625	1.365	1.771	0.400	0.294	0.349	0.609	0.874	1.013	1.688	1.238
6.875	1.441	1.698	0.430	0.252	0.374	0.612	0.992	0.954	1.568	1.250
7.125	1.497	1.650	0.483	0.254	0.419	0.626	1.122	0.944	1.507	1.308
7.375	1.528	1.614	0.554	0.294	0.477	0.655	1.216	0.981	1.508	1.393
7.625	1.539	1.580	0.630	0.359	0.538	0.702	1.240	1.025	1.552	1.482
7.875	1.537	1.538	0.702	0.438	0.597	0.759	1.198	1.050	1.623	1.561
8.125	1.530	1.486	0.765	0.530	0.650	0.810	1.116	1.058	1.702	1.617
8.375	1.520	1.429	0.825	0.646	0.697	0.837	1.035	1.066	1.775	1.650
8.625	1.504	1.378	0.885	0.793	0.741	0.849	0.987	1.080	1.829	1.661
8.875	1.485	1.339	0.941	0.970	0.792	0.874	0.986	1.082	1.843	1.657
9.125	1.464	1.316	0.987	1.157	0.857	0.937	1.024	1.051	1.807	1.639
9.375	1.441	1.307	1.013	1.325	0.936	1.033	1.074	0.988	1.734	1.605
9.625	1.418	1.310	1.019	1.435	1.020	1.130	1.109	0.916	1.660	1.556
9.875	1.396	1.323	1.014	1.472	1.105	1.192	1.112	0.871	1.626	1.494

Values of the xenon radial distribution functions - continued.

bin centre (Å)	C	CA	CC	CP1	CP2	CP3	CPH1	CPH2	CPT	CT1
10.125	1.373	1.342	1.012	1.438	1.186	1.210	1.090	0.878	1.651	1.427
10.375	1.348	1.364	1.024	1.367	1.266	1.206	1.063	0.931	1.714	1.362
10.625	1.318	1.382	1.052	1.295	1.334	1.218	1.052	1.008	1.779	1.303
10.875	1.284	1.394	1.093	1.259	1.379	1.266	1.058	1.076	1.800	1.252
11.125	1.245	1.394	1.136	1.258	1.390	1.328	1.074	1.117	1.765	1.211
11.375	1.205	1.379	1.170	1.272	1.368	1.364	1.093	1.126	1.672	1.178
11.625	1.168	1.348	1.190	1.264	1.316	1.337	1.127	1.117	1.548	1.152
11.875	1.135	1.307	1.194	1.214	1.246	1.254	1.184	1.106	1.417	1.129
12.125	1.107	1.265	1.185	1.126	1.171	1.147	1.264	1.108	1.302	1.107
12.375	1.083	1.229	1.169	1.026	1.104	1.056	1.342	1.124	1.208	1.087
12.625	1.061	1.198	1.152	0.947	1.053	0.998	1.400	1.154	1.133	1.071
12.875	1.040	1.168	1.139	0.912	1.017	0.977	1.429	1.193	1.073	1.062
13.125	1.022	1.132	1.134	0.926	0.989	0.978	1.439	1.245	1.038	1.057
13.375	1.008	1.095	1.135	0.974	0.969	0.987	1.436	1.306	1.034	1.056
13.625	0.999	1.061	1.143	1.031	0.959	0.987	1.424	1.369	1.060	1.054
13.875	0.995	1.036	1.154	1.073	0.961	0.970	1.397	1.410	1.095	1.047
14.125	0.994	1.019	1.165	1.088	0.974	0.939	1.353	1.416	1.118	1.031
14.375	0.992	1.005	1.169	1.080	0.994	0.909	1.292	1.386	1.116	1.008
14.625	0.989	0.988	1.162	1.060	1.019	0.895	1.225	1.336	1.094	0.981
14.875	0.984	0.965	1.141	1.035	1.048	0.908	1.155	1.278	1.061	0.952
15.125	0.976	0.938	1.112	1.007	1.075	0.945	1.090	1.210	1.022	0.927
15.375	0.967	0.912	1.082	0.979	1.092	0.994	1.035	1.117	0.974	0.908
15.625	0.955	0.891	1.059	0.955	1.092	1.037	0.994	0.995	0.913	0.896
15.875	0.942	0.874	1.042	0.946	1.076	1.066	0.965	0.865	0.840	0.889
16.125	0.930	0.861	1.029	0.957	1.054	1.077	0.944	0.760	0.764	0.883
16.375	0.921	0.849	1.018	0.994	1.037	1.077	0.923	0.708	0.698	0.877
16.625	0.916	0.838	1.011	1.056	1.031	1.072	0.898	0.719	0.651	0.872
16.875	0.914	0.826	1.012	1.130	1.034	1.067	0.869	0.778	0.627	0.871
17.125	0.914	0.813	1.020	1.189	1.041	1.054	0.843	0.861	0.625	0.872
17.375	0.915	0.800	1.031	1.211	1.046	1.033	0.820	0.947	0.641	0.875
17.625	0.915	0.785	1.038	1.189	1.046	1.009	0.799	1.019	0.667	0.878
17.875	0.914	0.771	1.037	1.139	1.041	1.000	0.781	1.071	0.695	0.879
18.125	0.910	0.758	1.026	1.083	1.032	1.018	0.772	1.093	0.718	0.876
18.375	0.903	0.746	1.011	1.035	1.026	1.056	0.781	1.079	0.733	0.868
18.625	0.891	0.733	0.995	0.994	1.025	1.090	0.814	1.027	0.742	0.852
18.875	0.870	0.716	0.974	0.951	1.022	1.093	0.858	0.945	0.741	0.825
19.125	0.826	0.680	0.930	0.889	0.991	1.043	0.882	0.842	0.719	0.775
19.375	0.739	0.610	0.840	0.790	0.903	0.925	0.848	0.720	0.652	0.690
19.625	0.577	0.477	0.659	0.615	0.713	0.712	0.697	0.549	0.513	0.537
19.875	0.371	0.307	0.426	0.397	0.461	0.452	0.465	0.353	0.329	0.346

Values of the xenon radial distribution functions - continued.

bin centre (Å)	CT2	CT3	CY	N	NC2	NH1	NH2	NH3	NR1	NR2
0.125	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.375	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.625	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.875	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.125	0.000	0.051	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.375	0.000	0.074	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
1.625	0.002	0.096	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000
1.875	0.007	0.122	0.001	0.000	0.000	0.001	0.018	0.000	0.002	0.002
2.125	0.018	0.170	0.004	0.000	0.000	0.004	0.050	0.002	0.007	0.009
2.375	0.046	0.278	0.015	0.003	0.001	0.012	0.110	0.007	0.020	0.031
2.625	0.104	0.500	0.049	0.012	0.005	0.033	0.203	0.018	0.046	0.084
2.875	0.212	0.893	0.128	0.039	0.015	0.080	0.322	0.039	0.095	0.187
3.125	0.387	1.494	0.279	0.097	0.039	0.165	0.449	0.072	0.179	0.354
3.375	0.634	2.265	0.523	0.195	0.085	0.295	0.559	0.114	0.316	0.563
3.625	0.924	3.071	0.854	0.321	0.155	0.459	0.628	0.163	0.517	0.768
3.875	1.203	3.717	1.227	0.438	0.244	0.633	0.636	0.215	0.775	0.895
4.125	1.409	4.035	1.570	0.514	0.332	0.781	0.586	0.271	1.053	0.899
4.375	1.508	3.984	1.796	0.536	0.399	0.886	0.499	0.325	1.294	0.782
4.625	1.507	3.648	1.842	0.541	0.430	0.951	0.412	0.367	1.448	0.614
4.875	1.442	3.184	1.702	0.569	0.431	1.000	0.353	0.387	1.494	0.469
5.125	1.349	2.731	1.441	0.639	0.418	1.061	0.329	0.384	1.448	0.404
5.375	1.251	2.363	1.175	0.714	0.412	1.144	0.331	0.372	1.345	0.424
5.625	1.157	2.096	1.017	0.748	0.428	1.243	0.347	0.368	1.238	0.507
5.875	1.067	1.913	1.035	0.712	0.471	1.339	0.365	0.380	1.156	0.624
6.125	0.983	1.789	1.216	0.632	0.532	1.417	0.376	0.396	1.102	0.756
6.375	0.911	1.702	1.475	0.559	0.592	1.475	0.380	0.400	1.048	0.888
6.625	0.858	1.636	1.702	0.533	0.631	1.517	0.387	0.385	0.979	1.007
6.875	0.836	1.585	1.819	0.558	0.644	1.547	0.413	0.365	0.889	1.101
7.125	0.850	1.552	1.830	0.610	0.643	1.564	0.468	0.357	0.803	1.160
7.375	0.899	1.541	1.790	0.668	0.644	1.568	0.550	0.370	0.744	1.193
7.625	0.971	1.556	1.771	0.724	0.653	1.559	0.644	0.393	0.733	1.214
7.875	1.051	1.589	1.809	0.781	0.668	1.546	0.725	0.409	0.764	1.236
8.125	1.126	1.630	1.906	0.842	0.684	1.534	0.770	0.410	0.819	1.260
8.375	1.187	1.664	2.042	0.903	0.698	1.525	0.780	0.399	0.879	1.263
8.625	1.234	1.679	2.191	0.963	0.719	1.521	0.766	0.389	0.939	1.219
8.875	1.268	1.671	2.326	1.013	0.759	1.523	0.754	0.395	1.011	1.115
9.125	1.289	1.635	2.416	1.054	0.823	1.529	0.758	0.424	1.110	0.974
9.375	1.297	1.578	2.434	1.085	0.904	1.536	0.784	0.473	1.231	0.838
9.625	1.289	1.506	2.370	1.118	0.981	1.535	0.824	0.538	1.342	0.749
9.875	1.268	1.434	2.237	1.162	1.031	1.519	0.872	0.609	1.400	0.724

Values of the xenon radial distribution functions - continued.

bin centre (Å)	CT2	CT3	CY	N	NC2	NH1	NH2	NH3	NR1	NR2
10.125	1.239	1.372	2.060	1.216	1.046	1.485	0.924	0.673	1.376	0.752
10.375	1.210	1.328	1.864	1.271	1.032	1.435	0.978	0.713	1.279	0.815
10.625	1.184	1.300	1.676	1.310	1.008	1.375	1.034	0.713	1.152	0.895
10.875	1.162	1.283	1.521	1.326	0.986	1.310	1.087	0.679	1.046	0.985
11.125	1.140	1.268	1.410	1.315	0.970	1.248	1.133	0.640	0.983	1.079
11.375	1.120	1.251	1.338	1.282	0.960	1.191	1.170	0.631	0.955	1.171
11.625	1.106	1.231	1.285	1.232	0.957	1.140	1.194	0.670	0.942	1.244
11.875	1.103	1.210	1.236	1.175	0.964	1.097	1.206	0.750	0.946	1.290
12.125	1.112	1.189	1.184	1.120	0.980	1.066	1.210	0.839	0.977	1.304
12.375	1.126	1.168	1.124	1.076	1.001	1.046	1.211	0.910	1.054	1.310
12.625	1.138	1.145	1.054	1.043	1.021	1.037	1.215	0.953	1.163	1.335
12.875	1.142	1.119	0.983	1.013	1.040	1.034	1.221	0.986	1.282	1.405
13.125	1.138	1.088	0.934	0.976	1.061	1.033	1.220	1.034	1.372	1.501
13.375	1.130	1.054	0.937	0.929	1.088	1.032	1.207	1.106	1.419	1.580
13.625	1.119	1.021	0.998	0.882	1.120	1.030	1.178	1.190	1.418	1.592
13.875	1.105	0.993	1.094	0.852	1.149	1.026	1.145	1.263	1.388	1.534
14.125	1.087	0.971	1.180	0.851	1.161	1.017	1.121	1.311	1.345	1.428
14.375	1.067	0.953	1.219	0.881	1.151	1.004	1.113	1.337	1.305	1.315
14.625	1.046	0.937	1.201	0.934	1.125	0.986	1.118	1.348	1.268	1.212
14.875	1.024	0.921	1.138	0.996	1.096	0.966	1.121	1.345	1.228	1.124
15.125	1.003	0.903	1.054	1.047	1.076	0.947	1.115	1.328	1.179	1.038
15.375	0.985	0.883	0.968	1.072	1.066	0.931	1.100	1.295	1.121	0.959
15.625	0.972	0.864	0.897	1.071	1.064	0.919	1.084	1.260	1.062	0.890
15.875	0.967	0.847	0.848	1.055	1.060	0.911	1.074	1.237	1.005	0.844
16.125	0.966	0.832	0.821	1.043	1.054	0.906	1.073	1.231	0.956	0.823
16.375	0.967	0.818	0.806	1.042	1.051	0.903	1.076	1.236	0.915	0.821
16.625	0.964	0.805	0.792	1.049	1.058	0.901	1.078	1.235	0.885	0.821
16.875	0.958	0.793	0.778	1.057	1.076	0.898	1.075	1.218	0.868	0.814
17.125	0.949	0.784	0.759	1.066	1.101	0.894	1.067	1.179	0.863	0.805
17.375	0.939	0.781	0.737	1.083	1.122	0.891	1.054	1.129	0.864	0.810
17.625	0.930	0.781	0.712	1.109	1.132	0.889	1.037	1.084	0.866	0.844
17.875	0.924	0.782	0.695	1.133	1.130	0.890	1.016	1.061	0.864	0.904
18.125	0.921	0.781	0.688	1.137	1.118	0.891	0.998	1.066	0.862	0.970
18.375	0.920	0.775	0.685	1.108	1.099	0.889	0.986	1.091	0.866	1.016
18.625	0.918	0.764	0.674	1.051	1.078	0.881	0.984	1.120	0.875	1.025
18.875	0.906	0.744	0.643	0.984	1.051	0.862	0.981	1.131	0.878	0.993
19.125	0.867	0.705	0.586	0.910	1.002	0.818	0.954	1.097	0.854	0.922
19.375	0.780	0.631	0.500	0.813	0.905	0.732	0.873	0.991	0.780	0.809
19.625	0.611	0.492	0.375	0.643	0.713	0.572	0.690	0.774	0.620	0.623
19.875	0.393	0.317	0.234	0.421	0.462	0.368	0.447	0.496	0.406	0.398

Values of the xenon radial distribution functions - continued.

bin centre (Å)	NY	O	OC	OH1	S
0.125	0.000	0.000	0.000	0.000	0.000
0.375	0.000	0.000	0.000	0.000	0.000
0.625	0.000	0.000	0.000	0.000	0.000
0.875	0.000	0.000	0.000	0.000	0.000
1.125	0.000	0.000	0.000	0.000	0.000
1.375	0.000	0.000	0.000	0.000	0.000
1.625	0.002	0.002	0.001	0.002	0.000
1.875	0.009	0.006	0.004	0.008	0.002
2.125	0.035	0.017	0.013	0.028	0.013
2.375	0.104	0.046	0.034	0.074	0.051
2.625	0.244	0.104	0.074	0.164	0.160
2.875	0.473	0.202	0.138	0.310	0.403
3.125	0.775	0.342	0.221	0.503	0.849
3.375	1.098	0.506	0.308	0.718	1.511
3.625	1.371	0.663	0.384	0.918	2.325
3.875	1.544	0.787	0.449	1.076	3.121
4.125	1.598	0.868	0.515	1.183	3.727
4.375	1.545	0.915	0.594	1.240	4.012
4.625	1.407	0.945	0.672	1.250	3.966
4.875	1.227	0.971	0.717	1.218	3.638
5.125	1.048	0.994	0.705	1.155	3.135
5.375	0.910	1.015	0.646	1.087	2.558
5.625	0.822	1.036	0.569	1.034	2.022
5.875	0.774	1.065	0.504	1.003	1.595
6.125	0.757	1.109	0.461	0.986	1.307
6.375	0.783	1.175	0.443	0.974	1.139
6.625	0.881	1.263	0.448	0.959	1.066
6.875	1.062	1.371	0.474	0.941	1.064
7.125	1.304	1.481	0.512	0.924	1.110
7.375	1.538	1.574	0.554	0.914	1.186
7.625	1.703	1.631	0.594	0.919	1.267
7.875	1.764	1.645	0.635	0.947	1.346
8.125	1.736	1.618	0.682	1.003	1.412
8.375	1.658	1.559	0.736	1.081	1.463
8.625	1.573	1.482	0.792	1.166	1.488
8.875	1.510	1.403	0.839	1.242	1.499
9.125	1.489	1.338	0.873	1.292	1.511
9.375	1.512	1.297	0.893	1.305	1.539
9.625	1.577	1.281	0.910	1.285	1.581
9.875	1.677	1.285	0.930	1.243	1.621

Values of the xenon radial distribution functions - continued.

bin centre (Å)	NY	O	OC	OH1	S
10.125	1.796	1.296	0.952	1.204	1.645
10.375	1.908	1.305	0.973	1.187	1.641
10.625	1.970	1.304	0.993	1.199	1.601
10.875	1.943	1.294	1.014	1.225	1.529
11.125	1.818	1.276	1.037	1.246	1.438
11.375	1.631	1.253	1.063	1.244	1.352
11.625	1.446	1.224	1.086	1.216	1.287
11.875	1.317	1.190	1.101	1.172	1.249
12.125	1.257	1.154	1.106	1.125	1.227
12.375	1.237	1.121	1.104	1.086	1.214
12.625	1.221	1.093	1.101	1.062	1.202
12.875	1.192	1.069	1.104	1.056	1.189
13.125	1.158	1.048	1.113	1.066	1.167
13.375	1.133	1.029	1.128	1.087	1.125
13.625	1.123	1.013	1.144	1.109	1.056
13.875	1.123	1.000	1.161	1.128	0.966
14.125	1.129	0.990	1.174	1.141	0.874
14.375	1.145	0.983	1.183	1.147	0.803
14.625	1.168	0.977	1.186	1.142	0.770
14.875	1.186	0.971	1.183	1.125	0.778
15.125	1.176	0.969	1.177	1.095	0.816
15.375	1.124	0.970	1.167	1.058	0.866
15.625	1.026	0.973	1.150	1.026	0.906
15.875	0.905	0.974	1.125	1.007	0.918
16.125	0.786	0.971	1.095	1.002	0.899
16.375	0.695	0.965	1.064	1.003	0.862
16.625	0.648	0.956	1.036	1.000	0.827
16.875	0.646	0.946	1.015	0.990	0.806
17.125	0.681	0.933	1.005	0.973	0.802
17.375	0.732	0.919	1.004	0.956	0.809
17.625	0.777	0.905	1.009	0.942	0.823
17.875	0.795	0.894	1.017	0.929	0.840
18.125	0.789	0.885	1.022	0.911	0.853
18.375	0.768	0.880	1.026	0.885	0.856
18.625	0.748	0.872	1.027	0.855	0.842
18.875	0.725	0.855	1.017	0.826	0.806
19.125	0.684	0.813	0.975	0.787	0.740
19.375	0.602	0.729	0.879	0.718	0.637
19.625	0.458	0.570	0.689	0.572	0.480
19.875	0.286	0.367	0.443	0.375	0.300

References

Numerical Python. 211

The PyMOL Molecular Graphics System. 23, 30, 156, 227

XPAND computer program. 156

Heatplus: Heatmaps with row and/or column covariates and colored clusters. 155

R: A Language and Environment for Statistical Computing. 155

C Abad-Zapatero and J T Metz. Ligand efficiency indices as guideposts for drug discovery. *Drug discovery today*, 2005. 5

J A Adams. Kinetic and catalytic mechanisms of protein kinases. *Chemical reviews*, 2001. 24

P Akamine, Madhusudan, J Wu, N H Xuong, L F Ten Eyck, and S S Taylor. Dynamic features of cAMP-dependent protein kinase revealed by apoenzyme crystal structure. *Journal of molecular biology*, 2003. 18, 22, 25

I Akritopoulou-Zanze and P J Hajduk. Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. *Drug discovery today*, 2009. 27

M M U Ali, S M Roe, C K Vaughan, P Meyer, B Panaretou, P W Piper, C Prodromou, and L H Pearl. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, 2006. 30, 31, 199, 200

E Anderson, Z Bai, C Bischof, S Blackford, J Demmel, J Dongarra, J Du Croz, A Greenbaum, S Hammarling, A McKenney, and D Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, 1999. 211

M R Arkin and J A Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature reviews Drug discovery*, 2004. 145

M Atefi, E von Euw, N Attar, C Ng, C Chu, D Guo, R Nazarian, B Chmielowski, J A Glaspy, B Comin-Anduix, P S Mischel, R S Lo, and A Ribas. Reversing melanoma cross-resistance to BRAF and MEK inhibitors by co-targeting the AKT/mTOR pathway. *PLOS ONE*, 2011. 229

A Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 2000. 17

M Baker. Fragment-based lead discovery grows up. *Nature reviews Drug discovery*, 2013. 3

X Barril, P Brough, M Drysdale, R E Hubbard, A Massey, A Surgenor, and L Wright. Structure-based discovery of a new class of Hsp90 inhibitors. *Bioorganic & medicinal chemistry letters*, 2005. 187, 213

J J Barrott and T A J Haystead. Hsp90, an unlikely ally in the war on cancer. *The FEBS journal*, 2013. 31

S Bartoschek, T Klabunde, E Defossa, V Dietrich, S Stengelin, C Griesinger, T Carlomagno, I Focken, and K U Wendt. Drug design for G-protein-coupled receptors by a ligand-based NMR method. *Angewandte Chemie*, 2010. 98

M Beck, A Schmidt, J Malmstroem, M Claassen, A Ori, A Szymborska, F Herzog, O Rinner, J Ellenberg, and R Aebersold. The quantitative proteome of a human cell line. *Molecular systems biology*, 2011. 30

G W Bemis and M A Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, 1996. 4

G W Bemis and M A Murcko. Properties of known drugs. 2. Side chains. *Journal of medicinal chemistry*, 1999. 4

A P Bento, A Gaulton, A Hersey, L J Bellis, J Chambers, M Davies, F A Krüger, Y Light, L Mak, S McGlinchey, M Nowotka, G Papadatos, R Santos, and J P Overington. The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, 2013. 17, 229

A Bergerat, B de Massy, D Gadelle, P C Varoutas, A Nicolas, and P Forterre. An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature*, 1997. 31, 165

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 2000. 53, 109, 154, 164, 215, 235

R B Best, T J Rutherford, S M V Freund, and J Clarke. Hydrophobic core fluidity of homologous protein domains: relation of side-chain dynamics to core composition and packing. *Biochemistry*, 2004. 55

M J J Blommers, W Stark, C E Jones, D Head, C E Owen, and W Jahnke. Transferred Cross-Correlated Relaxation Complements Transferred NOE: Structure of an IL-4R-Derived Peptide Bound to STAT-6. *Journal of the American Chemical Society*, 1999. 34

L C Blum and J L Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 2009. 3

P Blume-Jensen and T Hunter. Oncogenic kinase signalling. *Nature*, 2001. 26

R S Bohacek, C McMartin, and W C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 1996. 3

G Bollag, P Hirth, J Tsai, J Zhang, P N Ibrahim, H Cho, W Spevak, C Zhang, Y Zhang, G Habets, E A Burton, B Wong, G Tsang, B L West, B Powell, R Shellooe, A Marimuthu, H Nguyen, K Y J Zhang, D R Artis, J Schlessinger, F Su, B Higgins, R Iyer, K D'Andrea, A Koehler, M Stumm, P S Lin, R J Lee, J Grippo, I Puzanov, K B Kim, A Ribas, G A McArthur, J A Sosman, P B Chapman, K T Flaherty, X Xu, K L Nathanson, and K Nolop. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature*, 2010. 3, 7

G Bollag, J Tsai, J Zhang, C Zhang, P Ibrahim, K Nolop, and P Hirth. Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature reviews Drug discovery*, 2012. 3

K A Borkovich, F W Farrelly, D B Finkelstein, J Taulien, and S Lindquist. hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. *Molecular and cellular biology*, 1989. 28

I Borodina, P Krabben, and J Nielsen. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome research*, 2005. 26

W Bourguet, M Ruff, P Chambon, H Gronemeyer, and D Moras. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. *Nature*, 1995. 108

T Bremi, R Brüschweiler, and R R Ernst. A Protocol for the Interpretation of Side-Chain Dynamics Based on NMR Relaxation: Application to Phenylalanines in Antamanide. *Journal of the American Chemical Society*, 1997. 90

R Brenke, D Kozakov, G Y Chuang, D Beglov, D Hall, M R Landon, C Mattos, and S Vajda. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*, 2009. 15

C O Brostrom, J D Corbin, C A King, and E G Krebs. Interaction of the subunits of adenosine 3':5'-cyclic monophosphate-dependent protein kinase of muscle. *Proceedings of the National Academy of Sciences*, 1971. 17

R Brüschweiler, B Roux, M Blackledge, C Griesinger, M Karplus, and R R Ernst. Influence of rapid intramolecular motion on NMR cross-relaxation rates. A molecular dynamics study of antamanide in solution. *Journal of the American Chemical Society*, 1992. 38, 44, 90

T Carlomagno. Ligand-target interactions: what can we learn from NMR? *Annual review of biophysics and biomolecular structure*, 2005. 9, 10, 34

T Carlomagno. NMR in natural products: understanding conformation, configuration and receptor interactions. *Natural product reports*, 2012. 34, 35, 36

T Carlomagno, I C Felli, M Czech, R Fischer, M Sprinzl, and C Griesinger. Transferred Cross-Correlated Relaxation: Application to the Determination of Sugar Pucker in an Aminoacylated tRNA-Mimetic Weakly Bound to EF-Tu. *Journal of the American Chemical Society*, 1999. 34

R A E Carr, M Congreve, C W Murray, and D C Rees. Fragment-based lead discovery: leads by design. *Drug discovery today*, 2005. 35

O Carugo and D Bordo. How many water molecules can be detected by protein crystallography? *Acta crystallographica Section D*, 1999. 146

D A Case. Molecular dynamics and NMR spin relaxation in proteins. *Accounts of chemical research*, 2002. 47

J Cavanagh, W J Fairbrother, A G Palmer, and N J Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Elsevier Science, 2007. 82

A Chadli, I Bouhouche, W Sullivan, B Stensgard, N McMahon, M G Catelli, and D O Toft. Dimerization and N-terminal domain proximity underlie the function of the molecular chaperone heat shock protein 90. *Proceedings of the National Academy of Sciences*, 2000. 31, 199

J Chen, C L Brooks, and P E Wright. Model-free analysis of protein dynamics: assessment of accuracy and model selection protocols based on molecular dynamics simulation. *Journal of Biomolecular NMR*, 2004. 47, 72

H C Cheng, S M Van Patten, A J Smith, and D A Walsh. An active twenty-amino-acid-residue peptide derived from the inhibitor protein of the cyclic AMP-dependent protein kinase. *The Biochemical journal*, 1985. 19

H C Cheng, B E Kemp, R B Pearson, A J Smith, L Misconi, S M Van Patten, and D A Walsh. A potent synthetic peptide inhibitor of the cAMP-dependent protein kinase. *The Journal of biological chemistry*, 1986. 19

L K Chico, L J Van Eldik, and D M Watterson. Targeting protein kinases in central nervous system disorders. *Nature reviews Drug discovery*, 2009. 17

H K Choi and K Lee. Recent updates on the development of ganetespib as a Hsp90 inhibitor. *Archives of pharmacal research*, 2012. 32

S Chung, J B Parker, M Bianchet, L M Amzel, and J T Stivers. Impact of linker strain and flexibility in the design of a fragment-based inhibitor. *Nature chemical biology*, 2009. 7

M Cianci, P J Rizkallah, A Olczak, J Raftery, N E Chayen, P F Zagalsky, and J R Helliwell. Structure of lobster apocrustacyanin A1 using softer X-rays. *Acta crystallographica Section D*, 2001. 108

A Ciulli, G Williams, A G Smith, T L Blundell, and C Abell. Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. *Journal of medicinal chemistry*, 2006. 6, 14

M Clark, R D Cramer, and N Van Opdenbosch. Validation of the general purpose tripos 5.2 force field. *Journal of computational chemistry*, 1989. 115, 156

G M Clore and A M Gronenborn. Theory of the time dependent transferred nuclear Overhauser effect: Applications to structural analysis of ligand-protein complexes in solution. *Journal of Magnetic Resonance*, 1983. 35

J Cohen, A Arkhipov, R Braun, and K Schulten. Imaging the migration pathways for O_2, CO, NO, and Xe inside myoglobin. *Biophysical journal*, 2006. 119, 120, 135, 142, 143, 153, 155

P Cohen. Protein kinases – the major drug targets of the twenty-first century? *Nature reviews Drug discovery*, 2002. 17

P Cohen. Targeting protein kinases for the development of anti-inflammatory drugs. *Current opinion in cell biology*, 2009. 17

M Congreve, R Carr, C Murray, and H Jhoti. A 'rule of three' for fragment-based lead discovery? *Drug discovery today*, 2003. 3

K L Constantine, M S Friedrichs, M Wittekind, H Jamil, C H Chu, R A Parker, V Goldfarb, L Mueller, and B T Farmer. Backbone and side chain dynamics of uncomplexed human adipocyte and muscle fatty acid-binding proteins. *Biochemistry*, 1998. 55

S Cooper, F Khatib, A Treuille, J Barbero, J Lee, M Beenen, A Leaver-Fay, D Baker, Z Popovic, and Foldit Players. Predicting protein structures with a multiplayer online game. *Nature*, 2010. 105

G Cornilescu, J L Marquardt, M Ottiger, and A Bax. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *Journal of the American Chemical Society*, 1998. 116

G E Crooks, G Hon, J M Chandonia, and S E Brenner. WebLogo: a sequence logo generator. *Genome research*, 2004. 169

D W Cruickshank. Remarks about protein structure precision. *Acta crystallographica Section D*, 1999. 193, 201

J J Cui, M Tran-Dube, H Shen, M Nambu, P P Kung, M Pairish, L Jia, J Meng, L Funk, I Botrous, M McTigue, N Grodsky, K Ryan, E Padrique, G Alton, S Timofeevski, S Yamazaki, Q Li, H Zou, J Christensen, B Mroczkowski, S Bender, R S Kania, and M P Edwards. Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal-epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). *Journal of medicinal chemistry*, 2011. 229

S C Cullen and E G Gross. The anesthetic properties of xenon in animals and human beings, with additional observations on krypton. *Science*, 1951. 107

G D Dalton and W L Dewey. Protein kinase inhibitor peptide (PKI): a family of endogenous neuropeptides that modulate neuronal cAMP-dependent protein kinase function. *Neuropeptides*, 2006. 19

C Dalvit. NMR methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug discovery today*, 2009. 9

C Dalvit, G Fogliatto, A Stewart, M Veronesi, and B Stockman. WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability. *Journal of Biomolecular NMR*, 2001. 9

A C Dar and K M Shokat. The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annual review of biochemistry*, 2011. 27

A C Dar, M S Lopez, and K M Shokat. Small molecule recognition of c-Src via the Imatinib-binding conformation. *Chemistry & biology*, 2008. 23, 27, 28

D R Davies, B Mamat, O T Magnusson, J Christensen, M H Haraldsson, R Mishra, B Pease, E Hansen, J Singh, D Zembower, H Kim, A S Kiselyov, A B Burgin, M E Gurney, and L J Stewart. Discovery of leukotriene A4 hydrolase inhibitors using metabolomics biased fragment crystallography. *Journal of medicinal chemistry*, 2009. 5

T G Davies and I J Tickle. Fragment screening using X-ray crystallography. *Topics in current chemistry*, 2012. 2, 4, 8, 11, 12

M I Davis, J P Hunt, S Herrgard, P Ciceri, L M Wodicka, G Pallares, M Hocker, D K Treiber, and P P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 2011. 27, 28, 229

G E de Kloe, D Bailey, R Leurs, and I J P de Esch. Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug discovery today*, 2009. 3

J G Demaille, K A Peters, and E H Fischer. Isolation and properties of the rabbit skeletal muscle protein inhibitor of adenosine 3',5'-monophosphate dependent protein kinases. *Biochemistry*, 1977. 19

S Dennis, T Kortvelyesi, and S Vajda. Computational mapping identifies the binding sites of organic solvents on proteins. *Proceedings of the National Academy of Sciences*, 2002. 15, 149

C Doerig, O Billker, D Pratt, and J Endicott. Protein kinases as targets for antimalarial intervention: Kinomics, structure-based design, transmission-blockade, and targeting host cell enzymes. *Biochimica et biophysica acta*, 2005. 17

A Domling. Small molecular weight protein-protein interaction antagonists: an insurmountable challenge? *Current opinion in chemical biology*, 2008. 145

J S Edwards and B O Palsson. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 2000. 26

R J Ellis. Protein misassembly: macromolecular crowding and molecular chaperones. *Advances in experimental medicine and biology*, 2007. 30

P Emsley, B Lohkamp, W G Scott, and K Cowtan. Features and development of Coot. *Acta crystallographica Section D*, 2010. 189, 215

M Erdelyi, A Navarro-Vazquez, B Pfeiffer, C N Kuzniewski, A Felser, T Widmer, J Gertsch, B Pera, J F Diaz, K H Altmann, and T Carlomagno. The binding mode of side chain- and C3-modified epothilones to tubulin. *ChemMedChem*, 2010. 98

D A Erlanson. Introduction to fragment-based drug discovery. *Topics in current chemistry*, 2012. 2, 7

D A Erlanson, A C Braisted, D R Raphael, M Randal, R M Stroud, E M Gordon, and J A Wells. Site-directed ligand discovery. *Proceedings of the National Academy of Sciences*, 2000. 7

R R Ernst, G Bodenhausen, and A Wokaun. *Principles of NMR in one and two dimensions*. Clarendon Press, 1987. 37

P Evans. Scaling and assessment of data quality. *Acta crystallographica Section D*, 2006. 193, 214, 215, 216

R T Fielding and R N Taylor. Principled design of the modern Web architecture. *ACM Transactions on Internet Technology*, 2002. 154

T Fink and J L Reymond. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of chemical information and modeling*, 2007. 3

T Fink, H Bruggesser, and J L Reymond. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angewandte Chemie*, 2005. 3

H Fischer, I Polikarpov, and A F Craievich. Average protein density is a molecular-weight-dependent function. *Protein science*, 2004. 116

K T Flaherty, I Puzanov, K B Kim, A Ribas, G A McArthur, J A Sosman, P J O'Dwyer, R J Lee, J F Grippo, K Nolop, and P B Chapman. Inhibition of mutated, activated BRAF in metastatic melanoma. *The New England journal of medicine*, 2010. 3, 7

J Forster, I Famili, P Fu, B O Palsson, and J Nielsen. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome research*, 2003. 26

A Gaulton, L J Bellis, A P Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and J P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 2012. 17, 229

T Geppert, F Reisen, M Pillong, V Hähnke, Y Tanrikulu, C P Koch, A M Perna, T B Perez, P Schneider, and G Schneider. Virtual screening for compounds that mimic protein-protein interface epitopes. *Journal of computational chemistry*, 2012. 145

P J Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry*, 1985. 150, 152

O Guvench and A D MacKerell. Computational fragment-based binding site identification by ligand competitive saturation. *PLOS computational biology*, 2009. 15, 152

P J Hajduk and J Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery*, 2007. 3

P J Hajduk, J R Huth, and S W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of medicinal chemistry*, 2005. 6, 14

S J Hamill, A E Meekhof, and J Clarke. The effect of boundary selection on the stability and folding of the third fibronectin type III domain from human tenascin. *Biochemistry*, 1998. 56

D Hanahan and R A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 2011. 28

M M Hann. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*, 2011. 5

M M Hann and T I Oprea. Pursuing the leadlikeness concept in pharmaceutical research. *Current opinion in chemical biology*, 2004. 3

M M Hann, A R Leach, and G Harper. Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 2001. 4, 6, 14

O Hantschel, F Grebien, and G Superti-Furga. The growing arsenal of ATP-competitive and allosteric inhibitors of BCR-ABL. *Cancer research*, 2012. 27

P A Harris, A Boloor, M Cheung, R Kumar, R M Crosby, R G Davis-Ward, A H Epperly, K W Hinkle, R N Hunter, J H Johnson, V B Knick, C P Laudeman, D K Luttrell, R A Mook, R T Nolte, S K Rudolph, J R Szewczyk, A T Truesdale, J M Veal, L Wang, and J A Stafford. Discovery of 5-[[4-[(2,3-dimethyl-2H-indazol-6-yl)methylamino]-2-pyrimidinyl]amino]-2-methyl-b enzenesulfonamide (Pazopanib), a novel and potent vascular endothelial growth factor receptor inhibitor. *Journal of medicinal chemistry*, 2008. 229

F U Hartl and M Hayer-Hartl. Converging concepts of protein folding in vitro and in vivo. *Nature structural & molecular biology*, 2009. 30

M J Hartshorn, C W Murray, A Cleasby, M Frederickson, I J Tickle, and H Jhoti. Fragment-based lead discovery using X-ray crystallography. *Journal of medicinal chemistry*, 2005. 4

M Hendlich, P Lackner, S Weitckus, H Floeckner, R Froschauer, K Gottsbacher, G Casari, and M J Sippl. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *Journal of molecular biology*, 1990. 153

S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 1992. 155

M Hennig, A Ruf, and W Huber. Combining biophysical screening and X-ray crystallography for fragment-based drug discovery. *Topics in current chemistry*, 2012. 6, 8, 13

G Holdgate, S Geschwindner, A Breeze, G Davies, N Colclough, D Temesi, and L Ward. Biophysical methods in drug discovery from small molecule to pharmaceutical. *Methods in molecular biology*, 2013. 13

D S Hong, U Banerji, B Tavana, G C George, J Aaron, and R Kurzrock. Targeting the molecular chaperone heat shock protein 90 (Hsp90): lessons learned and future directions. *Cancer treatment reviews*, 2013. 28, 32

A L Hopkins, C R Groom, and A Alex. Ligand efficiency: a useful metric for lead selection. *Drug discovery today*, 2004. 5

V Hornak, R Abel, A Okur, B Strockbine, A Roitberg, and C Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 2006. 72

J S Hub and B L de Groot. Detection of functional modes in protein dynamics. *PLOS computational biology*, 2009. 33

S J Hubbard and J M Thornton. NACCESS. Computer program, 1993. 155

S R Hubbard, L Wei, L Ellis, and W A Hendrickson. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature*, 1994. 24

W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 1996. 89, 90

T Hunter. Signaling – 2000 and beyond. *Cell*, 2000. 17

M Huse and J Kuriyan. The conformational plasticity of protein kinases. *Cell*, 2002. 22, 24, 25

J R Huth, C Park, A M Petros, A R Kunzer, M D Wendt, X Wang, C L Lynch, J C Mack, K M Swift, R A Judge, J Chen, P L Richardson, S Jin, S K Tahir, E D Matayoshi, S A Dorwin, U S Ladror, J M Severin, K A Walter, D M Bartley, S W Fesik, S W Elmore, and P J Hajduk. Discovery and design of novel Hsp90 inhibitors using multiple fragment-based design strategies. *Chemical biology & drug design*, 2007. 7, 200

C Hyeon, P A Jennings, J A Adams, and J N Onuchic. Ligand-induced global transitions in the catalytic domain of protein kinase A. *Proceedings of the National Academy of Sciences*, 2009. 26

V A Jarymowycz and M J Stone. Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chemical reviews*, 2006. 25

V Jayalakshmi and N Rama Krishna. Complete relaxation and conformational exchange matrix (CORCEMA) analysis of intermolecular saturation transfer effects in reversibly forming ligand-receptor complexes. *Journal of magnetic resonance*, 2002. 34

W P Jencks. On the attribution and additivity of binding energies. *Proceedings of the National Academy of Sciences*, 1981. 2, 7

H Jhoti, A Cleasby, M Verdonk, and G Williams. Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Current opinion in chemical biology*, 2007. 11

D A Johnson, P Akamine, E Radzio-Andzelm, M Madhusudan, and S S Taylor. Dynamics of cAMP-dependent protein kinase. *Chemical reviews*, 2001. 18, 25

B Johnsson, S Lofas, and G Lindquist. Immobilization of proteins to a carboxymethyldextran-modified gold surface for biospecific interaction analysis in surface plasmon resonance sensors. *Analytical biochemistry*, 1991. 13

W L Jorgensen, J Chandrasekhar, J D Madura, R W Impey, and M L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 1983. 88, 146

W Kabsch. XDS. *Acta crystallographica Section D*, 2010. 213

W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 1983. 165, 169

A Kalk and H J C Berendsen. Proton magnetic relaxation and spin diffusion in proteins. *Journal of Magnetic Resonance*, 1976. 91

A Kamal, L Thao, J Sensintaffar, L Zhang, M F Boehm, L C Fritz, and J Burrows. A high-affinity conformation of Hsp90 confers tumour selectivity on Hsp90 inhibitors. *Nature*, 2003. 31

R Karlsson, J Zheng, N Xuong, S S Taylor, and J M Sowadski. Structure of the mammalian catalytic subunit of cAMP-dependent protein kinase and an inhibitor peptide displays an open conformation. *Acta crystallographica Section D*, 1993. 25

M Karplus and J Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 2005. 33

L E Kay, D R Muhandiram, G Wolf, S E Shoelson, and J D Forman-Kay. Correlation between binding and dynamics at SH2 domain interfaces. *Nature structural biology*, 1998. 33

J W Keepers and T L James. A theoretical study of distance determinations from NMR. Two-dimensional nuclear overhauser effect spectra. *Journal of Magnetic Resonance*, 1984. 91

J Kestin, M Sokolov, and W A Wakeham. Viscosity of liquid water in the range −8° C to 150° C. *Journal of Physical and Chemical Reference Data*, 1978. 82

F Khatib, F DiMaio, S Cooper, M Kazmierczyk, M Gilski, S Krzywda, H Zabranska, I Pichova, J Thompson, Z Popovic, M Jaskolski, and D Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 2011. 105

C Kim, N H Xuong, and S S Taylor. Crystal structure of a complex between the catalytic and regulatory (RIalpha) subunits of PKA. *Science*, 2005. 19

C Kim, C Y Cheng, S A Saldanha, and S S Taylor. PKA-I holoenzyme structure reveals a mechanism for cAMP-dependent activation. *Cell*, 2007. 19, 21

B Kirchner and M Reiher. The secret of dimethyl sulfoxide-water mixtures. A quantum chemical study of 1DMSO-*n*water clusters. *Journal of the American Chemical Society*, 2002. 83

D B Kitchen, H Decornez, J R Furr, and J Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 2004. 5, 12, 145

D R Knighton, J H Zheng, L F Ten Eyck, V A Ashford, N H Xuong, S S Taylor, and J M Sowadski. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 1991a. 17, 18

D R Knighton, J H Zheng, L F Ten Eyck, N H Xuong, S S Taylor, and J M Sowadski. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 1991b. 18, 19

D R Knighton, S M Bell, J Zheng, L F Ten Eyck, N H Xuong, S S Taylor, and J M Sowadski. 2.0 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with a peptide inhibitor and detergent. *Acta crystallographica Section D*, 1993. 18

P P Knowles, J Murray-Rust, S Kjaer, R P Scott, S Hanrahan, M Santoro, C F Ibanez, and N Q McDonald. Structure and chemical inhibition of the RET tyrosine kinase domain. *The Journal of biological chemistry*, 2006. 229

D E Knuth. *The art of computer programming, volume 2 (3^{rd} ed.): Seminumerical algorithms*. Addison-Wesley Longman Publishing, 1997. 95

J E Kohn, P V Afonine, J Z Ruscio, P D Adams, and T Head-Gordon. Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLOS computational biology*, 2010. 33

P Kolb and A Caflisch. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *Journal of medicinal chemistry*, 2006. 16

W A Koppensteiner and M J Sippl. Knowledge-based potentials – back to the roots. *Biochemistry (Mosc.)*, 1998. 153

A P Kornev, N M Haste, S S Taylor, and L F Ten Eyck. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences*, 2006. 21, 23, 24

A P Kornev, S S Taylor, and L F Ten Eyck. A helix scaffold for the assembly of active protein kinases. *Proceedings of the National Academy of Sciences*, 2008. 18

J K Kranz and C Schalk-Hihi. Protein thermal shifts to identify low molecular weight fragments. *Methods in enzymology*, 2011. 14

I Krimm. INPHARMA-based identification of ligand binding site in fragment-based drug design. *MedChemComm*, 2012. 98

K Kubicek, S K Grimm, J Orts, F Sasse, and T Carlomagno. The tubulin-bound structure of the antimitotic drug tubulysin. *Angewandte Chemie*, 2010. 98

R Kumar, V P Singh, and K M Baker. Kinase inhibitors for cardiovascular disease. *Journal of molecular and cellular cardiology*, 2007. 17

M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 2007. 155

A Leach. *Molecular Modelling: Principles and Applications (2^{nd} ed.)*. Prentice Hall, 2001. 153

D J Leahy, W A Hendrickson, I Aukhil, and H P Erickson. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, 1992. 53

A L Lee, P F Flynn, and A J Wand. Comparison of ^{2}H and ^{13}C NMR Relaxation Techniques for the Study of Protein Methyl Group Dynamics in Solution. *Journal of the American Chemical Society*, 1999. 55

S L Lee, R S Alexander, A Smallwood, R Trievel, L Mersinger, P C Weber, and C Kettner. New inhibitors of thrombin and other trypsin-like proteases: hydrogen bonding of an aromatic cyano group with a backbone amide of the P1 binding site replaces binding of a basic side chain. *Biochemistry*, 1997. 211

N M Levinson and S G Boxer. Structural and spectroscopic analysis of the kinase inhibitor bosutinib and an isomer of bosutinib binding to the Abl tyrosine kinase domain. *PLOS ONE*, 2012. 229

N M Levinson, O Kuchment, K Shen, M A Young, M Koldobskiy, M Karplus, P A Cole, and J Kuriyan. A Src-like inactive conformation in the abl tyrosine kinase domain. *PLOS biology*, 2006. 22

N M Levinson, M A Seeliger, P A Cole, and J Kuriyan. Structural basis for the recognition of c-Src by its inactivator Csk. *Cell*, 2008. 24

A J Li and R Nussinov. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins*, 1998. 142

E Liepinsh and G Otting. Organic solvents identify specific ligand binding sites on protein surfaces. *Nature biotechnology*, 1997. 14

G Lipari and A Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *Journal of the American Chemical Society*, 1982a. 38, 90

G Lipari and A Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society*, 1982b. 38, 90

C A Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 2001. 3

L T Liu, Y Xu, and P Tang. Mechanistic insights into xenon inhibition of NMDA receptors from MD simulations. *The Journal of Physical Chemistry B*, 2010. 119, 120, 135, 142, 143, 144, 153, 155

Y Liu and N S Gray. Rational design of inhibitors that bind to inactive kinase conformations. *Nature chemical biology*, 2006. 21, 27, 229

R E London. Theoretical analysis of the inter-ligand overhauser effect: a new approach for mapping structural relationships of macromolecular ligands. *Journal of magnetic resonance*, 1999. 91

A D MacKerell, D Bashford, M Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, F T K Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, W E Reiher, B Roux, M Schlenkrich, J C Smith, R Stote, J Straub, M Watanabe, J Wiórkiewicz-Kuczera, D Yin, and M Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, 1998. 15, 88, 117, 119, 120, 143, 153, 156

Madhusudan, P Akamine, N H Xuong, and S S Taylor. Crystal structure of a transition state mimic of the catalytic subunit of cAMP-dependent protein kinase. *Nature structural biology*, 2002. 23, 226

G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science*, 2002. 17

P R L Markwick, T Malliavin, and M Nilges. Structural biology by NMR: structure, dynamics, and interactions. *PLOS computational biology*, 2008. 47

L R Masterson, A Mascioni, N J Traaseth, S S Taylor, and G Veglia. Allosteric cooperativity in protein kinase A. *Proceedings of the National Academy of Sciences*, 2008. 25

L R Masterson, C Cheng, T Yu, M Tonelli, A Kornev, S S Taylor, and G Veglia. Dynamics connect substrate recognition to catalysis in protein kinase A. *Nature chemical biology*, 2010. 25, 73

C Mattos and D Ringe. Locating and characterizing binding sites on proteins. *Nature biotechnology*, 1996. 14

M Mayer and B Meyer. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *Journal of the American Chemical Society*, 2001. 9, 34

A J McClellan, S Tam, D Kaganovich, and J Frydman. Protein quality control: chaperones culling corrupt conformations. *Nature cell biology*, 2005. 30

M A McCoy and D F Wyss. Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. *Journal of the American Chemical Society*, 2002. 34

M McTigue, B W Murray, J H Chen, Y L Deng, J Solowiej, and R S Kania. Molecular conformations, interactions, and properties associated with drug efficiency and clinical performance among VEGFR TK inhibitors. *Proceedings of the National Academy of Sciences*, 2012. 229

A E Meekhof, S J Hamill, V L Arcus, J Clarke, and S M Freund. The dependence of chemical exchange on boundary selection in a fibronectin type III domain from human tenascin. *Journal of molecular biology*, 1998. 56

I Melnikova and J Golden. Targeting protein kinases. *Nature reviews Drug discovery*, 2004. 17, 27

X Y Meng, H X Zhang, M Mezei, and M Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 2011. 145

E G Mimnaugh, C Chavany, and L Neckers. Polyubiquitination and proteasomal degradation of the p185c-erbB-2 receptor protein-tyrosine kinase induced by geldanamycin. *The Journal of biological chemistry*, 1996. 28

D Ming and R Brüschweiler. Prediction of methyl-side chain dynamics in proteins. *Journal of Biomolecular NMR*, 2004. 47

A Mittermaier and L E Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 2006. 33

A Mittermaier, A R Davidson, and L E Kay. Correlation between 2H NMR sidechain order parameters and sequence conservation in globular proteins. *Journal of the American Chemical Society*, 2003. 55

A K Mittermaier and L E Kay. Observing biological dynamics at atomic resolution using NMR. *Trends in biochemical sciences*, 2009. 33

C Moler and C Van Loan. Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, 2003. 94

U Moran, R Phillips, and R Milo. SnapShot: key numbers in biology. *Cell*, 2010. 30

A L Morris, M W MacArthur, E G Hutchinson, and J M Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 1992. 193, 215

C W Murray and T L Blundell. Structural biology in fragment-based drug design. *Current opinion in structural biology*, 2010. 3, 6

G N Murshudov, A A Vagin, and E J Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta crystallographica Section D*, 1997. 189, 214

Z Nahleh, A Tfayli, A Najm, A El Sayed, and Z Nahle. Heat shock proteins in cancer: targeting the 'chaperones'. *Future medicinal chemistry*, 2012. 31

N Narayana, S Cox, X Nguyen-huu, L F Ten Eyck, and S S Taylor. A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility. *Structure*, 1997. 23

L Neckers and P Workman. Hsp90 molecular chaperone inhibitors: are we there yet? *Clinical cancer research*, 2012. 28, 31, 32

S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 1970. 155

M G Newlon, M Roy, D Morikis, D W Carr, R Westphal, J D Scott, and P A Jennings. A novel mechanism of PKA anchoring revealed by solution structures of anchoring complexes. *The EMBO journal*, 2001. 18

F Ni. Recent developments in transferred NOE methods. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 1994. 34

F Ni and H A Scheraga. Use of the Transferred Nuclear Overhauser Effect To Determine the Conformations of Ligands Bound to Proteins. *Accounts of chemical research*, 1994. 35

F H Niesen, H Berglund, and M Vedadi. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature protocols*, 2007. 14

M Nilges, J Habazettl, A T Brunger, and T A Holak. Relaxation matrix refinement of the solution structure of squash trypsin inhibitor. *Journal of molecular biology*, 1991. 37, 91

M E Noble, A Musacchio, M Saraste, S A Courtneidge, and R K Wierenga. Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *The EMBO journal*, 1993. 53

W M Obermann, H Sondermann, A A Russo, N P Pavletich, and F U Hartl. In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *The Journal of cell biology*, 1998. 28

N M O'Boyle, M Banck, C A James, C Morley, T Vandermeersch, and G R Hutchison. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 2011. 156

A Olczak, M Cianci, Q Hao, P J Rizkallah, J Raftery, and J R Helliwell. S-SWAT (softer single-wavelength anomalous technique): potential in high-throughput protein crystallography. *Acta crystallographica Section A*, 2003. 108

J Orts, S K Grimm, C Griesinger, K U Wendt, S Bartoschek, and T Carlomagno. Specific methyl group protonation for the measurement of pharmacophore-specific interligand NOE interactions. *Chemistry*, 2008a. 84, 85

J Orts, J Tuma, M Reese, S K Grimm, P Monecke, S Bartoschek, A Schiffer, K U Wendt, C Griesinger, and T Carlomagno. Crystallography-independent determination of ligand binding modes. *Angewandte Chemie*, 2008b. 34, 35, 36, 39, 40, 41, 42, 50, 62, 83, 91, 92

J Orts, C Griesinger, and T Carlomagno. The INPHARMA technique for pharmacophore mapping: A theoretical guide to the method. *Journal of magnetic resonance*, 2009. 36, 91

J P Overington, B Al-Lazikani, and A L Hopkins. How many drug targets are there? *Nature reviews Drug discovery*, 2006. 17

B Panaretou, C Prodromou, S M Roe, R O'Brien, J E Ladbury, P W Piper, and L H Pearl. ATP binding and hydrolysis are essential to the function of the Hsp90 molecular chaperone in vivo. *The EMBO journal*, 1998. 28

B Panaretou, G Siligardi, P Meyer, A Maloney, J K Sullivan, S Singh, S H Millson, P A Clarke, S Naaby-Hansen, R Stein, R Cramer, M Mollapour, P Workman, P W Piper, L H Pearl, and C Prodromou. Activation of the ATPase activity of Hsp90 by the stress-regulated cochaperone aha1. *Molecular cell*, 2002. 31

S Panjarian, R E Iacob, S Chen, J R Engen, and T E Smithgall. Structure and dynamic regulation of Abl kinases. *The Journal of biological chemistry*, 2013. 24

S Panjikar and P A Tucker. Phasing possibilities using different wavelengths with a xenon derivative. *Journal of Applied Crystallography*, 2002a. 107

S Panjikar and P A Tucker. Use of dry paraffin oil and Panjelly in the xenon derivatization of protein crystals. *Journal of Applied Crystallography*, 2002b. 108

S Panjikar and P A Tucker. Xenon derivatization of halide-soaked protein crystals. *Acta crystallographica Section D*, 2002c. 108

L H Pearl. Hsp90 and Cdc37 – a chaperone cancer conspiracy. *Current opinion in genetics & development*, 2005. 28

L H Pearl and C Prodromou. Structure and in vivo function of Hsp90. *Current opinion in structural biology*, 2000. 31

L H Pearl and C Prodromou. Structure, function, and mechanism of the Hsp90 molecular chaperone. *Advances in protein chemistry*, 2001. 31

L H Pearl and C Prodromou. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annual review of biochemistry*, 2006. 31

L H Pearl, C Prodromou, and P Workman. The Hsp90 molecular chaperone: an open and shut case for treatment. *The Biochemical journal*, 2008. 31

W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 1988. 154

S Pfeiffer, D Fushman, and D Cowburn. Simulated and NMR-derived backbone dynamics of a protein with significant flexibility: a comparison of spectral densities for the betaARK1 PH domain. *Journal of the American Chemical Society*, 2001. 69, 89

J C Phillips, R Braun, W Wang, J Gumbart, E Tajkhorshid, E Villa, C Chipot, R D Skeel, L Kale, and K Schulten. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 2005. 88

D Picard. Preface to Hsp90. *Biochimica et biophysica acta*, 2012. 28

L Polgar. The catalytic triad of serine peptidases. *Cellular and molecular life sciences*, 2005. 211

A N Popov and G P Bourenkov. Choice of data-collection parameters based on statistic modelling. *Acta crystallographica Section D*, 2003. 213

T Prange, M Schiltz, L Pernot, N Colloc'h, S Longhi, W Bourguet, and R Fourme. Exploring hydrophobic sites in proteins with xenon or krypton. *Proteins*, 1998. 99

A Prlic, T A Down, E Kulesha, R D Finn, A Kahari, and T J P Hubbard. Integrating sequence and structural biology with DAS. *BMC bioinformatics*, 2007. 154

C Prodromou and L H Pearl. Structure and functional relationships of Hsp90. *Current cancer drug targets*, 2003. 31

C Prodromou, S M Roe, R O'Brien, J E Ladbury, P W Piper, and L H Pearl. Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell*, 1997. 31, 165

C Prodromou, G Siligardi, R O'Brien, D N Woolfson, L Regan, B Panaretou, J E Ladbury, P W Piper, and L H Pearl. Regulation of Hsp90 ATPase activity by tetratricopeptide repeat (TPR)-domain co-chaperones. *The EMBO journal*, 1999. 31

C Prodromou, B Panaretou, S Chohan, G Siligardi, R O'Brien, J E Ladbury, S M Roe, P W Piper, and L H Pearl. The ATPase cycle of Hsp90 drives a molecular 'clamp' via transient dimerization of the N-terminal domains. *The EMBO journal*, 2000. 31, 199

M Punta, P C Coggill, R Y Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, E L L Sonnhammer, S R Eddy, A Bateman, and R D Finn. The Pfam protein families database. *Nucleic Acids Research*, 2012. 17, 235

M L Quillin and B W Matthews. Generation of noble-gas binding sites for crystallographic phasing using site-directed mutagenesis. *Acta crystallographica Section D*, 2002. 108

D H Rasmussen and A P MacKenzie. Phase diagram for the system water-dimethylsulphoxide. *Nature*, 1968. 81, 82

D C Rees, M Congreve, C W Murray, and R Carr. Fragment-based lead discovery. *Nature reviews Drug discovery*, 2004. 35

M Reese, V M Sanchez-Pedregal, K Kubicek, J Meiler, M J J Blommers, C Griesinger, and T Carlomagno. Structural basis of the activity of the microtubule-stabilizing agent epothilone a studied by NMR spectroscopy in solution. *Angewandte Chemie*, 2007. 35

P Rice, I Longden, and A Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics*, 2000. 155

D Rognan. Fragment-based approaches and computer-aided drug discovery. *Topics in current chemistry*, 2012. 7, 15

S Roughley, L Wright, P Brough, A Massey, and R E Hubbard. Hsp90 inhibitors and drugs from fragment and virtual screening. *Topics in current chemistry*, 2012. 31, 32, 200

S D Roughley and R E Hubbard. How well can fragments explore accessed chemical space? A case study from heat shock protein 90. *Journal of medicinal chemistry*, 2011. 32

B Rupp. *Biomolecular crystallography*. Garland Science, 2009. 99

T Ryckmans, M P Edwards, V A Horne, A M Correia, D R Owen, L R Thompson, I Tran, M F Tutt, and T Young. Rapid assessment of a novel series of selective CB(2) agonists using parallel synthesis protocols: A Lipophilic Efficiency (LipE) analysis. *Bioorganic & medicinal chemistry letters*, 2009. 5

N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 1987. 155

V M Sanchez-Pedregal, M Reese, J Meiler, M J J Blommers, C Griesinger, and T Carlomagno. The INPHARMA method: protein-mediated interligand NOEs for pharmacophore mapping. *Angewandte Chemie*, 2005. 34

V M Sanchez-Pedregal, K Kubicek, J Meiler, I Lyothier, I Paterson, and T Carlomagno. The tubulin-bound conformation of discodermolide derived by NMR studies in solution supports a common pharmacophore model for epothilone and discodermolide. *Angewandte Chemie*, 2006. 98

O Sauer, A Schmidt, and C Kratky. Freeze-Trapping Isomorphous Xenon Derivatives of Protein Crystals. *Journal of Applied Crystallography*, 1997. 107, 108

S A Schichman and R L Amey. Viscosity and local liquid structure in dimethyl sulfoxide-water mixtures. *The Journal of Physical Chemistry*, 1971. 81

C Schneider, L Sepp-Lorenzino, E Nimmesgern, O Ouerfelli, S Danishefsky, N Rosen, and F U Hartl. Pharmacologic shifting of a balance between protein refolding and degradation mediated by Hsp90. *Proceedings of the National Academy of Sciences*, 1996. 28

T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 1990. 165

T R Schneider, A T Brunger, and M Nilges. Influence of internal dynamics on accuracy of protein NMR structures: derivation of realistic model distance data from a long molecular dynamics trajectory. *Journal of molecular biology*, 1999. 44, 94

B P Schoenborn, H C Watson, and J C Kendrew. Binding of xenon to sperm whale myoglobin. *Nature*, 1965. 107

M N Schulz and R E Hubbard. Recent progress in fragment-based lead discovery. *Current opinion in pharmacology*, 2009. 14

B Schwanhausser, D Busse, N Li, G Dittmar, J Schuchhardt, J Wolf, W Chen, and M Selbach. Global quantification of mammalian gene expression control. *Nature*, 2011. 30

C D Schwieters, J J Kuszewski, N Tjandra, and G M Clore. The Xplor-NIH NMR molecular structure determination package. *Journal of magnetic resonance*, 2003. 85, 87

D E Scott, A G Coyne, S A Hudson, and C Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 2012. 4

J B Shabb. Physiological substrates of cAMP-dependent protein kinase. *Chemical reviews*, 2001. 17

Y Shan, M A Seeliger, M P Eastwood, F Frank, H Xu, M O Jensen, R O Dror, J Kuriyan, and D E Shaw. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences*, 2009. 21, 24

C E Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948. 169

S Shoji, D C Parmelee, R D Wade, S Kumar, L H Ericsson, K A Walsh, H Neurath, G L Long, J G Demaille, E H Fischer, and K Titani. Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase. *Proceedings of the National Academy of Sciences*, 1981. 17

S A Showalter and R Brüschweiler. Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field. *Journal of Chemical Theory and Computation*, 2007. 47

S A Showalter, E Johnson, M Rance, and R Brüschweiler. Toward quantitative interpretation of methyl side-chain dynamics from NMR by molecular dynamics simulations. *Journal of the American Chemical Society*, 2007. 47

S B Shuker, P J Hajduk, R P Meadows, and S W Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 1996. 2, 6, 9

J R Sierra, V Cepero, and S Giordano. Molecular mechanisms of acquired resistance to tyrosine kinase targeted therapy. *Molecular cancer*, 2010. 26

G Siligardi, B Panaretou, P Meyer, S Singh, D N Woolfson, P W Piper, L H Pearl, and C Prodromou. Regulation of Hsp90 ATPase activity by the co-chaperone Cdc37p/p50cdc37. *The Journal of biological chemistry*, 2002. 31

C C Smith, E A Lasater, X Zhu, K C Lin, W K Stewart, L E Damon, S Salerno, and N P Shah. Activity of ponatinib against clinically-relevant AC220-resistant kinase domain mutants of FLT3-ITD. *Blood*, 2013. 229

F Solca, G Dahl, A Zoephel, G Bader, M Sanderson, C Klein, O Kraemer, F Himmelsbach, E Haaksma, and G R Adolf. Target binding properties and cellular activity of afatinib (BIBW 2992), an irreversible ErbB family blocker. *The Journal of pharmacology and experimental therapeutics*, 2012. 229

D B Solit and G Chiosis. Development and application of Hsp90 inhibitors. *Drug discovery today*, 2008. 31

S M Soltis, M H B Stowell, M C Wiener, G N Phillips, and D C Rees. Successful flash-cooling of xenon-derivatized myoglobin crystals. *Journal of Applied Crystallography*, 1997. 108

J Stamos, M X Sliwkowski, and C Eigenbrot. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *The Journal of biological chemistry*, 2002. 229

B Stauch, J Orts, and T Carlomagno. The description of protein internal motions aids selection of ligand binding poses by the INPHARMA method. *Journal of Biomolecular NMR*, 2012. 33

C E Stebbins, A A Russo, C Schneider, N Rosen, F U Hartl, and N P Pavletich. Crystal structure of an Hsp90-geldanamycin complex: targeting of a protein chaperone by an antitumor agent. *Cell*, 1997. 30, 32

M J Stocks, S Barber, R Ford, F Leroux, S St-Gallay, S Teague, and Y Xue. Structure-driven HtL: design and synthesis of novel aminoindazole inhibitors of c-Jun N-terminal kinase activity. *Bioorganic & medicinal chemistry letters*, 2005. 39

M Taipale, D F Jarosz, and S Lindquist. Hsp90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature reviews Molecular cell biology*, 2010. 28, 30

M Tao, M L Salas, and F Lipmann. Mechanism of activation by adenosine 3':5'-cyclic monophosphate of a protein phosphokinase from rabbit reticulocytes. *Proceedings of the National Academy of Sciences*, 1970. 17

S S Taylor and A P Kornev. Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 2011. 17, 18, 21, 24, 25, 27

S S Taylor, J Yang, J Wu, N M Haste, E Radzio-Andzelm, and G Anand. PKA: a portrait of protein kinase dynamics. *Biochimica et biophysica acta*, 2004. 17, 21, 24, 25

S S Taylor, C Kim, D Vigil, N M Haste, J Yang, J Wu, and G S Anand. Dynamics of signaling by PKA. *Biochimica et biophysica acta*, 2005. 18

S S Taylor, R Ilouz, P Zhang, and A P Kornev. Assembly of allosteric macromolecular switches: lessons from PKA. *Nature reviews Molecular cell biology*, 2012. 17, 18, 21

R F Tilton, I D Kuntz, and G A Petsko. Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 Å. *Biochemistry*, 1984. 107

J S Tokarski, J A Newitt, C Y J Chang, J D Cheng, M Wittekind, S E Kiefer, K Kish, F Y F Lee, R Borzillerri, L J Lombardo, D Xie, Y Zhang, and H E Klei. The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer research*, 2006. 229

J Travers, S Sharp, and P Workman. Hsp90 inhibition: two-pronged exploitation of cancer dependencies. *Drug discovery today*, 2012. 31

N Trbovic, B Kim, R A Friesner, and A G Palmer. Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation. *Proteins*, 2008. 47

J Tsai, J T Lee, W Wang, J Zhang, H Cho, S Mamo, R Bremer, S Gillette, J Kong, N K Haass, K Sproesser, L Li, K S M Smalley, D Fong, Y L Zhu, A Marimuthu, H Nguyen, B Lam, J Liu, I Cheung, J Rice, Y Suzuki, C Luu, C Settachatgul, R Shellooe, J Cantwell, S H Kim, J Schlessinger, K Y J Zhang, B L West, B Powell, G Habets, C Zhang, P N Ibrahim, P Hirth, D R Artis, M Herlyn, and G Bollag. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proceedings of the National Academy of Sciences*, 2008. 3

UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 2012. 17, 39, 109, 154, 235

J Uppenberg, N Ohrner, M Norin, K Hult, G J Kleywegt, S Patkar, V Waagen, T Anthonsen, and T A Jones. Crystallographic and molecular-modeling studies of lipase B from Candida antarctica reveal a stereospecificity pocket for secondary alcohols. *Biochemistry*, 1995. 212

A A Vaguine, J Richelle, and S J Wodak. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta crystallographica Section D*, 1999. 193

E Valkov, T Sharpe, M Marsh, S Greive, and M Hyvonen. Targeting protein-protein interactions and fragment-based drug discovery. *Topics in current chemistry*, 2012. 3

W F van Gunsteren and H J C Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics*, 1977. 89

O P J van Linden. *Fragment-based drug design of small molecule EPHA4 kinase inhibitors*. PhD thesis, Vrije Universiteit Amsterdam, 2013. 17

G J P van Westen, J K Wegner, A Bender, A P Ijzerman, and H W T van Vlijmen. Mining protein dynamics from sets of crystal structures using "consensus structures". *Protein science*, 2010. 33

K Vanommeslaeghe, E Hatcher, C Acharya, S Kundu, S Zhong, J Shim, E Darian, O Guvench, P Lopes, I Vorobyov, and A D MacKerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry*, 2010. 89

S Vanwetswinkel, R J Heetebrij, J van Duynhoven, J G Hollander, D V Filippov, P J Hajduk, and G Siegal. TINS, target immobilized NMR screening: an efficient and sensitive method for ligand discovery. *Chemistry & biology*, 2005. 9

S Velankar, P McNeil, V Mittard-Runte, A Suarez, D Barrell, R Apweiler, and K Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 2005. 154

L Verlet and J J Weis. Perturbation theory for the thermodynamic properties of simple liquids. *Molecular Physics*, 1972. 142, 144

S Vijay-Kumar, C E Bugg, and W J Cook. Structure of ubiquitin refined at 1.8 Å resolution. *Journal of molecular biology*, 1987. 53

J Vitali, A H Robbins, S C Almo, and R F Tilton. Using xenon as a heavy atom for determining phases in sperm whale metmyoglobin. *Journal of Applied Crystallography*, 1991. 108

B Vögeli, T F Segawa, D Leitz, A Sobol, A Choutko, D Trzesniak, W van Gunsteren, and R Riek. Exact distances and internal dynamics of perdeuterated ubiquitin from NOE buildups. *Journal of the American Chemical Society*, 2009. 47, 49

D Wallach. Effect of Internal Rotation on Angular Correlation Functions. *The Journal of Chemical Physics*, 1967. 38

D A Walsh and S M Van Patten. Multiple pathway signal transduction by the cAMP-dependent protein kinase. *FASEB journal*, 1994. 17

D A Walsh, J P Perkins, and E G Krebs. An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle. *The Journal of biological chemistry*, 1968. 17

R Wang, R Kobayashi, and J M Bishop. Cellular adherence elicits ligand-independent activation of the Met cell-surface receptor. *Proceedings of the National Academy of Sciences*, 1996. 26

W H Ward and G A Holdgate. Isothermal titration calorimetry in drug discovery. *Progress in medicinal chemistry*, 2001. 13

D B Weiner, J Liu, J A Cohen, W V Williams, and M I Greene. A point mutation in the neu oncogene mimics ligand induction of receptor aggregation. *Nature*, 1989. 26

I B Weinstein and A Joe. Oncogene addiction. *Cancer research*, 2008. 26, 31

E Weisberg, P W Manley, W Breitenstein, J Bruggen, S W Cowan-Jacob, A Ray, B Huntly, D Fabbro, G Fendrich, E Hall-Meyers, A L Kung, J Mestan, G Q Daley, L Callahan, L Catley, C Cavazza, M Azam, D Neuberg, R D Wright, D G Gilliland, and J D Griffin. Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer cell*, 2005. 229

M Weisel, E Proschak, and G Schneider. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal*, 2007. 135, 151

B P Welford. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 1962. 95

S Whitehouse and D A Walsh. Purification of a physiological form of the inhibitor protein of the cAMP-dependent protein kinase. *The Journal of biological chemistry*, 1982. 19

S Whitehouse and D A Walsh. $Mg \cdot ATP^{2-}$-dependent interaction of the inhibitor protein of the cAMP-dependent protein kinase with the catalytic subunit. *The Journal of biological chemistry*, 1983. 19

L Whitesell and N U Lin. Hsp90 as a platform for the assembly of more effective cancer chemotherapy. *Biochimica et biophysica acta*, 2012. 28

L Whitesell, E G Mimnaugh, B De Costa, C E Myers, and L M Neckers. Inhibition of heat shock protein Hsp90-pp60v-src heteroprotein complex formation by benzoquinone ansamycins: essential role for stress proteins in oncogenic transformation. *Proceedings of the National Academy of Sciences*, 1994. 32

M D Winn, C C Ballard, K D Cowtan, E J Dodson, P Emsley, P R Evans, R M Keegan, E B Krissinel, A G W Leslie, A McCoy, S J McNicholas, G N Murshudov, N S Pannu, E A Potterton, H R Powell, R J Read, A Vagin, and K S Wilson. Overview of the CCP4 suite and current developments. *Acta crystallographica Section D*, 2011. 192, 193, 214, 215

E R Wood, A T Truesdale, O B McDonald, D Yuan, A Hassell, S H Dickerson, B Ellis, C Pennisi, E Horne, K Lackey, K J Alligood, D W Rusnak, T M Gilmer, and L Shewchuk. A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer research*, 2004. 229

J M Word, S C Lovell, J S Richardson, and D C Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, 1999. 88, 155

P Workman. Combinatorial attack on multistep oncogenesis by inhibiting the Hsp90 molecular chaperone. *Cancer letters*, 2004. 28

L Wright, X Barril, B Dymock, L Sheridan, A Surgenor, M Beswick, M Drysdale, A Collier, A Massey, N Davies, A Fink, C Fromont, W Aherne, K Boxall, S Sharp, P Workman, and R E Hubbard. Structure-activity relationships in purine-based inhibitor binding to Hsp90 isoforms. *Chemistry & biology*, 2004. 187, 189, 200, 213, 215

D F Wyss, Y S Wang, H L Eaton, C Strickland, J H Voigt, Z Zhu, and A W Stamford. Combining NMR and X-ray crystallography in fragment-based drug

discovery: discovery of highly potent and selective BACE-1 inhibitors. *Topics in current chemistry*, 2012. 9

W Xu, A Doshi, M Lei, M J Eck, and S C Harrison. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Molecular cell*, 1999. 24

Z Xu, D A Bernlohr, and L J Banaszak. The adipocyte lipid-binding protein at 1.6 Å resolution. Crystal structures of the apoprotein and with bound saturated and unsaturated fatty acids. *The Journal of biological chemistry*, 1993. 53

J Yang, P Cron, V Thompson, V M Good, D Hess, B A Hemmings, and D Barford. Molecular mechanism for the regulation of protein kinase B/Akt by hydrophobic motif phosphorylation. *Molecular cell*, 2002. 24

H Yin and A D Hamilton. Strategies for targeting protein-protein interactions with synthetic agents. *Angewandte Chemie*, 2005. 145

P Yip and D A Case. A new method for refinement of macro molecular structures based on nuclear overhauser effect spectra. *Journal of Magnetic Resonance*, 1989. 85

C H Yun, T J Boggon, Y Li, M S Woo, H Greulich, M Meyerson, and M J Eck. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer cell*, 2007. 229

J Zhang, P L Yang, and N S Gray. Targeting cancer with small molecule kinase inhibitors. *Nature reviews Cancer*, 2009. 17, 27, 229

P Zhang, E V Smith-Nguyen, M M Keshwani, M S Deal, A P Kornev, and S S Taylor. Structure and allostery of the PKA RIIβ tetrameric holoenzyme. *Science*, 2012. 18, 19

J Zheng, D R Knighton, N H Xuong, S S Taylor, J M Sowadski, and L F Ten Eyck. Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations. *Protein science*, 1993a. 25

J Zheng, E A Trafny, D R Knighton, N H Xuong, S S Taylor, L F Ten Eyck, and J M Sowadski. 2.2 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. *Acta crystallographica Section D*, 1993b. 19

M Zheng, Y Li, B Xiong, H Jiang, and J Shen. Water PMF for predicting the properties of water molecules in protein binding site. *Journal of computational chemistry*, 2012. 151

J Zhou and J A Adams. Participation of ADP dissociation in the rate-determining step in cAMP-dependent protein kinase. *Biochemistry*, 1997. 25

T Zhou, L Commodore, W S Huang, Y Wang, M Thomas, J Keats, Q Xu, V M Rivera, W C Shakespeare, T Clackson, D C Dalgarno, and X Zhu. Structural mechanism of the Pan-BCR-ABL inhibitor ponatinib (AP24534): lessons for overcoming kinase inhibitor resistance. *Chemical biology & drug design*, 2011. 229

Z Zou, L Cao, P Zhou, Y Su, Y Sun, and W Li. Hyper-acidic protein fusion partners improve solubility and assist correct folding of recombinant proteins expressed in Escherichia coli. *Journal of biotechnology*, 2008. 30

F Zuccotto, E Ardini, E Casale, and M Angiolini. Through the "gatekeeper door": exploiting the active kinase conformation. *Journal of medicinal chemistry*, 2010. 27, 229