

# *In Silico* Strategies to Modulate DNA Damage Response



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Vom Fachbereich Biologie  
der Technischen Universität Darmstadt  
zur Erlangung des akademischen Grades eines  
*Doctor rerum naturalium*  
genehmigte Dissertation von

**Dipl.-Biol. Sabine Knorr**

aus Flörsheim am Main

1. Referent: Prof. Dr. Kay Hamacher

2. Referent: Prof. Dr. Gerhard Thiel

Tag der Einreichung: 8. August 2014

Tag der mündlichen Prüfung: 6. Oktober 2014

Darmstadt 2015

D17

---



*Für Papa*



# Summary

The objective of this work was the investigation of DNA damage response in irradiated tumor cells at molecular level. Using different computational (*in silico*) approaches three proteins and protein complexes relevant to radiation biology were analyzed in terms of structural-dynamic and evolutionary aspects.

Inhibition of the **20S proteasome** was shown to selectively sensitize tumor cells to radiation-induced DNA damage in contrast to surrounding healthy tissue. Thus, proteasome inhibitors have a great potential as radiosensitizing agents. Simulation of the enzyme inhibition (protein-ligand docking) allows to investigate a ligand's conformation inside the  $\beta 5$  active site of the proteasome thereby supporting proteasome inhibitor optimization. In this study inhibitors are considered that form a covalent bond to the  $\beta 5$  active site. Here, modeling of the covalent interaction presented a major challenge. In collaboration with medicinal chemists of the lab of Prof. Schmidt (TU Darmstadt) it was possible to establish a general procedure for the docking of covalently bound ligands using the software MOE. In an extended cooperation with crystallographers (Prof. Groll, TU München) and biochemists (Prof. Kloetzel, Charité Berlin) we have succeeded in developing potent and highly selective  $\alpha$ -keto phenylamide proteasome inhibitors. These are characterized through a unique binding mode to the primed sites of the substrate binding channel.

In a follow-up project of this collaboration we focused on the elucidation of the natural proteolysis mechanism in the  $\beta 5$  subunit. It was hypothesized that substrates with large hydrophobic P1 residues interact with the Met45 side chain located in the  $\beta 5$  active site thus accelerating the cleavage mechanism. This trigger mechanism is considered to be analogous to that of a mouse trap. However, biochemical experiments have shown that the cleavage mechanism is also triggered by small residues which contradicts the hypothesis. Our results of extensive molecular dynamics (MD) simulations confirmed the Met45 side chain dynamics to be directly related to the substrate's residue size. This shows that the  $\beta 5$  binding pocket accommodates to a variety of differently-sized substrate residues. These findings allow for future rational design of proteasome inhibitors.

The DNA-dependent protein kinase (**DNA-PK**) consisting of the protein Ku (with subunits Ku70 and Ku80) and its catalytic subunit (DNA-PKcs) is the key complex responsible for DNA double-strand break (DSB) repair at early stages of the non-homologous recombination pathway (NHEJ). For the development of DNA-PK inhibitors which modulate the NHEJ pathway, structural knowledge of the DNA-PK complex is mandatory. This work addresses the question if molecular coevolution can provide information on the yet unknown three-dimensional architecture of the DNA-PK complex. Molecular coevolution defines the mutual evolution of

---

interacting amino acid residues located at interaction interfaces in order to ensure residue recognition and complex stability. The mutual information (MI), an information-theoretical measure, was applied to detect coevolutionary signals in sets of homologous sequences. Different MI correction procedures were evaluated with respect to their ability to predict interacting residues in the Ku70/Ku80 complex of which the crystal structure is known. It turned out that prediction quality is limited. Results show that the procedures tested must be further enhanced to extract those signals relevant to protein-protein interactions from other background signals.

In order to interfere with the homologous recombination pathway (HR), **Rad54** is a straightforward target since this protein plays a substantial role in this DNA repair pathway. This project was developed in collaboration with radiation biologists of the lab of Prof. Löbrich (TU Darmstadt) who discovered a specific phosphorylation reaction to be necessary for Rad54 activation. Using a reduced biophysical network model, putative phosphorylation-induced structural changes were investigated. Instead of the expected local response, dynamics intrinsic to the protein domain were observed. Most probably, these are related to the principal function of Rad54, namely the translocation along double-stranded DNA. This result led to a more intensive study on the structural basis of the translocation cycle which might be reproduced by currently available crystal structures. Based on these findings we were able to construct a three-dimensional model of zebrafish Rad54 which serves as a starting structure for further studies.

# Zusammenfassung

Diese Arbeit hatte zum Ziel, die DNA-Schadensantwort in bestrahlten Tumorzellen auf molekularer Ebene zu erforschen. Mithilfe von verschiedenen computergestützten (*in silico*) Verfahren wurden strukturdynamische und evolutionäre Aspekte von drei strahlenbiologisch relevanten Proteinen bzw. Proteinkomplexen untersucht.

Durch Inhibierung des **20S Proteasoms** kann erreicht werden, dass Tumorzellen auf strahleninduzierte DNA-Schäden sensitiver reagieren als umliegendes, gesundes Gewebe. Proteasominhibitoren besitzen daher großes Potential als sogenannte radiosensitivierende Substanzen. Mithilfe einer computergestützten Simulation der Inhibition am Enzym (Protein-Ligand Docking) kann die Konformation eines Liganden in der  $\beta$ 5-Bindetasche des Proteasoms untersucht und die Wirkstoffoptimierung unterstützt werden. Eine Herausforderung stellte hier die Modellierung der kovalenten Bindung dar, welche zwischen den von uns betrachteten Inhibitoren und der  $\beta$ 5-Bindetasche ausgebildet wird. In Zusammenarbeit mit Medizinalchemikern der Arbeitsgruppe von Prof. Schmidt (TU Darmstadt) gelang es, ein universelles Verfahren für das Docking von kovalent gebundenen Inhibitoren in der Software MOE zu etablieren. In einer erweiterten Kooperation mit Kristallographen (Prof. Groll, TU München) und Biochemikern (Prof. Kloetzel, Charité Berlin) gelang die Entwicklung von potenten und hochselektiven Proteasominhibitoren aus der Substanzklasse der  $\alpha$ -Keto-Phenylamide. Diese zeichnet sich durch eine einzigartige Ausnutzung der sogenannten *primed sites* des Substratbindekanals aus.

Ein weiterführendes Projekt dieser Kooperation hatte die Charakterisierung des natürlichen Proteolyse-Mechanismus in der  $\beta$ 5-Bindetasche zum Ziel. Es wurde die Hypothese postuliert, dass die Spaltungsreaktion ausschließlich durch lange, hydrophobe Reste des Substrats über Wechselwirkungen mit der Met45-Seitenkette der  $\beta$ 5-Bindetasche ausgelöst wird, vergleichbar mit der Funktionsweise einer Mausefalle. Diese Hypothese wurde jedoch durch weitere experimentelle Ergebnisse, die zeigten, dass auch kurze Reste die Spaltungsreaktion auslösen können, widerlegt. Durch unsere umfangreichen Molekulardynamik (MD) Simulationen konnte bestätigt werden, dass die Dynamik der Met45-Seitenkette direkt von der Länge des Substratrests abhängt. Dies zeigt, dass die  $\beta$ 5-Bindetasche Reste verschiedener Länge aufnehmen kann, denen sie sich jeweils anpasst. Diese Erkenntnisse ermöglichen ein zukünftiges, rationales Design von Proteasominhibitoren.

Der DNA-abhängige Protein-Kinase-Komplex (**DNA-PK**), bestehend aus den Proteinen Ku (mit den Untereinheiten Ku70 und Ku80) und dem katalytisch aktiven DNA-PKcs, bildet das strukturelle Grundgerüst für die Reparatur von DNA-Doppelstrangbrüchen in der frühen Phase des Reparaturwegs der nicht-homologen Endverknüpfung (NHEJ). Für die Entwicklung von DNA-PK-Inhibitoren, die der

---

NHEJ-Modulation dienen, ist strukturelles Wissen über den DNA-PK Komplex als solchen unabdingbar. Diese Arbeit beschäftigte sich daher mit der grundlegenden Fragestellung, ob molekulare Koevolution Aufschluss über den noch unbekanntem dreidimensionalen Aufbau des DNA-PK-Komplexes geben kann. Molekulare Koevolution bezeichnet die wechselseitige Anpassung von interagierenden Aminosäuren an der Oberfläche der am Komplex beteiligten Proteine, um die gegenseitige Erkennung und somit die Stabilität des Komplexes zu gewährleisten. Als ein informationstheoretisches Maß wurde die *Mutual Information* (MI) verwendet, um in homologen Sequenzdaten solche koevolutionären Signale zu detektieren. Verschiedene Korrekturverfahren der MI wurden bezüglich ihrer Vorhersagekraft von interagierenden Aminosäuren im Ku70/Ku80-Komplex, welcher aus der Kristallstruktur bereits bekannt ist, evaluiert. Es stellte sich heraus, dass die Vorhersage von Interaktionen nur beschränkt möglich ist. Die Ergebnisse zeigen, dass die verwendeten Verfahren weiter dahingehend optimiert werden müssen, die gewünschten Signale der Protein-Protein-Interaktion von dem Hintergrund weiterer Signale zu trennen.

Die Modulation des Proteins **Rad54** eröffnet die Möglichkeit, in die homologe Rekombination (HR) einzugreifen, da dem Protein in diesem DNA-Reparaturweg eine zentrale Schlüsselrolle zukommt. Dieses Projekt entstand in Kollaboration mit Strahlenbiologen der Arbeitsgruppe von Prof. Löbrich (TU Darmstadt), welche zeigen konnten, dass eine bestimmte Phosphorylierungsreaktion notwendig ist, um Rad54 zu aktivieren. Mit einem reduzierten biophysikalischen Netzwerkmodell wurde untersucht, welche strukturellen Änderungen durch die Phosphorylierung induziert werden. Anstelle der erwarteten lokalen Antwort wurde eine intrinsische Dynamik der Proteindomänen beobachtet, die vermutlich eng mit der hauptsächlichen Funktion von Rad54 verwandt ist: der Translokation entlang des DNA-Doppelstrangs. Dieses Resultat führte zu einer intensiveren Auseinandersetzung mit den strukturellen Grundlagen des Translokationszyklus. Dieser konnte durch die aktuell verfügbaren Kristallstrukturen dargestellt werden. Auf Basis dieser Erkenntnisse wurde ein dreidimensionales Modell des Rad54-Proteins aus dem Zebrafisch erstellt, welches nun als Ausgangsstruktur für weitere Studien dient.

# Contents

<b>Summary</b>	<b>ii</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>List of Contributions</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The Proteasome</b>	<b>5</b>
2.1 Background . . . . .	6
2.1.1 Molecular Biology . . . . .	6
2.1.2 Structure . . . . .	7
2.1.3 Proteasome Inhibitors . . . . .	9
2.2 Project I: Molecular Docking Studies . . . . .	12
2.2.1 Introduction . . . . .	12
2.2.2 Theory: Molecular Docking . . . . .	12
2.2.3 Methods . . . . .	14
2.2.4 Results & Discussion . . . . .	18
2.2.5 Conclusion . . . . .	27
2.3 Project II: Inhibitor Optimization . . . . .	30
2.3.1 Contributions . . . . .	30
2.3.2 Introduction . . . . .	30
2.3.3 Methods . . . . .	31
2.3.4 Results & Discussion . . . . .	32
2.3.5 Conclusion . . . . .	34
2.4 Project III: Proteasome Substrate Cleavage Mechanism . . . . .	35
2.4.1 Contributions . . . . .	35
2.4.2 Introduction . . . . .	35
2.4.3 Theory: Molecular Dynamics Simulation . . . . .	39
2.4.4 Methods . . . . .	41
2.4.5 Results & Discussion . . . . .	43
2.4.6 Conclusion . . . . .	48
<b>3 The DNA-PK Complex</b>	<b>51</b>
3.1 Background . . . . .	53
3.1.1 Molecular Biology . . . . .	53
3.1.2 Structure . . . . .	53

3.2	Theory: Mutual Information . . . . .	58
3.3	Methods . . . . .	62
3.4	Results & Discussion . . . . .	68
3.4.1	Intra Mutual Information . . . . .	68
3.4.2	Inter Mutual Information . . . . .	71
3.5	Conclusion . . . . .	81
<b>4</b>	<b>The Rad54 Protein</b>	<b>85</b>
4.1	Background . . . . .	86
4.1.1	Molecular Biology . . . . .	86
4.1.2	Structure . . . . .	86
4.2	Theory: Linear Response Theory . . . . .	89
4.3	Methods . . . . .	91
4.4	Results & Discussion . . . . .	96
4.5	Conclusion . . . . .	107
<b>5</b>	<b>Conclusion</b>	<b>111</b>
<b>A</b>	<b>Appendix</b>	<b>113</b>
	<b>List of Abbreviations</b>	<b>129</b>
	<b>Bibliography</b>	<b>133</b>
	<b>Danksagung</b>	<b>150</b>
	<b>Ehrenwörtliche Erklärung</b>	<b>152</b>
	<b>Curriculum Vitae</b>	<b>154</b>

# List of Contributions

Parts of this work have been published in journals, prepared as manuscripts or presented as conference posters:

- Voss C., Scholz C., Knorr S., Beck P., Stein M., Zall A., Kuckelkorn U., Kloetzel P.M., Groll M., Hamacher K. and Schmidt B.  
 **$\alpha$ -Keto Phenylamides as P1'-Extended Proteasome Inhibitors**  
This manuscript was accepted in *ChemMedChem* in June 2014. Contributions: statistical and structural analysis as described in *Project II: Inhibitor Optimization* (Section 2.3), manuscript preparation
- Scholz C., Voss C., Knorr S., Kuckelkorn U., Hamacher K., Kloetzel P.M. and Schmidt B.  
**Paradigm Caught by a Mouse Trap: 20S Proteasome's  $\beta$ 5 Subunit is Not Chymotrypsin-like**  
This manuscript is to be submitted to *Angew. Chem. Int. Ed.* in August 2014. Contributions: setup and analysis of molecular dynamics simulations as described in *Project III: Proteasome Substrate Cleavage Mechanism* (Section 2.4), manuscript preparation
- Knorr S., Keul F. and Hamacher K.  
**Information Theory Reveals Structural Knowledge on DNA Repair Complexes**  
This poster was presented at the 16<sup>th</sup> Annual Meeting of the German Society for Biological Radiation Research 2013 in Darmstadt and was awarded with the poster prize. Contributions: multiple sequence alignment preparation and coevolutionary analysis as described in Chapter 3, poster preparation and presentation
- Scholz C., Knorr S. and Schmidt B.  
**CovDock – A Highly Versatile Step-By-Step Workflow for Covalent Docking and Virtual Screening in MOE**  
This poster was presented at the *Chemical Computing Group Meeting & Conference 2014* in Strasbourg, France. Contributions: Scientific Vector Language (SVL) code implementation in the Molecular Operating Environment (MOE) software as described in *Project I: Molecular Docking* (Subsection 2.2.4)



# 1 Introduction

Cancer is a disease affecting all ages and socio-economic groups. Currently, cancer is among the leading causes of death worldwide, accounting for 8.2 million deaths in 2012 [49]. Anticancer therapies aim at increasing cytotoxicity for tumor cells while simultaneously sparing the surrounding tissue. Most current treatments are based on the generation of DNA damage in cancer cells by exogenous sources including ionizing radiation or chemotherapeutic agents [87].

Shortly after the discovery of X-rays in 1896, the first cancer patient was treated using irradiation [71]. In clinical radiotherapy, patients are exposed to ionizing radiation (IR), for example, photons ( $\gamma$ -rays, X-rays) or charged particles (protons, heavy ions). The high-energy radiation is able to remove tightly bound electrons from atoms, thus, creating ions. The energy that is deposited is enough to break chemical bonds [72].

In biological cells, ionizing radiation induces different types of DNA damage. Among them, DNA double-strand breaks (DSB) are most deleterious for the cell [72]. The complexity and spatial distribution of the lesions determine the efficiency of DNA repair and thereby the therapeutic outcome [122]. Ionizing radiation provokes multiple lesions arranged within clustered damage sites. These damages are structurally and chemically highly complex and have a reduced reparability when compared to that of individual lesions [122].

Cells have developed complex mechanisms to repair DNA damage to maintain genomic integrity and stability. In mammalian cells, DSBs are repaired by two major pathways: non-homologous end joining (NHEJ) and homologous recombination (HR). HR provides a high-fidelity repair for accurate resynthesis of the damaged DNA. Since the complementary sister chromatid is used as a template, this pathway is only available in the S and G2 phase of the cell cycle [170]. NHEJ is the more error-prone repair pathway with a the simple ligation of broken DNA ends but it is available throughout the full cell cycle.

Radiation therapy has various advantages: it is a non-invasive technique that precisely targets tumor tissue in almost any part of the body with minimal damage to healthy cells. In case of solid tumors, a good local tumor control is achieved. In contrast, the delivery of chemotherapeutic agents is usually applied for metastasis control. By targeting all rapidly dividing cells, chemotherapy is a non-selective approach. Healthy cells usually recover, but patients suffer from severe side effects. However, radiation therapy is an expensive and elaborate technology. The knowledge and experience of many specialists such as scientists, physicians and technicians is essential to develop a personalized treatment plan. For particle therapy, large high-energy accelerators are necessary.

In order to enhance the efficiency of radiation therapy, a promising approach is offered by combining the advantages of both radio- and chemotherapy. The delivery of chemotherapeutic agents concurrent with radiotherapy influences the response of tumor cells to ionizing radiation by selectively sensitizing tumor cells to radiation. Chemical compounds that increase the effect of radiation therapy are called radiosensitizing agents [201]. Beyond the expected additive effects [176] a significant enhancement is observed through synergistic ones [201]. Patients benefit from a reduced overall radiation dose and fewer fractions. This combination has another advantage: the therapy remains effective even if the tumor cells develop resistances against one of the components.

The selective targeting of proteins involved in DNA damage response offers several strategies to achieve radiosensitizing effects [12, 201], e.g., the modulation of DSB signaling pathways [128], the enhancement of programmed cell death (apoptosis) [107] or the interference with DNA repair [35, 87]. The latter approach is able to selectively target tumor cells, as their fate relies on proper DNA repair after IR-induced DNA damage. Moreover, repair proteins are often upregulated in tumor cells [20].

An example of a classical chemotherapeutic agent is cisplatin [160], an inorganic, platinum-containing drug that induces cross-linking of DNA and, thus, triggers apoptosis. It is the most widely used radiosensitizing agent that is clinically applied in combination with radiation therapy [201]. Later on, it was revealed that cisplatin interferes with the NHEJ repair pathway by binding to one of the NHEJ key proteins [52, 182]. As it was the case for cisplatin, the radiosensitizing ability of several other chemotherapeutic compounds such as 5-fluorouracil and gemcitabine has been identified afterwards [26, 191]. With detailed insights into the underlying DNA damage response pathways and the three-dimensional structures of the implicated proteins, the development of selective radiosensitizing compounds is now possible using rational drug design approaches.

## Aim of Work

In this thesis, three strategies are used to investigate the DNA damage response pathways after IR-induced DNA damage: the modulation of proteasomal protein degradation and the two major DSB repair pathways NHEJ and HR. This is achieved by combining structural biology and bioinformatics with radiation biology. By using computational (*in silico*) approaches, structural and dynamical features of three molecular systems will be investigated. This work focuses on contributing to the development of radiosensitizing inhibitors. The following three proteins and protein complexes are investigated:

1. **The proteasome** is a multicatalytic protein complex playing a substantial role in protein degradation. Proteasome inhibition was shown to selectively kill cancer cells by disrupting the homeostatic balance within tumor cells triggering apoptosis [159]. Moreover, it was shown that inhibitors also block

---

DNA repair and radiosensitize non-small cell lung cancer [34]. Several candidate compounds are in clinical trials for the treatment of multiple myeloma cells [110]. A major drawback consists in patients suffering from peripheral neuropathy due to undesired off-target effects [5]. Thus, the ultimate goal is to optimize proteasome inhibitors to develop potent agents being highly selective towards its  $\beta 5$  active site. In a classical chemoinformatic approach, molecular protein-ligand docking was used. In order to investigate the substrate cleavage mechanism of the proteasome's  $\beta 5$  active site, molecular dynamics simulations of protein-ligand complexes were performed.

2. **The DNA-PK complex** comprising the Ku protein (heterodimer of Ku70 and Ku80) and the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) is formed at a DSB in the early stages of the NHEJ repair pathway. It acts as a scaffold complex that orchestrates the assembly of various other repair factors [143]. NHEJ is the predominant pathway by which cells repair DSBs; it is estimated to repair up to 85% of IR-induced DSBs [170]. Modulation of the NHEJ pathway is a successful strategy to achieve radiosensitization [87]. At the moment, several DNA-PK inhibitors are in preclinical studies [114, 97]. Even though the structures of the individual protein components of the DNA-PK complex have been elucidated [192, 171], knowledge of the three-dimensional structure of the DNA-PK complex is essential for the development of DNA-PK inhibitors. Here, a sequence-based, information-theoretic approach was developed to identify interacting residues within the complex.
3. **Rad54** is a key protein in homologous recombination (HR) which is the principal repair pathway during the S phase of the cell cycle, and therefore required for the development of radioresistance during this phase [178, 81]. As a consequence, HR inhibitors are currently subject of preclinical studies [90]. Due to the diverse functions of Rad54, inhibition of Rad54 activation is the most promising approach. Prior biochemical results revealed a single phosphorylation site being responsible for Rad54 activation [173]. In order to study the underlying structural consequences that are responsible for the activation, a coarse-grained biophysical model was developed.



## 2 The Proteasome

Since dysregulation of cell cycle control and growth-death balance are hallmarks of cancer, research in the last decade focused on the ubiquitin-proteasome system (UPS), as it is essential for maintaining the protein turnover [4]. In 2004, the Nobel prize was awarded to Ciechanover, Hershko and Rose for the discovery of the ubiquitin-mediated protein degradation [141]. In the UPS, the key proteolytic core complex is the 20S proteasome, that degrades proteins into smaller oligopeptide fragments. Thus, this ubiquitous and abundant protein complex can be considered as a molecular destruction machine.

In cancer therapy, targeting the 20S proteasome by small-molecule inhibitors was shown to be a promising therapeutic approach to selectively destroy cancer cells, while healthy tissue stays unaffected [11, 80]. The highest response to proteasome inhibition was observed in patients with hematological malignancies, e.g., multiple myeloma and mantle cell lymphoma [148].

Bortezomib (Velcade<sup>®</sup>), a boron acid derivative approved by the US Food and Drug Administration in 2003, was the first proteasome inhibitor used in multiple myeloma treatment [105]. In general, proteasome inhibition is much less toxic than standard chemotherapy but especially bortezomib was shown to induce peripheral neuropathy [5]. In order to overcome these severe side effects, there is an ongoing search for new structures [110] and five promising candidate compounds have entered clinical trials. Importantly, inhibitors were shown to be potent radiosensitizers: in combination with radiation therapy, proteasome inhibitors increased sensitivity to ionizing radiation in malignant cells [57].

Binding of small molecules to the catalytic centers of the 20S proteasome in a reversible or irreversible manner leads to deleterious effects for the cell [167]. The consequence of this induced proteasomal dysfunction, in particular for fast cellular growth, which is characteristic for cancer cells, is protein accumulation and hence apoptosis [96]. Interestingly, most eukaryotic proteasome inhibitors were derived from natural products since proteasome inhibition is an effective defense strategy developed in microorganisms [110].

It is a major purpose of this work to understand the underlying structural aspects of 20S proteasome substrate cleavage mechanism and its inhibition for the development of potent and specific proteasome inhibitors that cause fewer side effects. Therefore, three projects are presented that treat different structural aspects of the particular  $\beta 5$  active site of the eukaryotic 20S proteasome: I) molecular docking studies, II) optimization of proteasome inhibitors and III) investigation of the proteasome substrate cleavage mechanism. The topics range from protein-ligand docking of covalently bound inhibitors towards a comprehensive structure-based and rational

inhibitor design approach. In a final step, we have benefited from inhibitor findings and focused on structural mechanisms of the general 20S proteasome's substrate cleavage.

## 2.1 Background

### 2.1.1 Molecular Biology

The proteasome is a multicatalytic protease complex ubiquitous in all three kingdoms of life: bacteria, archaea and eukaryotes [65]. Eukaryotic 26S proteasomes consist of a cylindrical 20S core particle (CP, 700 kDa, also termed 20S proteasome) and two 19S regulatory particles (RP, 900 kDa).

Protein turnover is the well-regulated balance between protein synthesis and degradation, required to ensure a functional proteome [181]. Hence, proteolysis is necessary for maintaining biological homeostasis and regulation of different cellular processes. In eukaryotes, the major non-lysosomal protein degradation pathway is the cytosolic UPS with the proteasome being the main enzymatic component [83]. This protein degradation pathway is ubiquitin- and ATP-dependent. Protein substrates are marked for degradation by a polyubiquitin chain that is recognized by the 19S RP. It is responsible for ubiquitin-chain cleavage, ATPase function and substrate unfolding. The unfolded proteins are degraded in the interior of the 20S proteasome into small polypeptide fragments.

The UPS is the major quality control pathway [33] being responsible for the removal of abnormal, misfolded or improperly assembled proteins. Here, proteasomes act jointly together with chaperones that recognize proteins of non-native structure and pass them to degradation [95]. Beside aberrant proteins, the UPS also degrades numerous regulatory proteins necessary for the regulation of diverse cellular and physiological pathways, e.g., the cell cycle progression via the degradation of regulatory proteins, cyclins and cyclin-independent kinase inhibitors, whose timely destruction is vital for controlled cell division [139]. In general, the efficient removal of short-lived and regulatory proteins permits a rapid metabolic adaptation to new physiological conditions [119]. For example, the activation of the key transcription factor NF- $\kappa$ B which is involved in the inflammatory response, is initiated by the signal-induced degradation of I $\kappa$ B proteins by the proteasome. Free NF- $\kappa$ B enters into the nucleus and induces the expression of several genes involved in promoting cell survival and proliferation [150].

In addition to the described constitutive proteasome, two alternative proteasomes are expressed in vertebrates, the immunoproteasome and the thymoproteasome. They differ by incorporating different sets of catalytic  $\beta$ -subunits and thus exhibit modified cleavage patterns [17]. The immunoproteasome plays an important role in cellular immune response by an enhanced generation of antigenic peptides that are presented to the immune system by the major histocompatibility complex (MHC)

class I molecules. The thymoproteasome is exclusively expressed in cortical thymic epithelial cells and supposed to regulate CD8<sup>+</sup> T cell development [138].

## 2.1.2 Structure

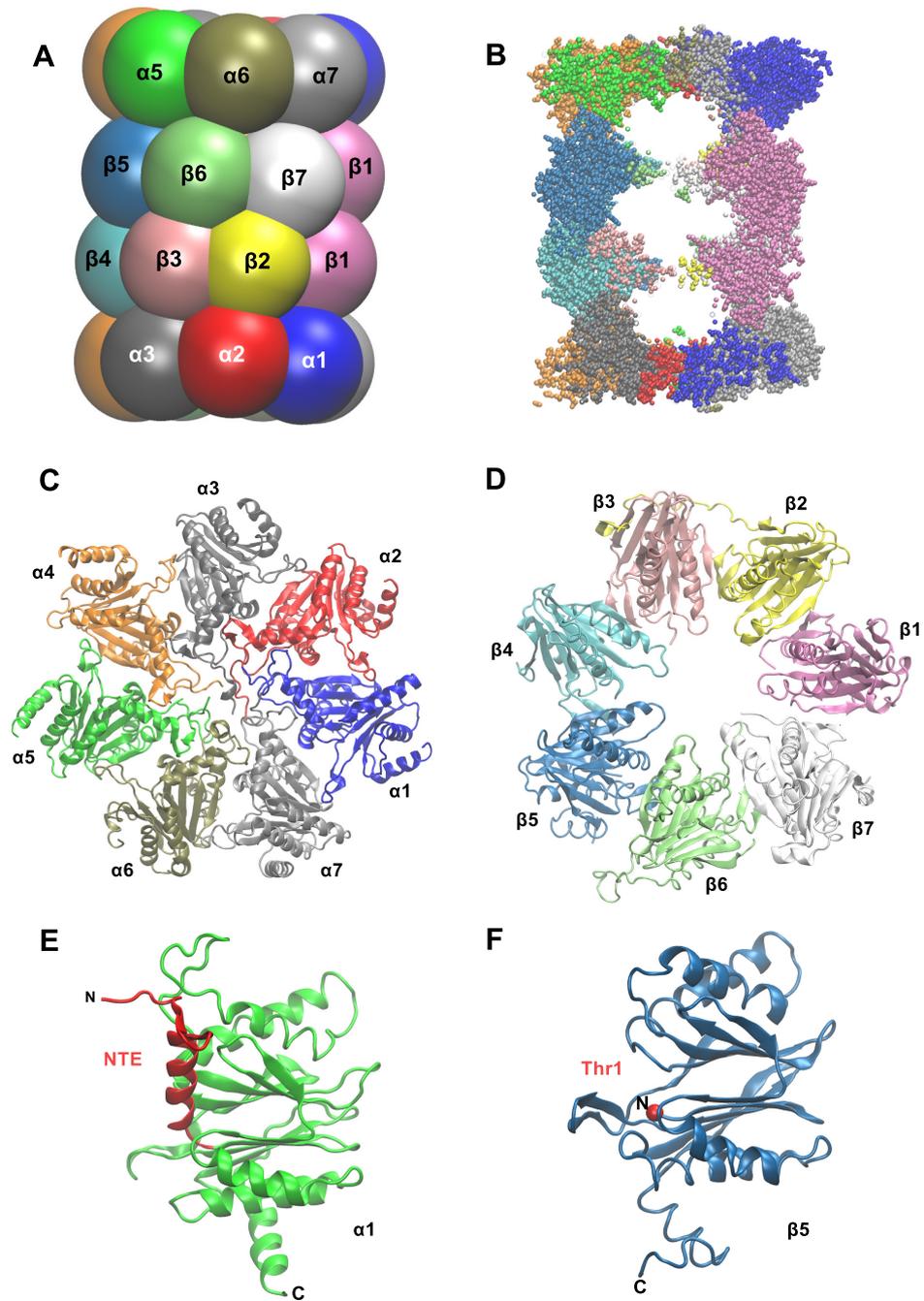
### General Architecture

Whereas the complex architecture of the 19S RP is only roughly known [109], the structure of the 20S CP is well-characterized. The elucidation of the yeast 20S CP crystal structure at a resolution of 2.4 Å [65] provided the first insights into the structural organization of the eukaryotic proteasome that is well-conserved even within higher eukaryotes (Figure 2.1). The CP forms a cylinder having two entrance sites. This barrel is made out of fourteen individual subunits: seven different  $\alpha$ -type subunits ( $\alpha$ 1-7) and seven  $\beta$ -type ( $\beta$ 1-7) ones.

During CP assembly, the  $\alpha$ -subunits form a ring first followed by the addition of single  $\beta$ -subunits to form half proteasomes. These dimerize along the  $\beta$ -rings to form the CP [109]. Consequently, the CP consists of 28 subunits in total that are uniquely arranged in a  $\alpha_{1-7}\beta_{1-7}\beta_{1-7}\alpha_{1-7}$  stoichiometry. This process was shown to be assisted by chaperones that promote specific subunit interactions while blocking other undesired ones [118]. Due to a high sequence similarity, the  $\alpha$ - and  $\beta$ -subunits probably evolved from a common ancestor [203]. Both subunits fold into a  $\beta$ -sandwich structure typical for proteins belonging to the N-terminal nucleophile (Ntn) hydrolases superfamily [65, 145] (Figure 2.1). The subunits of the yeast CP reveal an  $\alpha\beta\beta\alpha$ -core structure of two five-stranded antiparallel  $\beta$ -sheets flanked by two  $\alpha$ -helical layers [65]. Within their classes, the subunits differ in turns, insertions connecting secondary structure elements and in the termini, to ensure specific intersubunit contacts.

The sequences of the  $\alpha$ -subunit's N-terminal extensions are highly conserved and were found to close the barrel-shaped CP on both ends forming an entry gate to the interior of the CP [63] (Figure 2.1). By that fact, the proteasome is an inherently repressed enzyme and binding of activators to the  $\alpha$ -ring is necessary to induce a rearrangement of the N-termini thereby regulating substrate access [175]. In order to ensure substantially unfolded polypeptide chains, the substrate must additionally pass through a narrow channel (13 Å in diameter) termed  $\alpha$ -annulus.

In eukaryotes, three out of the seven different  $\beta$ -subunits ( $\beta$ 1,  $\beta$ 2 and  $\beta$ 5) are catalytically active with the proteolytically active threonine residue 1 (Thr1) located inside the barrel (six active sites in total). In order to prevent undesired proteolysis during assembly, the  $\beta$ 1,  $\beta$ 2 and  $\beta$ 5 polypeptide chains are synthesized as inactive precursor proteins. Only after proteasome assembly they are processed further to the mature forms via intrasubunit autolysis leading to the liberation of the active site [123, 39].



**Figure 2.1:** Crystal structure of the yeast 20S proteasome (PDB code: 1RYP, at 2.4 Å resolution) [65]. A) Schematic view of the overall architecture showing 28 subunits composed of  $\alpha$ 1-7 and  $\beta$ 1-7; each subunit is colored differently. B) The cross section reveals the view into the proteolytic chamber shown in VDW representation. C+D) Top view of an  $\alpha$ -subunit ring and a  $\beta$ -subunit ring. E+F) Closeup of single subunits:  $\alpha$ 1 and the catalytically active  $\beta$ 5. The subunits in C-F) are shown in new cartoon representation. N: N-terminus, C: C-terminus, NTE: N-terminal extension, Thr1: catalytically active residue threonine 1. Structures were rendered in VMD.

## Active Sites

The proteasome is an endoprotease being able to break peptide bonds within the substrate's polypeptide chain. During natural proteolysis, the catalytically active, nucleophilic oxygen atom Thr1O<sup>γ</sup> reacts with the electrophilic carbonyl carbon atom of the substrate's scissile bond forming an acyl ester intermediate that is stabilized by a hydrogen bond network of adjacent conserved residues (Glu17 and Lys33) [123]. A nucleophilic water molecule is suggested to transfer a proton from the hydroxyl group of Thr1O<sup>γ</sup> to the nucleophilic Thr1N also involved in the cleavage of the acyl ester intermediate for the regeneration of Thr1O<sup>γ</sup> [65, 67].

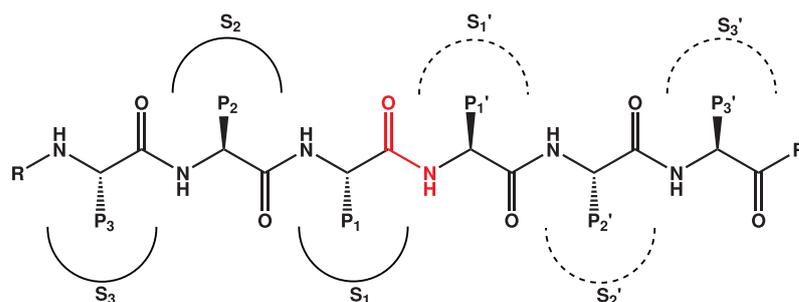
Prior to reaction, the substrate polypeptide chain docks into preformed channels near active sites, that are formed by two adjacent subunits ( $\beta$ 1 and  $\beta$ 2,  $\beta$ 2 and  $\beta$ 3,  $\beta$ 5 and  $\beta$ 6) [123]. The channel harbors substrate binding pockets termed non-primed (S1, S2, S3) and primed (S1', S2', S3') with respect to the cleavage site (Figure 2.2). Those pockets accommodate the different side chains of the substrate that are denoted accordingly with (P1, P2, P3) and (P1', P2', P3').

The proteasome produces polypeptide fragments with an average length of 8-12 aa due to 5 distinct cleavage preferences commonly found in proteases: caspase-like (CL), trypsin-like (TL), chymotrypsin-like (ChTL), branched chain amino acid preferring (BrAAP) and small neutral amino acid preferring (SNAAP) activity. The affinity of a substrate to a specific site is mainly determined by the character of the S1 specificity pocket that accommodates the P1 residue. This specificity depends on the chemical properties of residue 45 shaping the S1 pocket and is different in each of the active sites:

- $\beta$ 1 The positively charged Arg45 prefers to accommodate acidic residues in the S1 pocket, e.g., Glu as P1 residue, and its activity is therefore attributed CL. Moreover, this subunit was shown to cleave after hydrophobic amino acids and thereby also contributes to the BrAAP activity [27, 147].
- $\beta$ 2 The small Gly45 causes a large S1 pocket, constrained by Glu53 at the bottom. It favors a TL activity by accommodating large residues of basic character [66].
- $\beta$ 5 The S1 pocket preferentially accommodates large, hydrophobic residues due to the hydrophobic character of the Met45 side chain. Thus, a ChTL activity is conferred to the  $\beta$ 5 active site. Nevertheless, there is evidence that  $\beta$ 5 also exhibits SNAAP as well as BrAAP activity [66].

### 2.1.3 Proteasome Inhibitors

Dozens of potent proteasome inhibitors emerged over the last years that selectively target the 20S proteasome's  $\beta$ 5 active site. By mimicking the natural polypeptide substrate, many proteasome inhibitors are based on a peptidic scaffold and thus perfectly match the binding pocket. Nevertheless, peptide inhibitors show a decreased bioavailability due to the degradation of endogenous proteases [13].



**Figure 2.2:** Scheme of a polypeptide chain inside the substrate binding channel in the 20S proteasome's active sites: the channel is divided into the primed and non-primed region with respect to the scissile bond (red). Substrate binding pockets are termed S1-S3 and S1'-S3' respectively, and accommodate the polypeptide side chain's P1-P3 and P1'-P3'.

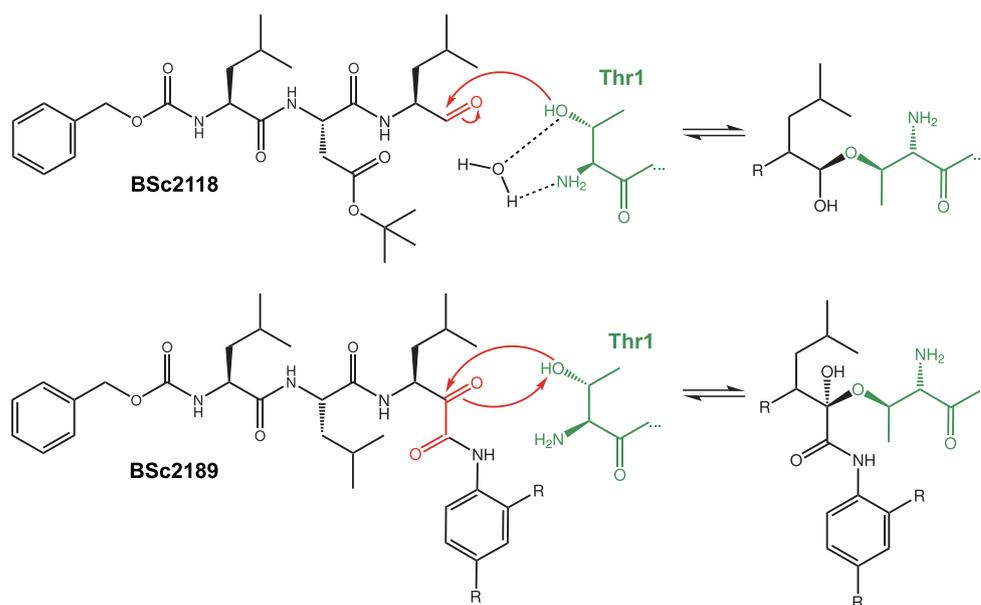
The inhibitors differ in their binding mode: non-covalent inhibitors lack a reactive functional group and bind reversibly to the active site through a network of interactions (hydrophobic, electrostatic and hydrogen bonds, van der Waals forces). The formation of a covalent bond between the ligand and a reactive side chain in the protein pocket increases the binding affinity. In contrast, the binding mode of covalent inhibitors is slowly reversible or irreversible. The disadvantage of a high-affinity binder with a slow dissociation rate is evident as it cannot be released from the target.

Covalent proteasome inhibitors differ in their functional groups, the so-called warheads, that attack the nucleophilic Thr1. The warhead determines the underlying binding mode and thus six major structural classes of inhibitors are defined: aldehydes, boronates, epoxyketones,  $\alpha$ -keto aldehydes,  $\beta$ -lactones and vinyl sulfones. The binding modes of all those inhibitor classes to the yeast 20S proteasome have been elucidated by X-ray crystallography (Table 2.1).

In this work, the focus lies on two covalently binding inhibitors, BSc2118 featuring an aldehyde warhead and BSc2189 with an  $\alpha$ -keto phenylamide functional group similar to that of the  $\alpha$ -keto aldehydes. Their chemical structures and the respective binding modes are shown in Figure 2.3.

**Table 2.1:** Major classes of 20S proteasome inhibitors listed together with a representative and its active site preference (subunit). Their binding modes could be elucidated by the co-crystal structures of the yeast 20S proteasome. Crystal structure details (PDB code, resolution and references) are indicated.

class	inhibitor	subunit	PDB code	resolution [Å]	reference
aldehyde	BSc2118	$\beta$ 1,2,5	–	2.8	[62]
boronate	bortezomib	$\beta$ 5	2F16	2.8	[64]
$\alpha,\beta$ -epoxyketone	epoxomicin	$\beta$ 2,5	1G65	2.3	[68]
$\alpha$ -keto aldehyde	glyoxal	$\beta$ 1,2,5	3OKJ	2.7	[61]
$\alpha$ -keto phenylamide	BSc2189	$\beta$ 5	4NO8	2.9	[177]
$\beta$ -lactone	homobelactosin c	$\beta$ 5	3E47	3.0	[69]
vinyl sulfone	LU-102	$\beta$ 2	4INR	2.7	[55]



**Figure 2.3:** Covalent 20S proteasome inhibitors used in this study: the aldehyde BSc2118 and the  $\alpha$ -keto phenylamide BSc2189 [21]. The electrophilic warhead is shown in red. The underlying chemical reaction upon binding to the Thr1 side chain (green) is shown.

## 2.2 Project I: Molecular Docking Studies

### 2.2.1 Introduction

With detailed structural knowledge of the binding site and binding mode of several proteasome inhibitors it is possible to support the identification and optimization of proteasome inhibitors by computational methods. The underlying idea is to accelerate drug design and discovery processes by supporting expensive and time-consuming wet-lab experiments. Protein-ligand docking is a classical *in silico* approach applied in chemoinformatics to identify the correct conformation (pose) and orientation of a small organic compound (ligand) in a binding site of a target protein and to quantify their interaction strength.

For the covalently bound proteasome inhibitors the challenge is to find a proper way to model the covalent bond. Among the great variety of docking software we chose the commonly used AutoDock [137] because it has been applied successfully in several studies [88, 157, 134]. Furthermore, as an open source project the program code is extensible. The development of an optimal docking parameter set for docking of proteasome inhibitors presented a major task. To this end, redocking studies of general protein-ligand complexes were conducted first. After that, redocking studies of the covalently-bound proteasome inhibitor BSc2118 [21] to the 20S proteasome's  $\beta 5$  subunit was performed. Different search algorithms (LGA, SA and STUN) were compared. A general study on ligand flexibility was performed.

Finally, with the comprehensive knowledge acquired, it was possible to develop an optimal procedure for covalent docking using the software MOE [133].

### 2.2.2 Theory: Molecular Docking

Molecular docking is a computational technique applied in structure-based drug design. The general goal of docking is both the prediction of the most favorable protein-ligand complex geometry and the prediction of the free energy of binding. Thus, molecular docking requires two components:

**Conformational Sampling** First, possible ligand conformations (docking poses) are generated. The conformational space of a ligand is very large since the total possible conformations increase exponentially with the number of rotatable bonds. A systematical search of the conformational space to fully explore all degrees of freedom would be computationally impossible. The degrees of freedom involve translation, rotation and dihedral angles. Therefore, different heuristics exist for global optimization to efficiently search the potential energy surface (PES): 1) empirically-based algorithms, 2) stochastic Monte Carlo methods and 3) evolutionary-based optimization. In this study, three different search algorithms are applied:

- LGA - The Lamarckian genetic algorithm (LGA) is a variant of a genetic algorithm [86]. The evolutionary concept is adapted from nature: here, an

individual defines a conformation that is represented by a vector of dihedral angles as well as entries for translation and rotation that define a certain ligand conformation. Crossover and mutational events occur with a certain probability analogous to biological evolution. Individuals undergo iterative cycles of reproduction and selection. Those vectors with conformations that exhibit the lowest free energy of binding are selected. The Lamarckian aspect states that individuals adapt to their environment and pass this information to the offsprings. Here, conformations are first optimized and this information is passed to next generations [137].

- SA - simulated annealing (SA) is a Monte Carlo based optimization algorithm [108]. Starting at an initial annealing temperature  $T_0$ , the temperature  $T$  is iteratively decreased by a reduction factor  $rtrf$  during each cycle during SA. Random changes are made to the ligand's current location, orientation and conformation. The probability  $p$  of a new conformation to be accepted is determined by the Metropolis criterion [129] defined as

$$p = e^{-\Delta E/k_B T} \quad (2.1)$$

where  $\Delta E$  is energy difference between the current and the last step,  $k_B$  is the Boltzmann constant and  $T$  is the temperature. The energy of the protein-ligand complex is calculated after each step. If the energy decreases, the move is accepted with the Boltzmann-distributed probability depending on current temperature according to Equation 2.1. If the temperature is high, the acceptance is favored.

- STUN - Stochastic tunneling (STUN) is an alternative algorithm to SA [198]. It avoids the freezing problem of SA. This occurs if the optimization process is trapped in local minima with decreasing temperature. To this end, in STUN the PES is modified through a non-linear transformation

$$f_{STUN} = 1 - e^{-\gamma(f(x)-f_0)} \quad (2.2)$$

where the tunneling parameter  $\gamma$  determines the transformation strength,  $f(x)$  is the current location on the PES and  $f_0$  is iteratively adjusted to the lowest minimum found so far within the optimization process. The transformation eliminates irrelevant high-energy regions of the PES, while those of low-energy are still preserved. The process is allowed to tunnel through high-energy barriers and therefore local minima traps are avoided. In STUN, the temperature is constant and thus, the process only depends on the single parameter  $\gamma$ .

**Scoring Function** Out of the set of generated conformations, the poses are evaluated by the scoring function with respect to a certain protein cavity (local docking) and the most favorable being selected. Scoring functions estimate the free energy of binding between protein and ligand under consideration of the physicochemical parameters derived A) by molecular mechanics force fields B) empirically functions and C) knowledge-based ones.

In this study, a semiempirical free energy force field is used [91]. The estimated free energy of binding energy is formulated as the sum of individual energy terms

$$V = W_{\text{vdw}} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{\text{hb}} \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \quad (2.3)$$

$$+ W_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{\text{sol}} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}$$

where the weighting constants  $W$  were parameterized using a large number of protein-inhibitor complexes for which both structure and experimentally-determined affinity constants are known. The four terms refer to a typical 6/12 Lennard-Jones potential [102] for dispersion/repulsion interactions (vdw), directional H-bond term based on a 10/12 potential (hb), Coulomb electrostatic energy (elec) [32] and a desolvation potential (sol). This is based on the volume  $V$  of atoms that surround a given atom and shelter it from solvent, weighted by a solvation parameter  $S$  and exponential term with distance-weighting factor  $\sigma$ .

The docking procedure is performed via several iterative steps of conformational prediction and subsequent scoring. Usually the major drawback of docking predictability are limitations of the scoring functions [19]. Although the most important energetic contributions are taken into account, the free energy of binding is estimated simplified thermodynamics. Entropic and solvation contributions are often approximated or neglected. The correlation between the estimated free energy of binding and experimentally-measured binding affinities is usually low [19]. More sophisticated methods such as free energy perturbation and quantum mechanical scoring are needed to overcome this problem, but are – in almost all cases – prohibitively expensive.

## 2.2.3 Methods

### Root Mean Square Deviation

The root mean square deviation (RMSD) is a measure for structural distance and is defined as:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i^a - x_i^b)^2 + (y_i^a - y_i^b)^2 + (z_i^a - z_i^b)^2} \quad (2.4)$$

where the Cartesian coordinates  $x$ ,  $y$  and  $z$  of a certain atom  $i$  are compared between two different conformations  $a$  and  $b$  of a molecule. The overall displacement is averaged over all  $N$  atoms and usually measured in Å. Here, the RMSD is used to compute the deviation between the ligand's docking poses relative to the conformation that observed in the crystal structure.

### Correlation Coefficients

A correlation coefficient is a normalized measure of dependence. By assuming a linear dependence between two random variables  $X$  and  $Y$ , the Pearson product-moment correlation coefficient  $r_{xy}$  is a measure of the linear correlation between the two variables

$$r_{xy} = \frac{s_{xy}}{\sigma_x \cdot \sigma_y} \quad (2.5)$$

where  $s_{xy}$  is the covariance of the random variables  $X$  and  $Y$  and  $\sigma_x$  and  $\sigma_y$  the standard deviation of  $X$  and  $Y$ . An ideal positive correlation results in a value of  $r = 1$  while a negative linear relation results in  $r = -1$ . In case of  $r = 0$  the two variables do not share a linear dependency.

Spearman's rank correlation coefficient  $R_{xy}$  is an extension to the Pearson coefficient also accounting for non-linear relationships. The data is ranked to its value and the correlation is calculated of the ranked data as follows

$$R_{xy} = r_{\text{rank}(x)\text{rank}(y)} \quad (2.6)$$

### General Redocking of Protein-Ligand Complexes

Protein-ligand complex structures were taken from the core set of the PDBBind2007 database [193] which actually comprises 70 diverse high-resolution structures (1.38-2.35 Å) with corresponding experimental binding affinities extracted from publications. Proteins and ligands containing unusual atom types or residues other than the standard 20 amino acids were removed. Several structures showed problems in later SA docking runs resulting in a program crash and were therefore discarded. It has to be noted that none of the remaining 50 structures established a covalent bond between protein and ligand. Necessary input files were created for docking jobs with AutoDock 4.2. The ligand and protein coordinates were put in separated files, water atoms were deleted and hydrogen atoms were added. Gasteiger charges were assigned as electrostatic partial charges and AutoDock-specific atom types were assigned. Subsequently, charges of both non-polar hydrogens and lone pairs were merged and non-polar hydrogens were removed. All dihedral angles of the ligand were allowed to rotate whereas flexibility of the protein's side chains was neglected. The docking area was centered on the ligand and the box size was set to the PDB-based dimensions plus 1 Å to ensure sufficient space for docking runs.

For the box region, grid maps for each receptor and ligand atom types were created using AutoGrid 4.0.

Three search algorithms were used in this study: LGA, SA and STUN. Default functions for LGA and SA in AutoDock 4.2 were utilized, STUN was implemented by modifying the SA routine. The parameter variation leads to a total of 27 setups for LGA, 9 for SA and 15 for STUN (appendix Table A.2). Different key parameters of conformational search algorithms were varied: population size, mutational rate and crossover rate (LGA); initial annealing temperature  $T_0$ , temperature reduction factor  $rtrf$  (SA); temperature  $T$  and tunneling parameter  $y$  (STUN). For SA, the annealing schedule was chosen to be geometric. For every setup, 100 individual runs were performed starting from an initial random location and conformation.

The progression of the estimated free energies of binding  $E_t^i$  (AutoDock score) were monitored for every step  $t$  within each run and averaged over all runs. For every structure  $i$ , the mean relative standard error  $\epsilon_{rel}^i(t)$  was calculated with regard to the estimated free energy of binding  $E_0^i$  of the initial crystal complex:

$$\epsilon_{rel}^i(t) = \frac{E_t^i - E_0^i}{|E_0^i|} \quad (2.7)$$

Here, the assumption is that  $E$  reached in any docking results never falls below the crystal structure initial energy, therefore  $\epsilon_{rel}^i$  values never get negative. Since the crystal structure complex is supposed to reveal the optimal binding mode, the estimated free energy of binding  $E_0^i$  should be the minimum energy. The estimated energies of conformations found during docking runs should not be lower.  $E_0^i$  was calculated by starting an AutoDock run using the *epdb* command. According to the AutoDock manual, this command can be used to calculate the energy of a particular ligand conformation before performing the docking runs. The coordinates of the ligand's conformation observed in the crystal structure were automatically maintained and several random initial translational or rotational steps were switched off. The  $E_0^i$  were extracted from the output file in the section *Total Intermolecular + Intramolecular Energy* in kcal/mol. An alternative approach for  $E_0^i$  determination was to setup a docking run consisting only of 1 docking step. The parameter file was manually modified to disable any ligand movement.

Further,  $\epsilon_{rel}^i(t)$  was averaged over all  $I$  structures leading to an overall relative error  $\bar{\epsilon}_{rel}$  for each parameter setup:

$$\bar{\epsilon}_{rel}(t) = \frac{1}{I} \sum_{i=1}^I \epsilon_{rel}^i(t) \quad (2.8)$$

### Redocking of BSc2118

Protein and ligand coordinates were taken from the complex structure of the yeast 20S proteasome co-crystallized with the tripeptidic aldehyde BSc2118 at 2.8 Å by M. Groll [62]. Only those two subunits of the proteasome were chosen that shape the pocket around the  $\beta 5$  active site (subunits PRE2 and C5, chains K and L). Hydrogen

atoms were added using Babel 2.3.2 [142] and charges were added to the heavy atoms in AutoDock. Electric partial charges were added with the Gasteiger-Marsili PEOE charges. Depending on the setup, crystal water was kept or removed and calculation of solvation parameters was done. Receptor flexibility regarding the side chains of Thr1 and Met45 was varied in setups 5-7 (Table 2.2). For the ligand, the rotation around 17 dihedral angle was allowed.

The covalent bond between the ligand's C32 and the receptor's O $\gamma$  atom of Thr1 was treated with the grid-based approach. This implicates an additional Gaussian grid map centered on the receptor's atom forming the covalent bond. The docking area's dimension was 60 $\times$ 60 $\times$ 60 Å centered on the O $\gamma$  atom with a grid spacing of 0.375 Å. Parameters of conformational search algorithms were kept constant:

- LGA: population size (150), mutational rate (0.02) and crossover rate (0.8)
- SA: initial annealing temperature  $T_0$  (1000) and temperature reduction factor  $rtrf$  (0.95)
- STUN: temperature  $T$  (1000) and tunneling parameter  $y$  (0.95)

For SA, the annealing schedule was chosen to be geometric. A total of 50 temperature cycles were applied. Apart from that, standard parameters were used. Different setups refer to the general docking strategy (Table 2.2). For every setup, 100 individual runs were performed starting from an initial random location and conformation.

Final estimated free energies of binding (AutoDock score) for each run were extracted from the output file. Coordinates of the final poses were extracted to calculate the RMSD with respect to the crystal structure. All heavy atoms were taken into account, no fit was done prior RMSD calculation to also include the translation deviation. Structures were visualized and rendered in AutoDockTools 1.5.4 [162].

**Table 2.2:** Setups chosen for redocking of BSc2118. All seven setups were conducted for the three search algorithms: LGA, SA and STUN. Setups 1 and 2 can be considered as the most basic redocking approaches (non-covalent and covalent docking). The degrees of freedom increases with setup numbers.

	parameter setup						
	1	2	3	4	5	6	7
grid-based covalent ligand docking	-	✓	✓	✓	✓	✓	✓
crystal water kept	✓	✓	✓	-	-	-	-
flexible ligand	-	-	✓	✓	✓	✓	✓
flexible side chains	-	-	-	-	Thr1	Met45	Thr1, Met45

## Dihedral Angle Dynamics

Molecular dynamics simulation of the free ligand BSc2118 in water was performed using Gromacs 4.5.1 [155] to derive the dihedral angles' flexibilities. The ligand was parameterized using the PRODRG Server [166] producing a topology file for the GROMOS69-53a6 force field [146]. Hydrogens were added and the SPC216 water model [10] was used. The system contained 46 ligand and 23,199 solvent atoms. The system was energy-minimized using the steepest descent integrator. A convergence was reached when the maximum force acting upon an atom was smaller than 1 kJ/(mol nm) with a step size of 0.01 nm. The MD production run was performed over a time range of 50 ns with a time step of 2 fs. Frames were written every 5000<sup>th</sup> steps.

Dihedral angles were extracted using *mk\_angndx*, absolute angles as well as autocorrelation were measured using *g\_angle* routine of Gromacs 4.5.1. A total of 20 dihedral angles were considered (Figure 2.8) where angles 18-20 define quasi-planar peptide bonds and are used for control purposes. Mean and standard deviation of angle distributions were calculated, the first 10 ns were neglected due to the initial equilibration phase. According to Spoel et al. [174], the dihedral angle autocorrelation function is defined as

$$C(t) = \langle \cos[\theta(\tau) - \theta(\tau + t)] \rangle_{\tau} \quad (2.9)$$

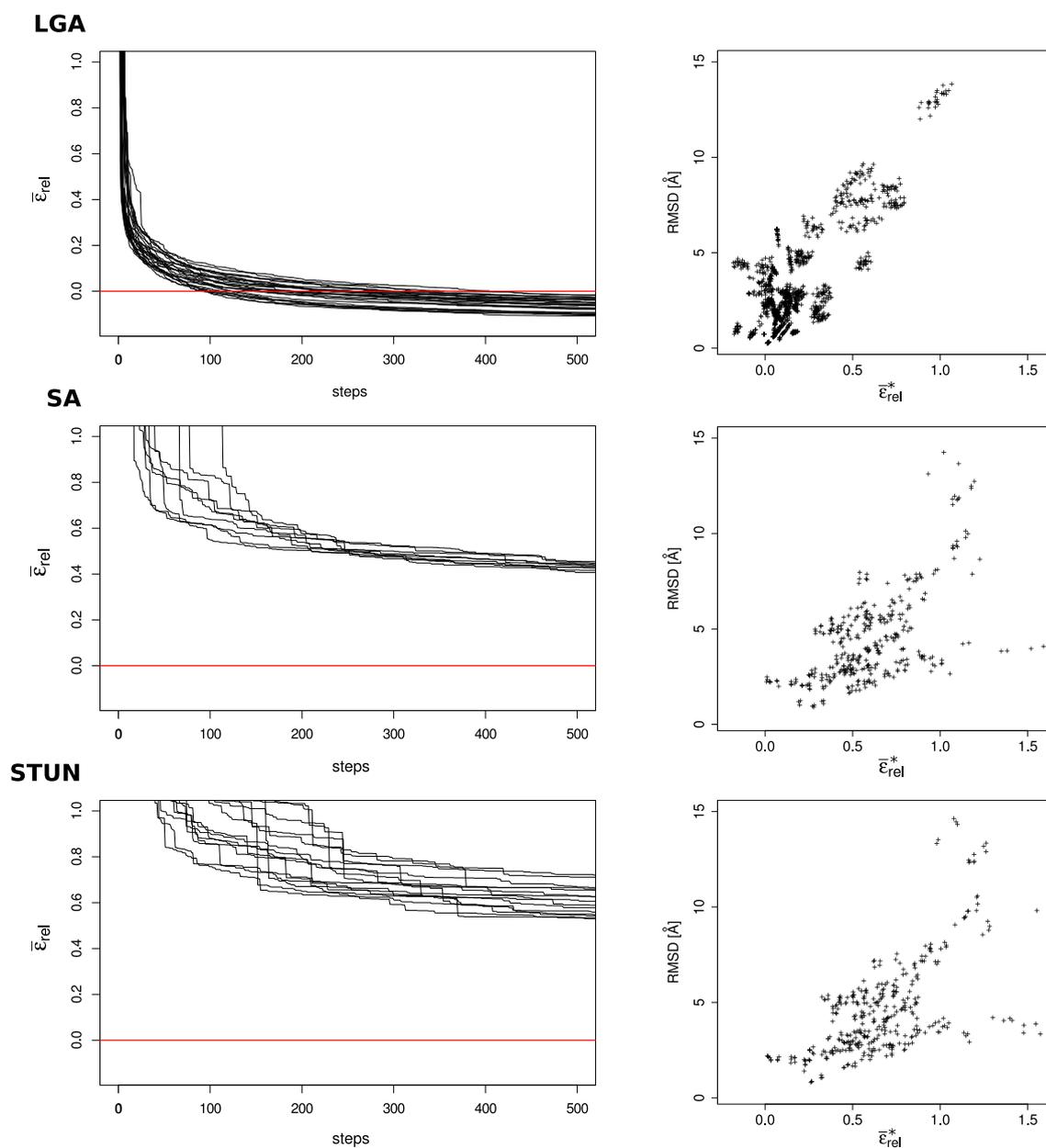
where the dihedral angle  $\theta$  is measured within every frame  $\tau$  and the cosine is taken with reference to the dihedral angle measure in frame  $\tau + t$ . The window size is defined by  $t$ . The use of cosines rather than absolute angles themselves resolves the problem of periodicity.

## 2.2.4 Results & Discussion

### Redocking Evaluation of PDBBind2007 Database

In order to assess the general docking capacity of AutoDock, we started an automated docking procedure of 50 protein-ligand complexes from the PDBBind2007 database [193]. This data set contains high-quality crystal structures of diverse enzymes with inhibitors not covalently bound (appendix Table A.1). The aim is the comparison of the performance of three search algorithms LGA, SA and STUN. For this purpose, different setups were chosen for each algorithm by variation of their core parameters (appendix Table A.2). The progression of the estimated free energy of binding (AutoDock score) is monitored over the iterative steps within a docking run.

For each setup, Figure 2.4 shows the progression of the mean relative error  $\bar{\epsilon}_{rel}$  of Equation 2.8 of predicted free energy compared to the initial energy  $E_0^i$  of the crystal structure. The mean relative error is averaged over 100 independent docking runs and averaged again over all 50 structures. The overall convergence behavior is visible during the first 500 steps. LGA shows a homogeneous picture with all setups



**Figure 2.4:** Convergence behavior (left) of the three search algorithms LGA (top), SA (middle) and STUN (bottom) using different parameter settings: 27 for LGA, 9 for SA and 15 for STUN. The estimated free energy of binding is extracted after every step and related to  $E_0^i$  yielding the relative error  $\epsilon_{rel}^i$ . For each setup, the mean relative error  $\bar{\epsilon}_{rel}$  is averaged over 100 runs and 50 complex structures. The correlation (right) between the RMSD and  $\bar{\epsilon}_{rel}^*$  is shown for all setups and all complex structures averaged over 100 runs. Note that the estimated free energy  $\bar{\epsilon}_{rel}^*$  is averaged over 100 runs only.

reaching negative values between -0.1 and 0. The pattern for SA and STUN is a rather step-like pattern. After 500 steps, SA setups converge against a relative error of 0.5, whereas STUN setups are spread between 0.5 and 0.8. The closeups show the convergence pattern at 1000 steps. For LGA, 1000 steps seem to be sufficient whereas SA and especially STUN relative mean error still continue to decrease even after 1000 steps.

Two alternative approaches were conducted to determine the initial energies  $E_0^i$  as it is described in Section 2.2.3. The  $E_0^i$  values achieved by the second approach are consistently lower. Still, in 29 out of 51 crystal structures the calculated  $E_0^i$  values are higher than those found during docking runs and therefore  $\bar{\epsilon}_{rel}$  for LGA setups reach negative values as shown in Figure 2.4. This excludes that this observation is dependent of a certain protein-ligand complex. Most probably, the AutoDock scoring function might be inadequate to accurately score the poses.

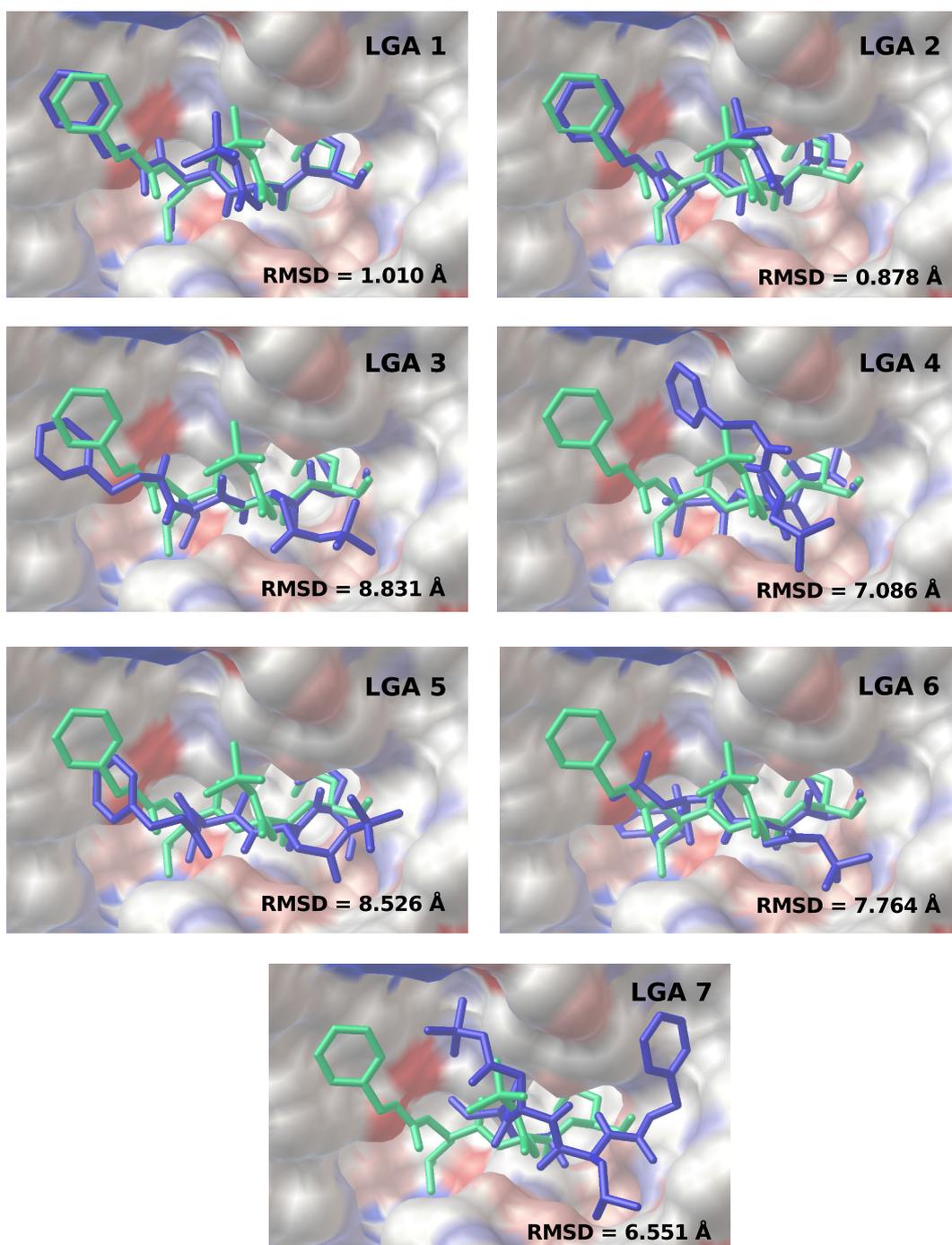
Although the initial energies  $E_0^i$  are actually lower, the convergence patterns at step 1000 suggest that the variation of core parameters do not seem to have a great influence on the prediction quality. Nevertheless, the LGA is the recommended conformational search algorithm and STUN does not perform better than SA.

### Redocking Evaluation of BSc2118

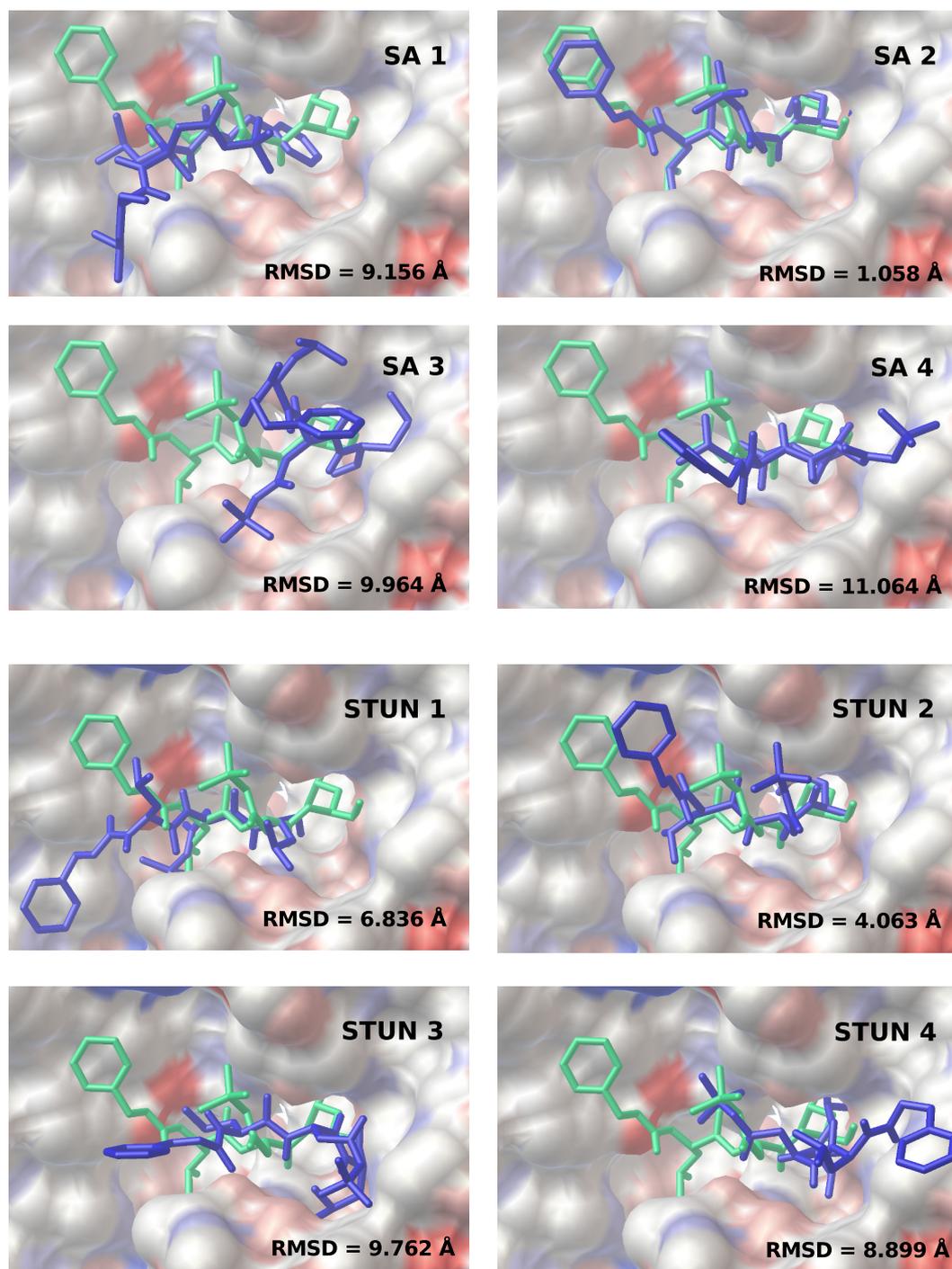
In order to dock inhibitors of the 20S proteasome's  $\beta 5$  active site in AutoDock, we chose to focus on the aldehyde lead structure BSc2118 being our most active compound at that time. The challenge here lies in the covalent bond between the Thr1O $\gamma$  and the hemiacetal moiety of the ligand. In AutoDock, this problem is tackled by using a grid-based approach, i.e. the covalent affinity of the ligand to the protein is modeled by a grid map. Seven different setups were defined, with the main focus lying on the structural setup and the degree of flexibility in general (Table 2.2). Additionally, the setups were conducted for the three conformational search algorithms LGA, SA and STUN using standard parameters as it was mentioned in the section before.

The orientation of the best poses using LGA are visualized together with the reference crystal structure in Figure 2.5. The basic setups 1 and 2 of non-covalent and covalent docking with the rigid ligand can resemble the crystal structure to a high extend (RMSD values of 1.010 Å and 0.878 Å). Setting the ligand's dihedral angles flexible, the crystal structure is well reproduced (setup 2). Obviously, it does not make a difference if water molecules from the crystal structure are kept (setup 3) or not (setup 4) since both approaches exhibit high deviations. Docking runs with only one flexible side chain Thr1 (setup 5) or Met45 (setup 6) show also high RMSD values. The docking attempt with the two side chains Thr1 and Met45 being both flexible failed due to a program crash. Applying the quality criterion for poses of RMSD < 2 Å, setups 3 to 7 also failed for this reason.

The orientation of the best poses using SA are visualized in Figure 2.6 (top). The predicted pose for the non-covalent docking (setup 1) shows the inverse orientation of the ligand highlighting the importance of the covalent attachment of the ligand to



**Figure 2.5:** Orientation of the top predicted pose in BSc2118 redocking with LGA setups 1-7. The predicted best pose (blue) is shown together with the reference crystal structure (green) embedded in the  $\beta 5$  pocket of the 20S proteasome colored acidic (red) and basic (blue) [62]. The predicted best pose is selected due to the lowest estimated free energy of binding out of the structures from 100 independent docking runs. Structures were rendered using AutoDockTools.



**Figure 2.6:** Orientation of the top predicted pose in BSc2118 redocking for SA (top) and STUN (bottom) setups 1-4. The pose (blue) is shown together with the reference crystal structure (green) embedded in the 20S proteasome's  $\beta 5$  pocket that is colored acidic (red) and basic (blue) [62]. The predicted best pose is selected due to the lowest estimated free energy of binding out of the structures from 100 independent docking runs. Structures were rendered using AutoDockTools.

the protein. Indeed, the covalent docking shows a high accordance to the reference structure with an RMSD of 1.058. Docking runs using the flexible ligand fail to reproduce the crystal structure for both setups with and without water molecules from the crystal structure (setups 3 and 4, RMSD > 9.9). Docking attempts with any flexible protein side chain(s) (setups 5-7) could not be evaluated due to a program crash. A similar picture is drawn by the poses produced with the STUN algorithm (Figure 2.6 bottom), but the covalent redocking indicates a rather high RMSD of 4.063. Still, the RMSD values of setups 1, 3 and 4 are even lower than observed with SA.

Taking a look at the entirety of all predicted structures of the 100 independent docking runs, the relation between the estimated free energy of binding and their RMSD to the reference crystal structure was investigated. Using Spearman's ranking correlation coefficient, no significant correlation could be detected in any of the setups independent from conformational search algorithms (Table 2.3). That implicates that the estimated free energy of binding does not have a predictive power regarding the prediction of poses being similar to the reference crystal structure and is not reliable.

**Table 2.3:** Spearman's ranking correlation coefficient between AutoDock score (estimated free energy of binding [kcal/mol]) and RMSD relative to the starting crystal structure. The 20S proteasome inhibitor BSc2118 was redocked into the  $\beta 5$  pocket over 100 individual docking runs for each algorithm and setup. LGA 1-7, SA 1-4, STUN 1-4. The setups 5-7 of SA and STUN could not be evaluated due to program crash most probably to a freezing problem. Details for parameter setups are listed in Table 2.2.

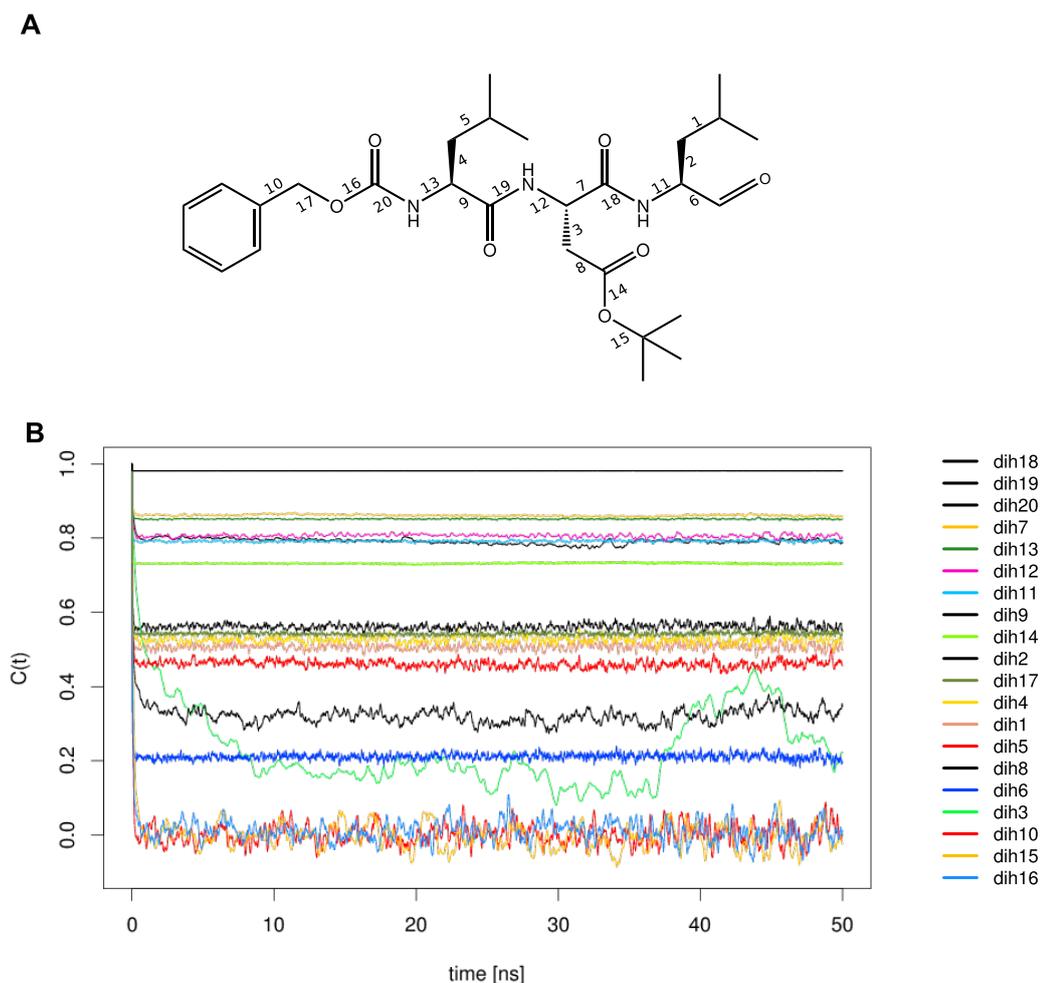
algorithm	parameter setup						
	1	2	3	4	5	6	7
LGA	-0.636	0.546	0.380	0.311	0.048	0.414	0.145
SA	0.187	0.268	0.194	0.295	–	–	–
STUN	0.261	0.305	0.009	0.033	–	–	–

## Dihedral Angle Dynamics

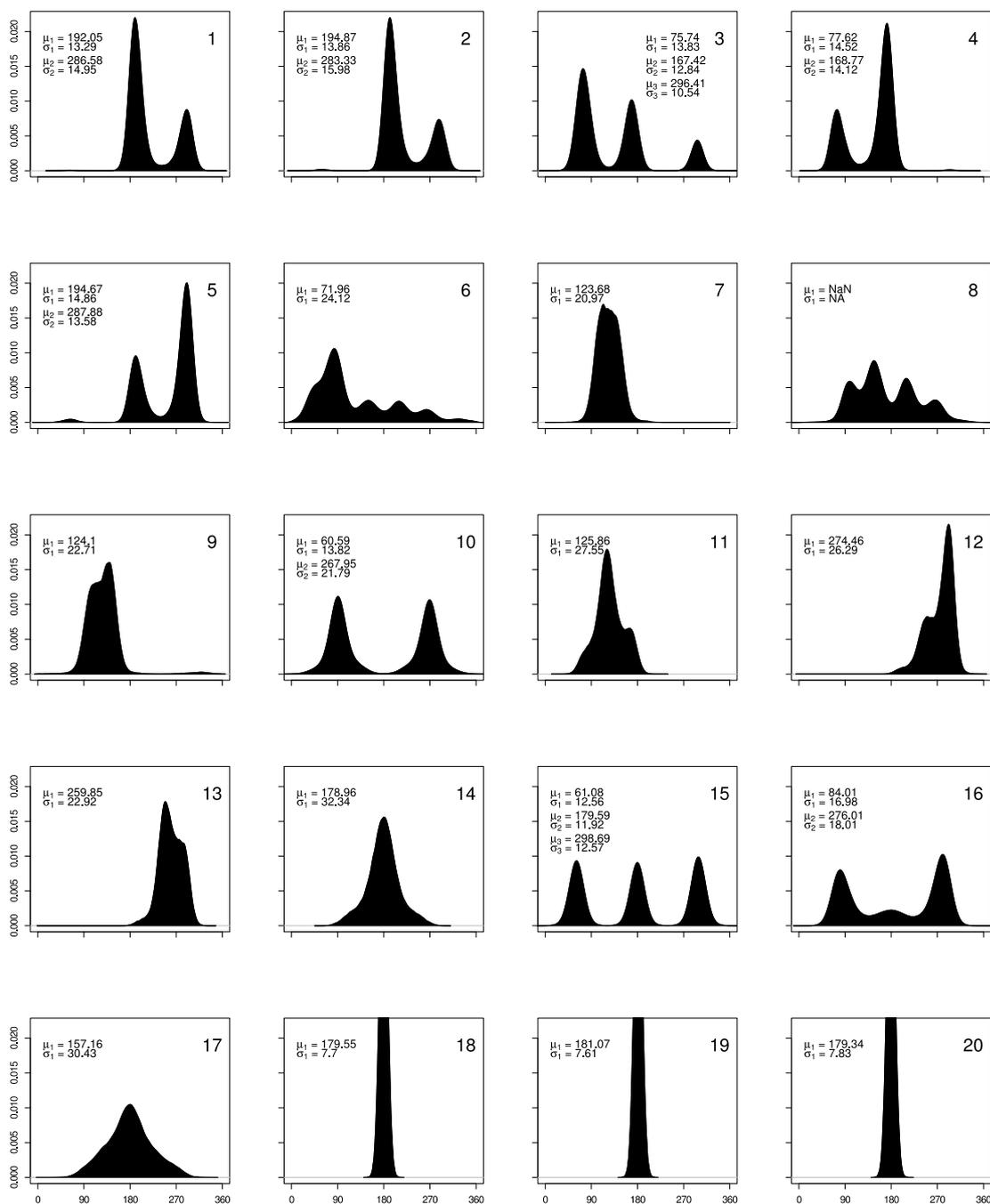
Ideal protein-ligand docking procedures would take the entire ligand's and protein's flexibility into account. This leads to a dramatical increase in complexity of the conformational search space accompanied by high computational cost. As it was shown in the section before, the flexibility of just a few side chains led to inaccurate predicted docking poses. Due to this fact, we decided to investigate the BSc2118

dihedral angle flexibility with the aim of possibly reducing the ligand's degrees of freedom.

In order to study torsion angle dynamics, we computed the autocorrelation of each dihedral angle to detect periodicities within a time series (Figure 2.7). It indicates the angle's frequency to switch to another state. In combination with the density plots of the absolute values (Figure 2.8) where the different states are visible, it gives rise to the flexibility of dihedral angles. Different classes of angles can be described by inspecting the autocorrelation plot:



**Figure 2.7:** A) Chemical structure of the 20S proteasome inhibitor BSc2118 with dihedral angles defining the rotation around certain bonds (enumerated from 1 to 20). B) Correlation of the 20 dihedral angles of the 20S proteasome inhibitor BSc2118 observed over time in a 50 ns molecular dynamics trajectory. The correlation of each dihedral angle was calculated according to Equation 2.9.



**Figure 2.8:** Density plots of 20 dihedral angles of 20S proteasome inhibitor BSc2118 observed in a 20 ns molecular dynamics trajectory. Dihedral angles 18-20 are located at peptide bonds and serve as a control, for compatibility purposes the y-axis limit is set to 0.02.

- The dihedral angles 18, 19 and 20 defining the rotation around peptide bonds are considered to be non-flexible and were included for control purposes. Indeed, they show high autocorrelation values with a very low standard deviation ( $\mu = 0.981$ ,  $\sigma < 10^{-4}$ ). Their angle distributions confirm a very sharp peak at  $180^\circ$ .
- Dihedral angles 7, 9, 11, 12, 13 and 14 are located in the second region with average autocorrelation values between 0.731 and 0.861 and low standard deviations ranging from 0.002 to 0.009. Corresponding density plots reveal one peak only. All angles are located in the ligand's peptide backbone, except dihedral 14 in the butyl aspartate side chain of the P2 residue.
- Dihedral angles 1, 2, 4, 5 and 17 (average autocorrelation values between 0.461-0.561 and standard deviations 0.010-0.012) and dihedral 17 ( $\sigma = 0.006$ ) are within the same region. All density plots show two peaks, one of them being dominant. The angle 17 has only one peak with a wide base. Angles 1 and 2 are equivalent to 5 and 4 being located at the base of the two leucyl side chains. Dihedral 17 controls the rotation of the phenyl ring.
- Dihedral angles 3, 6 and 8 show a low mean autocorrelation around 0.211-0.323 and relatively high standard deviations around 0.025-0.110. The density plots reveal several peaks without a clear dominance. Dihedral angles 3 and 8 define the rotation of the P2 residue and angle 6 of the carbonyl head group.
- Dihedral angles 10, 15 and 16 show mean autocorrelation values around zero while the standard deviation is between 0.034-0.046. Density plots show two or three well-defined states. Angles 10 and 16 are located near the phenyl group, dihedral 15 rotates the end of the P2 residue.

Hence, it can be concluded that the simulation based on the GROMOS69-53a6 force field was appropriate to sample the ligand's entire conformational space. The density plots show that energy barriers were surmounted forming different defined states for every angle. The analysis reveals that some dihedral angles are less rigid than others. Still, the flexibility is present in almost all angles and thus none of them can be set rigid during a docking run. Contrary to the flexible bonds the peptide bond's flexibility however is harshly restricted and these angles can be set to rigid.

### Development of a Covalent Docking Protocol in MOE

The project described in the following section was done in collaboration with C. Scholz (TU Darmstadt, Organic Chemistry and Biochemistry), see *List of Contributions*. C. Scholz and S. Knorr implemented the method in MOE by scientific vector language (SVL) scripting. A poster summarizing the principal method was created and presented by C. Scholz at the MOE User Group Meeting & Conference in 2014. A manuscript describing the following method in detail and its applications is in preparation.

It has been shown that the docking of those ligands being covalently bound to the protein is not straightforward (Figures 2.5 and 2.6) and not adequately implemented in most of the docking programs available today. To this end, we developed a docking protocol for the Molecular Operating Environment (MOE) [133], a platform for structural modeling. The procedure works for any kind of covalently bound protein-ligand complexes, but will be explained here for covalent docking of 20S proteasome inhibitors to the  $\beta 5$  pocket:

1. *Warhead Screening*: In the first step, the location of the electrophilic warhead within the ligand is identified by searching a database of predefined warheads. The electrophilic atom is tagged and transformed to the configuration according to the bound state. In case of BSc2118, the electrophilic carbon atom is converted from a  $sp^2$ -hybridized ketone to the bound hemiacetal with  $sp^3$ -geometry.
2. *Side Chain Attachment*: The covalent bond is established between the tagged carbon atom and the protein's Thr10 $\gamma$ . Being separated from the protein, the side chain is now considered as a part of the ligand leading to the formation of a chimeric ligand.
3. *Pharmacophore Model*: This model of the side chain's atom types and locations is constructed. By constraining the atom movements to a maximum deviation of 0.2 Å and additionally setting up a high pharmacophore force constraint, the side chain is ensured to maintain its original place.
4. *Docking and Rescoring*: Several pharmacophore-guided docking cycles are performed, the ligand flexible and the pocket side chains fixed. The poses are scored by standard non-covalent scoring functions of MOE (Affinity  $\Delta G$  or London  $\Delta G$ ). Subsequently, the poses are further refined by energy minimization with the ligand kept flexible. The side chain and the ligand are disconnected and the complex is rescored by external scoring functions such as DSX [140] to obtain the final estimated free energy of binding.

Although the covalent bond's energy is still not treated explicitly in the scoring function, it can be neglected when comparing the docking poses with each other. The protocol was implemented in the SVL scripting language of MOE. The process ensures even high-throughput docking of covalently bound ligands so that a large library of ligand candidates can be assessed in an automated way. Once the ligand library is transformed to the intermediate state and the pharmacophore is prepared, the docking runs can be performed automatically.

### 2.2.5 Conclusion

Protein-ligand docking is a widely used computational method in drug design. Beside the requirement of structural knowledge of the protein-ligand complex,

the very first step to assess reliable docking results is to perform a redocking study. Ideally, a ligand's conformation is predicted being most similar to the one observed in the crystal structure. Further, a correlation should exist between estimated free energies of binding (the docking score) and experimentally measured binding affinities. In order to obtain successful redocking results, two conditions are required:

1. The crystal structure's conformation must be within the set of generated conformations suggested by the conformational search algorithm.
2. The interaction between these conformations and the protein must be evaluated correctly by the scoring function.

In order to check the performance of the popular docking program AutoDock, we first conducted a redocking study of arbitrary protein-ligand complexes. The redocking in this context can be considered as a relatively simple task, because the protein is treated as a rigid body and only the ligand is flexible. Thus, finding the crystal structure conformation is somehow trivial, because the ligand is put in an already preshaped protein pocket. Different search algorithms (LGA, SA and STUN) were compared to accomplish this task. LGA showed the best performance. Moreover, results demonstrate that the variation of core parameters does not effect the docking result significantly.

In case of docking covalent proteasome inhibitors, modeling of the covalent bonds is challenging. In none of the docking programs the covalent bond is included properly in the scoring functions. AutoDock uses an indirect variant for covalent docking, a grid-based method. The redocking analysis of the proteasome inhibitor BSc2118 shows severe problems in predicting the crystal structure with an increasing amount of degrees of freedom. The ligand alone has 17 torsional degrees of freedom and with flexible protein side chains, a complex conformational space arises and an effective search to find deep local minima (or even the global minimum) is difficult. Eventually, the crystal structure's conformation is not returned by any of the tested search algorithms (LGA, SA and STUN). Further, it turned out that the estimated free energies of binding and similarity to the crystal structure are highly uncorrelated. This implicates conformations not to be scored adequately.

In order to tackle the dimensionality the attempt was to focus on the ligand's dihedral angle dynamics and eventually restrain some of the dihedral angles. According to the AutoDock manual, current implemented search algorithms are only efficient up to a maximum of 10 rotatable bonds, the ligand's flexibility must be reduced, particularly for the docking of larger proteasome inhibitor candidates. The analysis revealed that none of the dihedral angles has a restricted flexibility that it can be neglected.

Because of the results shown, docking of large ligands is not straightforward and only possible by reducing the degrees of freedom. Still, a major drawback of protein-ligand docking in general is the rigid character of the protein. Proteins are

not fixed in nature and the crystal structure is only a snapshot of the protein's entire conformational ensemble [40]. Docking of arbitrary ligands into a native protein structure exhibited by the crystal structure conformation is questionable. Flexible side chains try to account for that problem but larger, globally conformational changes upon ligand binding cannot be investigated.

One possibility to reduce the system's complexity is partial docking: while most of the ligand is maintained rigid, only certain sites of interest are flexible. This is useful in comparing different substitutions. Thus, protein-ligand docking can achieve successful results in the context of ligand optimization rather than predicting a ligand pose *de novo*. Still, it is not possible to solely rely on docking results. Protein-ligand docking must be considered as just one branch in the overall framework of a drug-design project.

All in all, the results show that AutoDock is not suitable for docking 20S proteasome inhibitors under the setups tested in this work. A great variety of diverse docking programs exists with different conformational search algorithms and scoring functions, all performing differently depending on the protein-ligand complex. For every complex the optimal choice is different.

Finally, we established a protocol for covalent protein-ligand docking in MOE that accounts for the formation of covalent bond and full ligand flexibility within a reasonable time. This approach is to be pursued further and future large-scale evaluation studies are to be made. Preliminary results show that redocked covalently-bound inhibitors are in good agreement with the crystal structures.

## 2.3 Project II: Inhibitor Optimization

### 2.3.1 Contributions

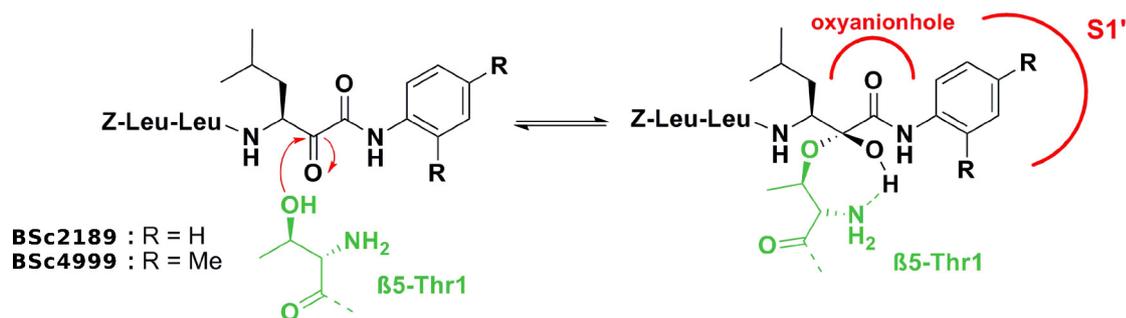
The project described in the following section was done in collaboration with several other groups and the contents were submitted to ChemMedChem in June 2014, see *List of Contributions*. Following contributions have been made by the authors: C. Voss, C. Scholz, A. Zall and B. Schmidt (TU Darmstadt, Organic Chemistry and Biochemistry) developed the project plan, chemical compound synthesis, covalent docking and SAR study. S. Knorr and K. Hamacher (TU Darmstadt, Computational Biology & Simulation) did the statistical and structural analyses. P. Beck, M. Stein and M. Groll (TU Darmstadt, Biochemistry) elucidated the cocrystal structure of BSc4999 bound to the yeast 20S proteasome. U. Kuckelkorn and P.M. Kloetzel (Charité Berlin, Biochemistry) performed the *in vitro* and *in vivo* fluorescence-based activity measurements and the reversibility assay. The overall project is summarized in the following section. C. Voss, C. Scholz and S. Knorr prepared the manuscript.

### 2.3.2 Introduction

$\beta$ -Lactone proteasome inhibitors such as belactosin A and homobelactosin C constitute a promising class for the development of covalently bound proteasome inhibitors [6, 111, 106]. Classical known inhibitors exclusively target the non-primed substrate channel S (Figure 2.2) whereas recent structural studies showed that homobelactosin C additionally exploit the primed region S': the peptidic backbone lies in the non-primed channel (S1-S3) whereas simultaneously the phenylamide moiety lies in the primed site S1' [69]. Targeting the S' region could constitute an important selectivity criterion because, e.g., the S' sites differ between the consecutive proteasome and immunoproteasome in size and polarity [17]. High selectivity is an important goal in drug design because a low specificity leads to undesirable side effects in patients.

$\alpha$ -Keto phenylamides were first described as proteasome inhibitors by Chatterjee et al. [29] and their predicted binding mode of a hemiacetal formation in the  $\alpha$ -position of the phenylamide moiety could be confirmed by the crystal structure of BSc2189 (Figure 2.9) [177]. Prior studies have shown the  $\alpha$ -keto phenylamide moiety to be a promising warhead [177]. The lack of a second strong electrophile (as in epoxyketones) is compensated by binding to the S' channel.

BSc2189 was identified as a highly active ( $IC_{50,\beta5}$ : 72 nM)  $\alpha$ -keto phenylamide 20S proteasome inhibitor preferentially targeting the  $\beta5$  active site [21]. The stepwise optimization procedure of this inhibitor is described in the following using structure-based and direct rational design. As a result, an enhanced inhibitor BSc4999 was identified ( $IC_{50,\beta5}$ : 38 nM) that is still highly selective and moreover able to penetrate cells. The crystal structure reveals the inhibitor's orientation inside the primed channel.



**Figure 2.9:** Binding mode of  $\alpha$ -keto phenylamide inhibitors to Thr1 of the 20S proteasome's  $\beta 5$  subunit (green). The notable feature is the occupancy of the S' substrate channel by the aromatic phenylamide ring. BSc4999 differs from BSc2189 by two methyl substitutions in *para*- and *ortho*-position.

### 2.3.3 Methods

The compound series was synthesized from peptidic aldehydes in Passerini reactions with aromatic isonitriles. The resulting alcohol-intermediates were then oxidized utilizing the reagent IBX to result in alpha-keto phenylamides [21].

The covalent docking was done in MOE 2012.10 [133] using the SVL script *conflexdock* [31]. The cocrystal of BSc2189 in complex with the yeast 20S proteasome was used [177], only the  $\beta 5$  and  $\beta 6$  subunits were taken. Water atoms were removed and hydrogen atoms added. The ligand conformational library was prepared containing BSc2189 derivatives with five possible methyl substitutions in *ortho*- and *para*-locations at the phenyl ring. For each of the derivatives, a set of 8 different conformations were constructed by altering the dihedral angle  $\theta$  that defines the rotation around the phenylring in 45°-steps. The covalent bond was established between the ligand's  $\alpha$ -carbonyl site and the O $^{\gamma}$  of Thr1 of the  $\beta 5$  subunit. The ligand's peptidic backbone was kept fix and only the dihedral angle  $\theta$  was allowed to rotate. After docking runs, the poses were refined by an energy minimization and ranked using the MOE energy score. Poses were then rescored and compared by the estimated binding free energies calculated with the London  $\Delta G$  scoring function.

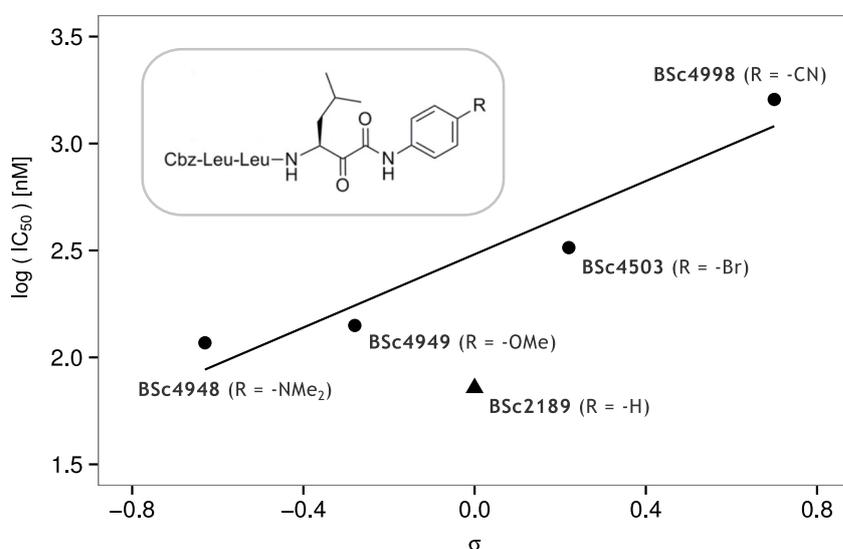
For structural elucidation of the BSc4999 complex, yeast constitutive 20S proteasome crystals were grown [65] and soaked with the inhibitors and measured with synchrotron radiation at a resolution of 2.5 Å.

The site-specific inhibitory activity was measured in a competitive assay using different inhibitor concentrations: 1) *in vitro* using 20S proteasome isolated from human red blood cells and 2) *in vivo* assays in HeLa cells. If the  $\beta 5$ -specific substrate Suc-LLVY-AMC is cleaved, a fluorogenic group is released and detected at 460 nm. Highly active inhibitors block the active site and the fluorescence signal decreases. Reversibility of proteasome inhibition was done *in vitro* using dialysis experiments

with a stepwise reduction of inhibitor concentration. The fluorescence signal was measured at different time steps to detect recovery of proteasome activity.

### 2.3.4 Results & Discussion

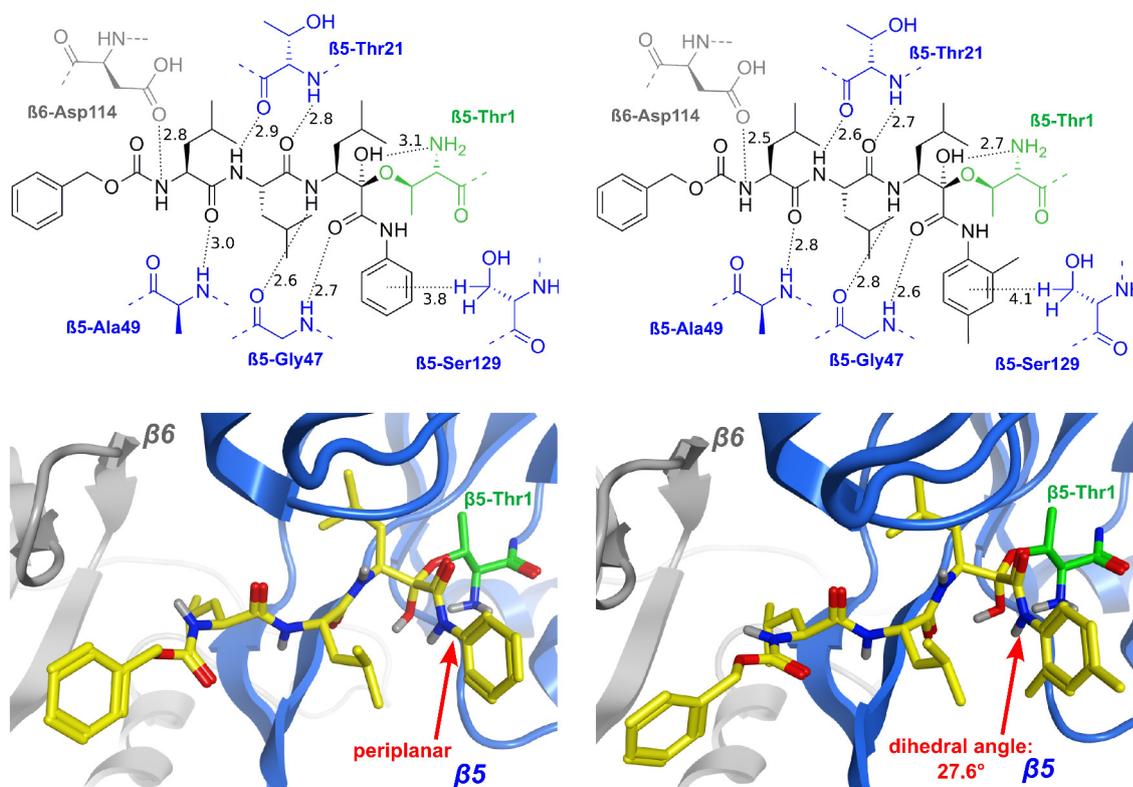
Based on the binding mode of  $\alpha$ -keto phenylamide inhibitors revealed by the crystal structure analysis of BSc2189, we chose to investigate the electronic situation of the aromatic system. As visualized in Figure 2.11 (bottom left), the phenylamide moiety of BSc2189 lies almost planar in the binding channel. A compound series was synthesized with different electron-donating and -withdrawing groups in *para*-position of the phenyl ring to alter the electron density.



**Figure 2.10:** Relation of logarithmic  $IC_{50}$  values of inhibitors against the Hammett constants  $\sigma$  [76]. The closeup shows the peptidic scaffold, the inhibitors differ in substituents at the *para*-position. A linear model  $\log_{10}(IC_{50}) = 0.856 \text{ nM} \cdot \sigma + 2.482 \text{ nM}$  was fitted to the data of the *para*-substituted derivatives showing a significant correlation ( $p \leq 0.041$ ). For illustration purposes, the data point of the lead structure BSc2189 was added to the plot but was excluded from the correlation analysis. Note, that the Hammett constant  $\sigma$  is dimensionless.

Based on *in vitro* activity measurements, a structure-activity relationship (SAR) study was performed investigating the correlation between logarithmic  $IC_{50}$  values and Hammett constants  $\sigma$  [76, 169, 202] for the *para*-substituted moieties. Here, the linear correlation indicates that the binding process is mainly effected by *para*-induced electronic effects: high  $\sigma$  correspond to higher  $IC_{50}$  values i.e. lower

inhibitory activities (Figure 2.10). Still, none of the derivatives reach the activity of the original unsubstituted BSc2189. The question arises why BSc4948 and BSc4949 do not reveal lower IC<sub>50</sub> values. Noticeable, electron-donating groups (-NMe<sub>2</sub>, -OMe) lead to higher activities than withdrawing groups (-Br, -CN). We hypothesize, that the electron-donating groups participate in hydrogen bonding with water molecules and that this effect causes a lower activity of BSc4948 and BSc4949 [115] than possible due to the mere electron-donating effects without hydrogen bonding. To diminish this effect, we suggest to introduce electron-donating groups without the ability to form hydrogen bonds. Therefore, we further considered methyl groups as substituents.



**Figure 2.11:** Crystal structures of BSc2189 (left) [177] and BSc4999 (right) in complex with the  $\beta 5$  active site of the yeast 20S proteasome. 2D chemical structure (upper panels) with their respective binding mode with interactions of amino acids belonging to subunit  $\beta 5$  (blue) and  $\beta 6$  (gray). The catalytically active Thr1 of  $\beta 5$  subunit amino acid is green. Distances in Å are indicated with dotted lines. 3D stick model of ligand and new cartoon representation of protein (lower panels). The dihedral angle turning around the bond connecting the peptidic backbone to the phenylamide moiety is marked with red arrows. Picture was reprinted with permission [190].

An *in silico* docking approach should give rise to the optimal arrangement of methyl substitutions at the phenyl ring. To this end, all possible configurations of methyl groups arranged in *ortho*- and *para*-position were realized in a set of newly-constructed ligand derivatives. For each derivative, different conformations were prepared with varying dihedral angles. The results of covalent docking to the  $\beta 5$  active site state BSc4999 to be the most active inhibitor of this series suggesting that the aromatic system should be dimethylated.

BSc4999 was synthesized and *in vitro* assays confirmed its proposed high activity ( $IC_{50,\beta 5}$ : 34 nM) and moreover exhibits a clear preference for the  $\beta 5$  active site. The elucidated cocrystal structure of BSc4999 revealed the phenyl moiety not lying perfectly coplanar in the S' channel by showing a rotation around the dihedral of  $27.6^\circ$  (Figure 2.11). *In vivo* studies confirmed the results from *in vitro* measured activities and showed that BSc4999 is cell-permeable. These promising results were further enhanced by the dialysis experiments that revealed a slowly reversible inhibition with the proteasomal activity completely restored after 72 hours.

### 2.3.5 Conclusion

This work represents a straightforward inhibitor optimization approach leading throughout different experimental steps. It was conducted in a large-scale collaboration combining the knowledge of medicinal chemistry, modeling, crystallography and biochemistry.

The elucidation of the  $\alpha$ -keto phenylamide binding mode confirms the covalent bond between the catalytically active Thr1 and the  $\alpha$ -site of the  $\alpha$ -keto hemiacetal formation. As a remarkable feature this inhibitor class targets the primed and non-primed site simultaneously with the  $\alpha$ -keto moiety lying in the primed channel. Based on the observed coplanarity of BSc2189's  $\alpha$ -keto moiety inside the channel, a systematic comparison of *para*-substituted derivatives was conducted. With BSc2189 still being the most active inhibitor of the initial compound series, it was postulated that hydrogen bond formation of electron-donating substituents diminishes the activity. Finally, this led to the identification of the highly-active BSc4999 inhibitor that retains ligand efficiency and exploits the S' channel, thereby improving selectivity to the 20S proteasome. This study presents new opportunities in high-selective proteasome inhibitor design.

## 2.4 Project III: Proteasome Substrate Cleavage Mechanism

### 2.4.1 Contributions

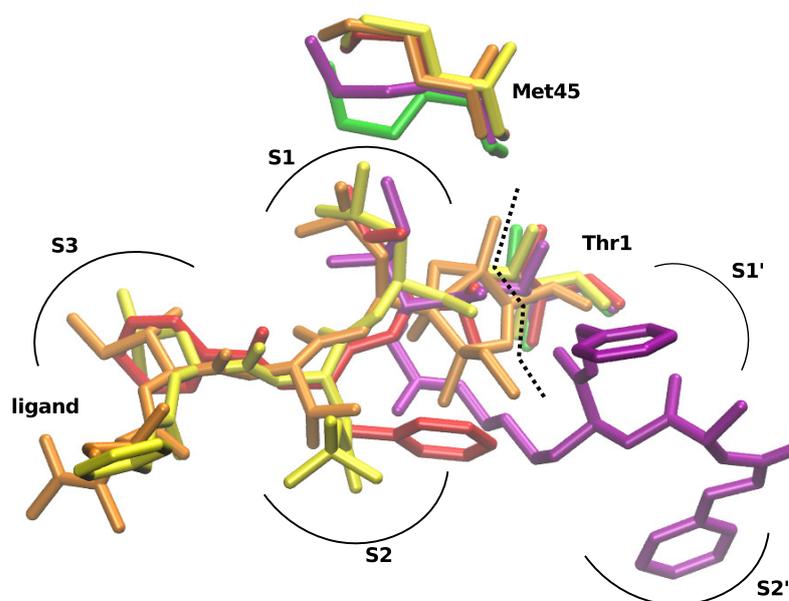
The project described in the following section was done in collaboration with other groups, the manuscript is prepared and will be submitted for publication, see *List of Contributions*. Following contributions have been made by the authors: C. Voss, C. Scholz and B. Schmidt (TU Darmstadt, Organic Chemistry and Biochemistry) proposed the mechanism, developed the project plan, synthesized the chemical compounds and performed activity measurements by competitive *in vitro* assays. S. Knorr and K. Hamacher (TU Darmstadt, Computational Biology & Simulation) did the setup and analysis of the molecular dynamics (MD) simulations at the high-performance computing cluster (HHLR) in Darmstadt. U. Kuckelkorn and P.M. Kloetzel (Charité Berlin, Biochemistry) performed the *in vitro* site-specific fluorescence substrates activity measurements. The cocrystal structure of BSc2118 bound to the yeast 20S proteasome was kindly provided by M. Groll (TU München, Biochemistry). The following section exclusively treats the MD study. C. Voss, C. Scholz and S. Knorr prepared the manuscript.

### 2.4.2 Introduction

Apart from the proteasome inhibitor optimization in Section 2.3, we investigated the natural proteasome cleavage mechanism. Due to the fast cleavage reaction it is impossible to crystallize the complex of the proteasome with a natural substrate. Here, several inhibitor cocrystals indicate the binding mechanism: inhibitors are trapped at the cleavage site and so it is possible to gain insights in the nature of the  $\beta 5$  active site [68, 64, 69, 61, 14, 111, 106].

Comparing the side chain conformations of the  $\beta 5$  subunit in different crystal structures of the 20S proteasome bound to several proteasome inhibitors to the native state, an extraordinary flexibility of the methionine residue 45 (Met45) side chain is observed. Met45 is located in the vicinity of the catalytic center. This structural rearrangement of the  $\beta 5$  active site was first reported by Groll et al. [64] (RMSD 2.7 Å in the bortezomib cocrystal) suggesting that it creates an induced fit by enlargement of the S1 pocket.

Obviously, covalent inhibitors are trapped at the  $\beta 5$  active site, as the crystals indicate. Inhibitors as well as substrates both being attacked by the catalytically active Thr1, exploit the same binding mechanism. If we observe a Met45 shift in inhibitor binding (Figure 2.12), we conclude that Met45 also plays an important role in natural substrate cleavage. In order to understand both, inhibitor- and substrate-dependent mechanisms, and also to improve covalent inhibitor development we focused on that mechanism.



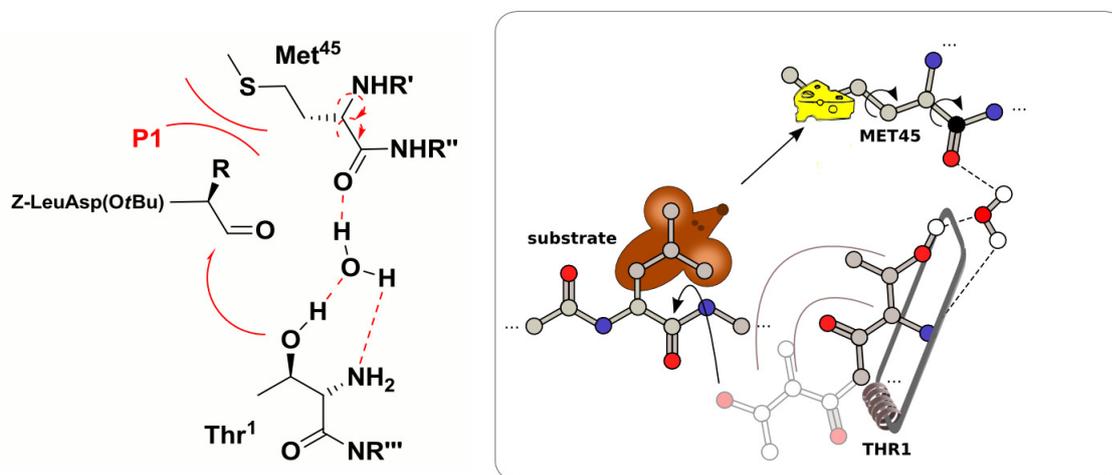
**Figure 2.12:** Met45 side chain conformations of the  $\beta 5$  subunit observed among structures of the yeast 20S proteasome cocrystallized with different inhibitors with reference to the ligand-free native state (green) [63]. Met45 side chain reveals a visible shift upon binding of the following ligands: BSc2118 (yellow) [62], Bortezomib (red) [64], homobelactosin C (purple) [69] and epoxomicin (orange) [68], for PDB codes see Table 2.1. The structural fit was done by taking several pocket  $C^\alpha$  atoms into account. The structure was rendered in VMD.

It is known that the 20S proteasome's  $\beta 5$  subunit cleaves peptide bonds after large, hydrophobic residues and is thus referred to as chymotrypsin-like activity [200]. Combining that fact with the structural knowledge, we proposed that a substrate must possess a P1 residue that is sufficiently large to interact hydrophobically with the Met45 side chain triggering the cleavage mechanism.

The proposed mechanism works according to the "induced fit" principle where the ligand induces a conformational change in the protein partner thereby drastically enhancing their binding affinity. The following consecutive steps are involved:

1. The substrate's P1 residue interacts hydrophobically with the Met45 side chain.
2. The Met45 backbone is pushed away leading to a conformational rearrangement of the Met45 backbone.
3. A water molecule is stabilized via a hydrogen bond network, connecting the backbone to the Thr1 side chain, which increases the Thr1O $^\gamma$  nucleophilicity.
4. Thr1 acting as a nucleophilic base, is now able to attack the substrate's or inhibitor's carbonyl center.

In analogy, the mechanism can be illustrated by a mouse trap (Figure 2.13): the mouse (substrate/inhibitor) is attracted by the cheese (sulfur  $S^\delta$  atom of Met45 side chain), triggering the trip (Met45 backbone), releasing the spring-loaded bar (Thr1) and finally being caught. The information drawn from the crystal structures shows inhibitors "caught in the mouse trap".



**Figure 2.13:** The proposed mechanism of substrate binding to the  $\beta 5$  active site of the 20S proteasome (left), the scheme was kindly provided by C. Scholz. Inside the S1 pocket, the Met45 side chain is pushed away through interactions with the substrate's P1 residue leading to a conformational rearrangement. Illustration of the mechanism in analogy to that of a mouse trap (right) [165].

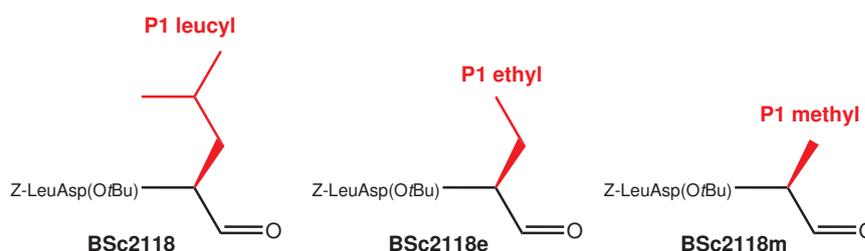
Because the crystal structures only exhibit static structural information we chose an *in silico* approach to shed light on the Met45 dynamics. We thereby focus on the first and most critical step of the mechanism: the suggested interaction of the P1 residue with the Met45 side chain.

The idea was to conduct a comprehensive MD study of a protein-ligand complex, with two subunits of the 20S proteasome shaping the  $\beta 5$  active site that harbor ligands with iteratively enlarged P1 residues. The 20S proteasome inhibitor BSc2118 was chosen as scaffold structure because the chemical synthesis of P1 derivatives was feasible. Based on BSc2118, two compounds with different P1 residues were constructed (Figure 2.14). Together, a total of eight MD simulations (Figure 2.15) was performed:

- The *LIG series* involves four simulations based on the protein-ligand complex structure of the yeast 20S proteasome cocrystallized with the tripeptidic aldehyde BSc2118 at 2.8 Å by the Groll group (unpublished structure). Those four simulations reflect the set with the ligand's P1-residue iteratively *shortened*:

the original ligand with leucyl, ethyl and methyl residue. The fourth simulation reflects the native state of the protein without any ligand. It has to be noted that the covalent bond between the Thr1 residue of the PRE2-subunit and the ligand was broken to mimic the state before ligand binding.

- The *NAT series* involves four simulations based on the native yeast 20S proteasome structure crystallized without any ligand at 2.4 Å (PDB code: 1G0U) [70]. Those four simulations reflect the set with the ligand's P1-residue iteratively *extended*: for three simulations, the original BSc2118 ligand with leucyl residue, the methyl and ethyl derivatives were placed inside the pocket of the native structure, after a structural fit of both proteasome structures. The fourth simulation reflects the unaltered, native state of the protein. The covalent bond between the Thr1 residue of the PRE2-subunit and the ligand was not established according to the *LIG series*.



**Figure 2.14:** The three ligands with different P1 residues applied in this MD study: the original ligand BSc2118 with a leucyl residue (left) and its derivatives BSc2118e with ethyl residue (middle) and BSc2118m with a methyl residue (right).

The key questions to be solved by this *in silico* approach are as follows:

- Is the Met45 side chain movement related to the inhibitor's P1 residue?
- Do the simulations from the two different sets converge to each other independently from the initial proteasome conformation? More precisely, does the native structure with the full ligand included resemble the ligand crystal structure and, reversely, does the ligand crystal structure with its ligand removed resemble the native crystal structure?



**Figure 2.15:** Overview of the eight MD simulations performed in this work: the LIG series (left) is based on the BSc2118 cocrystal structure [62] whereas the NAT series (right) is based on the native crystal structure (PDB code: 1G0U).

### 2.4.3 Theory: Molecular Dynamics Simulation

Biologically relevant events of proteins such as ligand binding, ion transport in channels or protein folding take place on different time scales [164]. In order to investigate the underlying dynamics of such processes, molecular dynamics (MD) simulations provide information on the atomic motions. In principle, an accurate calculation would be possible by using quantum mechanics by describing the exact behavior of the electrons. For large biomolecules, the computational cost would be dramatically high and thus, approximations are inevitable. In classical molecular mechanics, electrons are neglected and motions of nuclei are calculated exclusively [92].

The reduction of an accurate quantum description to a classical mechanical potential underlies three assumptions [135]:

- The separation between electronic and atomic motions is justified by the assumption that slow nucleic degrees of freedom are different to the fast motion of the light electrons (Born-Oppenheimer approximation) [18].
- The potential energy of the system may be written as a sum of different potentials treating the atom cores as classical particles (additivity).
- The potential energy function that has been validated for a small set of molecules is still valid for a wider range of molecules having similar chemical groups (transferability).

In MD, the biomolecule is typically surrounded by a solvent and centered in a box with periodic boundary conditions. The system contains  $N$  atoms, each with a vector  $\mathbf{r}_i$  for atom  $i \in 1..N$  of  $x$ ,  $y$  and  $z$  coordinates defining the atom's position in the Cartesian space. The position of every particle is updated at every discrete time step  $t$  by integration of Newton's equation of motion

$$\mathbf{f}_i = m_i \cdot \frac{\partial^2 \mathbf{r}_i}{\partial t^2} \quad \mathbf{f}_i = - \frac{\partial V(\mathbf{r}_i)}{\partial \mathbf{r}_i} \quad (2.10)$$

where the vector  $\mathbf{f}_i$  is the force acting on atom  $i$ ,  $m_i$  denotes the mass of a particle  $i$  that is multiplied by the second order derivative with respect to time  $t$  of the atomic coordinate vector  $r_i$ . In MD, forces  $\mathbf{f}_i$  are evaluated at every time step by the potential energy function  $V$

$$V = \sum V_{bonded} + \sum V_{non-bonded} \quad (2.11)$$

that is the sum of the potentials  $V_{bonded}$  for bonded and  $V_{non-bonded}$  for non-bonded interactions. The bonded interaction potential is defined as

$$V_{bonded} = \sum_{\text{bnd}} \frac{k_d}{2} (d - d_0)^2 + \sum_{\text{ang}} \frac{k_\theta}{2} (\theta - \theta_0)^2 + \sum_{\text{dih}} \frac{k_\phi}{2} (1 + \cos(n\phi - \phi_0)) + \sum_{\text{imp}} \frac{k_\psi}{2} (\psi - \psi_0)^2 \quad (2.12)$$

including the potentials for deformation of bonds  $d$  (bnd), bending of angles  $\theta$  (ang) and improper dihedral angles  $\psi$  (imp). These are approximated by harmonic potentials that model the deviation from the reference  $d_0$ ,  $\theta_0$  or  $\psi_0$  with their respective harmonic force constants  $k$ . The potential energy of the rotation of dihedral angles  $\phi$  (dih) is modeled through a simple periodic function where  $k_\theta$  is related to the barrier height that is defined as the difference between the minimum and maximum potential energy. Further,  $n$  is the number of minima and  $\phi_0$  determines the positions of the minima.

The non-bonded interaction potential is defined by pair-wise, long-range interactions between several non-covalently bound atoms

$$V_{non-bonded} = \sum_{\text{vdw}} 4\epsilon \left[ \left( \frac{\sigma}{s_{ij}} \right)^{12} - \left( \frac{\sigma}{s_{ij}} \right)^6 \right] + \sum_{\text{elec}} \frac{q_i q_j}{4\pi\epsilon_0 s_{ij}} \quad (2.13)$$

where the left term is the potential for dispersion and repulsion interactions (van der Waals forces, vdw) between a pair of neutral atoms  $i$  and  $j$ . It is modeled by a Lennard-Jones 12-6 potential [102] where  $s_{ij}$  is the distance between the atoms  $i$  and  $j$ ,  $\sigma$  is the distance at which the potential energy is minimal and  $\epsilon$  is the dielectric constant of the solvent. The right term defines the electrostatic potential (elec) between a pair of charged atoms and is derived from Coulomb's law [32] where  $q_i$  and  $q_j$  are the point charges of atoms  $i$  and  $j$ . The constant  $\epsilon_0$  is the dielectric constant of vacuum.

Equations 2.11 - 2.13 define a force field that is empirically derived approximating the actual atomic force in biomolecular systems. A force field is supposed to model the potential energy surface (PES) that otherwise would result from quantum mechanics calculations [135].

The Verlet algorithm [189] and leapfrog algorithm [183] are usually used to integrate the function of Equation 2.10. In this section, the software Nanoscale Molecular Dynamics (NAMD) [152] in combination with the CHARMM27 force-field [124] for proteins and the CHARMM general force field (CGenFF 2b7) [184] for small molecules was used to conduct MD runs. NAMD performs particular well in large-scale parallel simulations [152]. Further, in Section 2.2 the Groningen Machine for Chemical Simulations (GROMACS) [155] was used together with the GROMOS69-53a6 united-atom force field [146].

## 2.4.4 Methods

### Construction of Ligand Structures

The ligand BSc2118 structure was simply extracted from the crystal structure of the protein-ligand complex of the yeast 20S proteasome at 2.8 Å [62]. The force field files (parameter and topology files) of BSc2118 were obtained using ParamChem 0.9.6 [185, 186] that returns a parameter and topology file specific for each fragment and for the CHARMM general force field (CGenFF 2b7).

The preparation of the methyl and ethyl derivatives of BSc2118 in Figure 2.14 is more complex and described in the following: First, the ligand's C-terminus is extracted from the crystal structure and hydrogens are added using Babel 2.3.2 [142]. In Molden 5.0 [163], the C-terminal fragment is tailored leading to the methyl and ethyl variant, respectively. The fragments are converted from .pdb to .xyz format, where MM3 force field [2] atom types are assigned. The MM3 is a small molecule force field and is used here for structural optimization as to C-C bond lengths and hydrogen orientation. In Tinker 6.1 [153], the fragments are energy-minimized to an RMSD gradient of  $10^{-4}$  with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimizer. This is followed by a manual assignment of hydrogen and carbon atom numbers and conversion to the mol2 format using Babel 2.3.2. Subsequently, the structure was passed to ParamChem. These files were merged together with the force field files of the main ligand scaffold. Accordingly, the fragments were fitted to the ligand scaffold (on atoms C32, C33 and C37) to get the overall coordinates of the ligand derivatives with ethyl and methyl P1 residues. The structures were inspected visually and saved in PDB format using VMD 1.9 [93].

### Molecular Dynamics Simulation

The eight different simulations performed have different protein/ligand combinations that are divided into the LIG and NAT series (Figure 2.15). Only those two subunits of the proteasome forming the pocket around the  $\beta 5$  active site (subunits  $\beta 5$ /PRE2 and  $\beta 6$ / PRS3) and the ligand BSc2118 were chosen as input structure. In order to ensure that the ligand stays inside the pocket, harmonic constraints with a force constant of 99 kcal/mol/Å<sup>2</sup> were applied to three atoms of each protein subunit (backbone atoms of PRE2 residue Gly107 and C5 residue Arg29, both

distant to the protein pocket) and two atoms of the the ligand (O7 and C9). The simulations were performed in NAMD 2.9 utilizing the CHARMM27 forcefield for the protein and the CHARMM general force field (CGenFF 2b7) for the ligand that was described before. The TIP3P water model [104] was used, and 8 ions were added to neutralize the system. The system was composed of parts with following number of atoms: BSc2118 cocystal subunits  $\beta 5$  (3238 atoms) and  $\beta 6$  (3468 atoms), native crystal structure' subunit  $\beta 5$  (3192 atoms) and  $\beta 6$  (3153 atoms), BSc2118 (81 atoms), BSc2118e (75 atoms) and BSc2118m (72 atoms), water (approximately 42000 atoms). The overall box dimension was  $74 \times 81 \times 86 \text{ \AA}$  with periodic boundary conditions applied.

The system was energy-minimized for 20 ps at a temperature of 310 K and was followed by two consecutive equilibration procedures: In the first run, harmonic constraints were applied to protein and ligand, with water and ions remaining flexible. In the second, harmonic constraints were applied on the protein's  $C^\alpha$  atoms only and the ligand's backbone atoms and the immobilization constraints as described above. Both equilibration runs were conducted over 20 ps. The system was consecutively heated up by temperature reassignment from an initial temperature of 0 to a final value of 310 K with constant temperature and pressure control switched off. The MD production runs were performed over a time period of 50 ns each (with a time step of 1 fs) using a maximum of 64 CPUs in parallel, output frames were written every 100 fs.

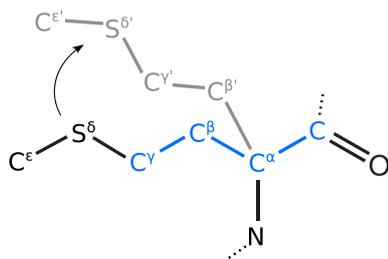
### Trajectory Analysis

In order to save computational time, every 10<sup>th</sup> frame is selected for analysis only. Additionally, the first 10 ns were discarded from analysis due to the initial equilibration phase. Prior to all measurements, every frame of the trajectory is fitted onto the structure in the initial frame in VMD 2.9. There are different possibilities to define the atom set for structural overlay: all  $C^\alpha$  atoms, only  $C^\alpha$  atoms beyond a distance of 15  $\text{\AA}$  to the central pocket and Met45 backbone atoms. The RMSD was calculated according to Equation 2.4.

In contrast to the RMSD, the root mean square fluctuation (RMSF) is the deviation of atomic positions of a molecule averaged over time:

$$\text{RMSF}_i = \sqrt{\frac{1}{T} \sum_{t=1}^T (x_i^t - \tilde{x}_i)^2 + (y_i^t - \tilde{y}_i)^2 + (z_i^t - \tilde{z}_i)^2} \quad (2.14)$$

The RMSF of a specific atom  $i$  of a certain protein gives the mean fluctuation. The deviation is measured for each frame  $t$  with respect to the average coordinates  $\tilde{x}_i$ ,  $\tilde{y}_i$  and  $\tilde{z}_i$ . Here, the deviations are averaged over all  $T$  frames analyzed in a MD trajectory. For the RMSD, RMSF, angle and dihedral calculation, customized Tcl scripts were created for the use in VMD 2.9.



**Figure 2.16:** Nomenclature of the Met45 side chain atoms. The RMSD is measured considering the five atoms  $C^\alpha$ ,  $C^\beta$ ,  $C^\gamma$ ,  $S^\delta$  and  $C^\epsilon$ . The dihedral angle  $\theta$  defines the rotation around the  $C^\alpha$ - $C^\beta$  bond involving the four atoms  $C$ ,  $C^\alpha$ ,  $C^\beta$  and  $C^\gamma$  shown in blue.

## 2.4.5 Results & Discussion

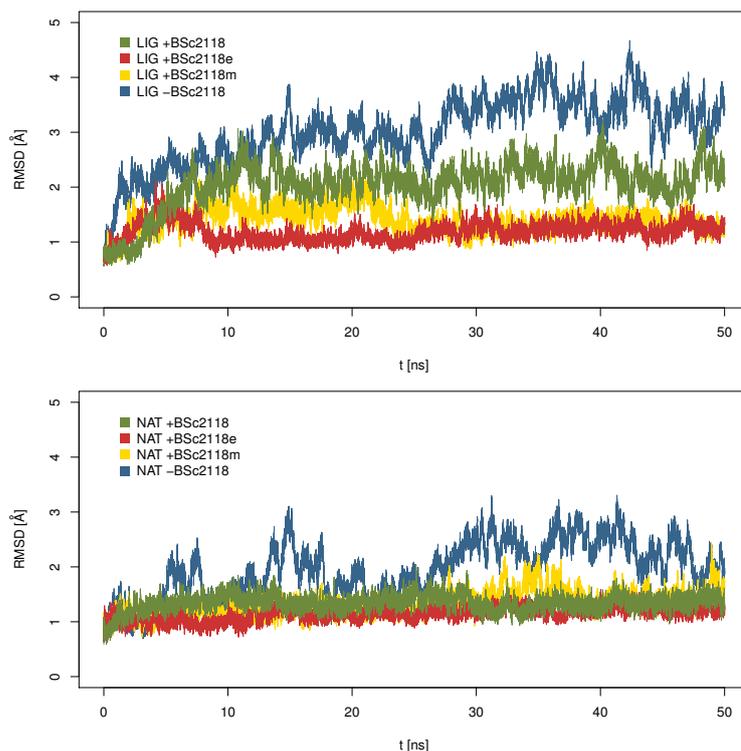
In this section, various structural measures were selected and applied to the trajectories to compare the dynamical properties of the eight different simulations. At first, the general stability and fluctuation of the complexes were analyzed. Then we focused on the detailed behavior of the Met45 side chain.

### Overall Stability

In order to follow the general course of the simulations' trajectories, the structural change within the protein part ( $\beta 5$  and  $\beta 6$  subunits) is monitored over the whole 50 ns time range. For this purpose, the RMSD is calculated for every 10<sup>th</sup> frame with respect to the BSc2118 cocrystal structure that is considered as the *on-state*. Here, the RMSD indicates the protein's fluctuation and if the simulation reached a certain level of equilibrium. Figure 2.17 shows that the RMSD values of all simulations increase during the first 10 ns and then fluctuate around certain RMSD levels. The first 10 ns deviate due to initial inconsistencies and are therefore excluded from all further analyses.

A general difference between the two simulation series is noticed regarding the RMSD range: the LIG series shows increased and different final RMSD values whereas the NAT series indicates a more homogeneous RMSD progression with less variability.

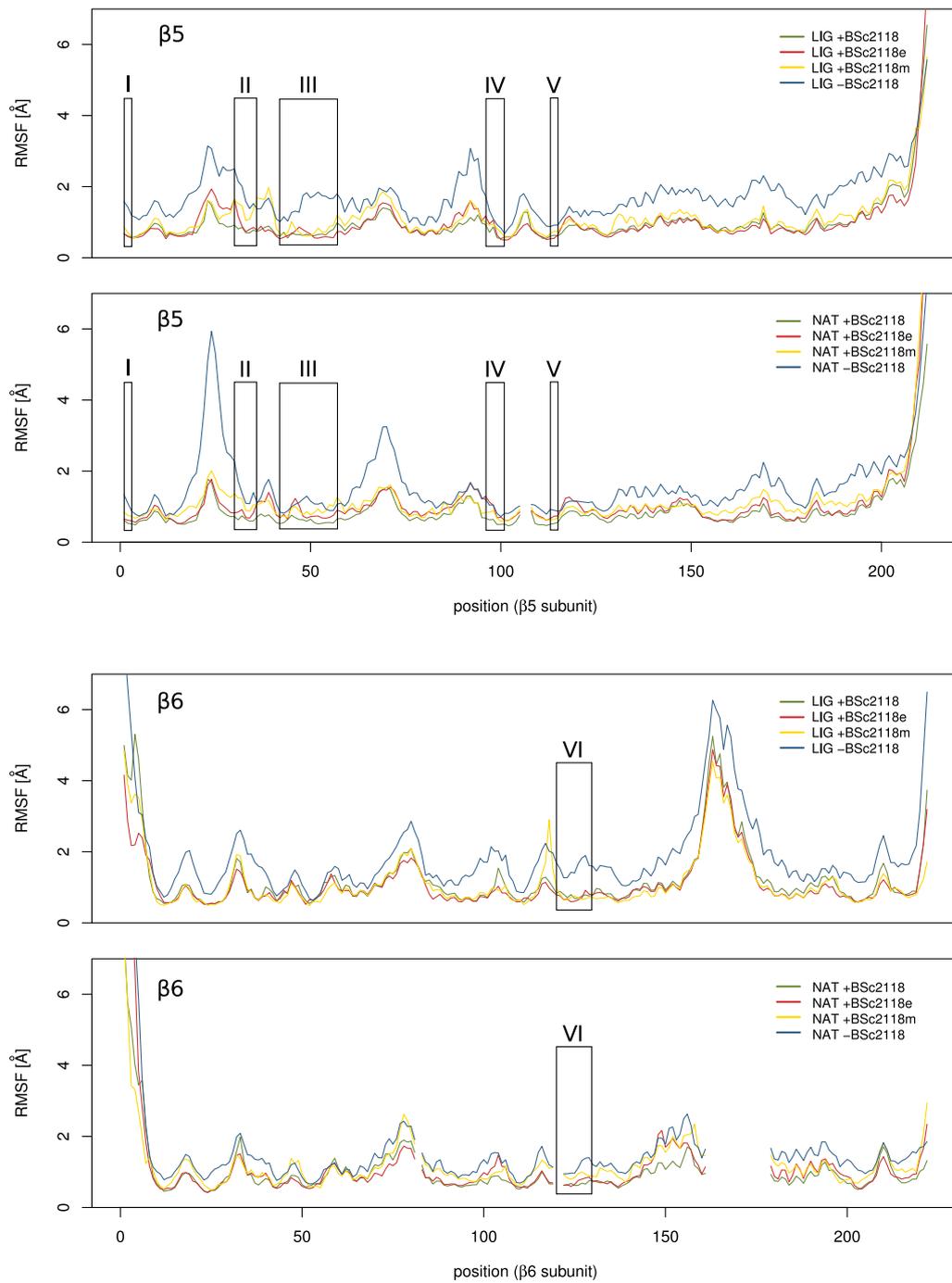
However, the simulations without any ligand (-BSc2118) of both series, NAT and LIG, show the highest variability with RMSD values ranging from 2-5 and 1-3 Å, respectively. This is expected because a ligand bound to the pocket contributes to the stability of the complex whereas in its absence the overall fluctuation is enhanced. Reversely, this would imply a low variability of RMSD values with the full-ligand setup (+BSc2118). This assumption is confirmed in the NAT simulations but not with those of LIG. Nevertheless, in both series the methyl variant (+BSc2118m) shows the expected increased RMSD fluctuation compared to the ethyl variant (+BSc2118e). Notably, the LIG +BSc2118m shows a decrease after 30 ns.



**Figure 2.17:** Root mean square deviation (RMSD) averaged over the two protein subunits ( $\beta 5$  and  $\beta 6$ ) shown for each simulation in the LIG (upper panel) and the NAT series (lower panel) throughout 50 ns simulation time. The RMSD was calculated with respect to the BSc2118 cocrystal structure and measured of each 10<sup>th</sup> frame which equals 1 ps. The identical set of core C $^{\alpha}$  atoms (not being at the surface and outside a radius of 10 Å away from the binding pocket) were used in structural fitting for all 8 simulations.

## Overall Fluctuations

The root mean square fluctuation is a measure of the atomic mobility averaged over the entire trajectory. For each simulation, the RMSF of the C $^{\alpha}$  atoms representing each amino acid residue is shown in Figure 2.18. Defined regions with high fluctuation peaks are visible. Generally, high fluctuations are measured at the terminal ends of both chains. This is not observed for the N-terminal region of  $\beta 5$ , because it constitutes the catalytic center and is highly stabilized inside the binding pocket. As expected, the atoms in the native simulations without ligand (-BSc2118) show consistently the highest mobility in the NAT and LIG series. For the  $\beta 5$  subunit, the fluctuation decreases further for (+BSc2118m), (+BSc2118e) and (+BSc2118) in accordance to the ligand size. This is due to the fact that most of the substrate binding channel is shaped by residues of the  $\beta 5$  subunit. This order is not reflected by the  $\beta 6$  subunit. Herewith, those simulations containing a ligand show similar fluctuation patterns.



**Figure 2.18:** Root mean square fluctuation (RMSF) of the  $C^\alpha$  atoms in the 20S proteasome's  $\beta 5$ -subunit (top, 212 atoms) and  $\beta 6$ -subunit (bottom, 222 atoms) observed in the LIG and NAT simulation series. Regions belonging to the binding pocket are shown: I, II, III, IV, V and VI. NAT simulations reveal certain gaps ( $\beta 5$ : positions 106-107 and  $\beta 6$ : positions 80, 117-118 and 157-173) because residues are missing in the native crystal structure (PDB code: 1G0U). The RMSF was measured for each position and each  $10^{\text{th}}$  frame which equals 1 ps.

In order to observe ligand-related effects we concentrate on binding pocket regions (I-VI): the N-terminal domain with the Thr1 catalytic site (region I, residues 1-3), two  $\beta$ -strands forming the S2 pocket (region II, residues 30-36), the loop where Met45 is attached to (region III, residues 42-57), a  $\beta$ -strand neighboring Met45 (region IV, residues 96-101), a farther located area neighboring Met45 (region V, residues 113-115) and the  $\beta$ 6 subunit harboring the S3 pocket (region VI, residues 125-130). The fluctuations measured in region I, II and VI show a clear dependence on P1 residue size. Regions III and V sites show low fluctuations because they are located too far from the binding pocket.

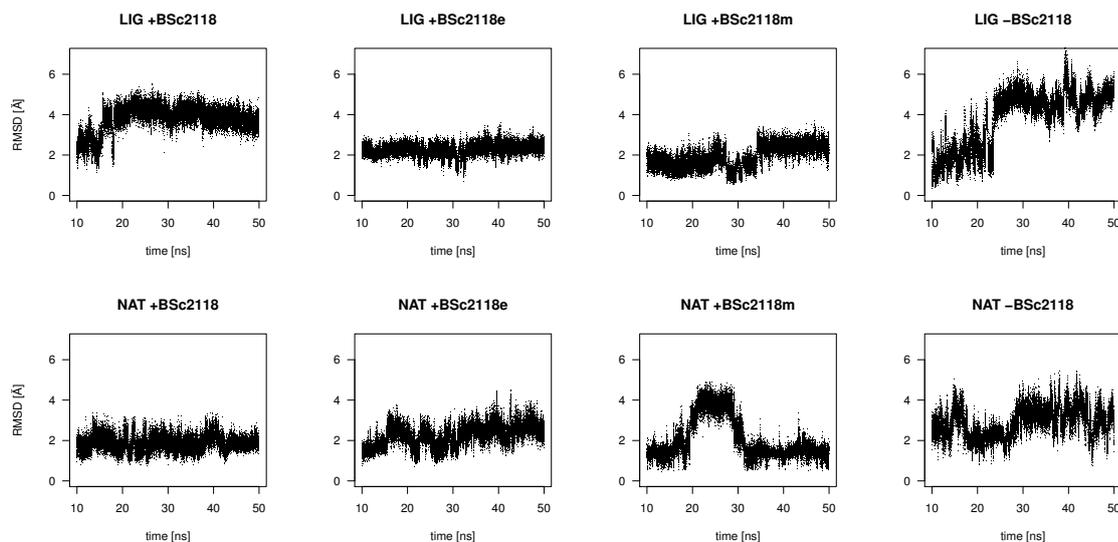
Interestingly, the  $\beta$ 6 subunit residues 157-173 that are absent in the native crystal structure show, in contrast, a high fluctuation in the cocrystal structure.

### Displacement of Met45 Side Chain

The mean displacement of the Met45 side chain atoms is plotted over time (10-50 ns, Figure 2.19). The first 10 ns are omitted due to the initial non-equilibrium phase of MD simulations (Figure 2.17). Prior to the RMSD calculation, the structural superposition was conducted over the Met45 backbone atoms C, N, C $^{\alpha}$  and O, in order to exclusively cover the side chain movement. An alternative fit over the core C $^{\alpha}$  atoms leads to very similar patterns (data not shown) indicating that the Met45 backbone localization does not change much. Figure 2.19 shows the RMSD progression observed in the LIG (upper panels) and the NAT simulation series (lower panels). The RMSD values fluctuate around certain levels leading to different visible states.

Considering the fact that the RMSD is measured with respect to the BSc2118 crystal structure, it is surprising that the full-ligand simulation (LIG +BSc2118) shows a permanent RMSD level around 4 Å. This finding suggests that in presence of ligands with long P1 residues (like BSc2118) the Met45 side chain is indeed kept at a fixed position but not in the conformation observed in the crystal structure. Probably, the Met45 displacement towards the S1 pocket is even larger. The RMSD is lowered with the ethyl and methyl variants (LIG +BSc2118e and +BSc2118m) to a mean level of 2 Å. The methyl variant simulation (LIG +BSc2118m) exhibits an RMSD switch within 25-35 ns revealing at least two states. The ligand-free simulation (LIG -BSc2118) reveals a considerable variability (RMSD values range from 0-6.5 Å) compared to the other simulations in the series offering several different conformational states.

The NAT series also shows a decreased variability (one state) with the full-ligand simulation (NAT +BSc2118). However, here the RMSD values fluctuate around 2 Å which is not the level observed in the LIG counterpart. The ethyl variant (NAT +BSc2118e) exhibits two states. The RMSD progression of the smaller methyl variant (NAT +BSc2118m) reveals two states between 20-30 ns and therefore differs significantly from the LIG counterpart. The ligand-free form (NAT -BSc2118) shows an increased variability but within a lower range (RMSD between 1.5-4.5 Å) which does not correspond to the LIG counterpart (LIG -BSc2118).

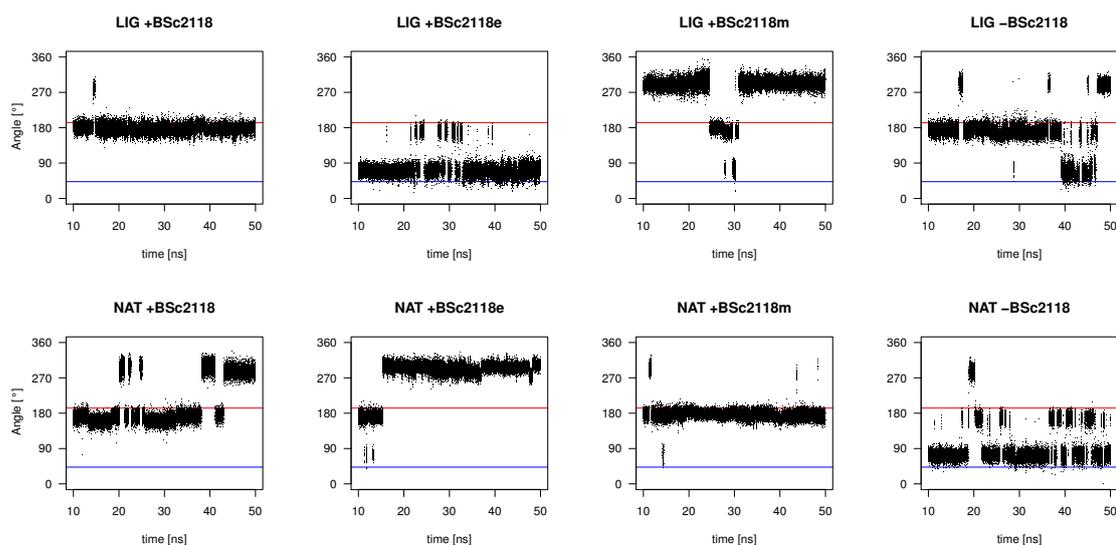


**Figure 2.19:** Root mean square deviation (RMSD) of the four heavy atoms of the Met45 side chain ( $C^\beta$ ,  $C^\gamma$ ,  $S^\delta$ ,  $C^\epsilon$ ) within the simulation time range 10-50 ns. The RMSD was measured with respect to the BSc2118 cocrystal structure and the Met45 backbone atoms were used for structural fitting. LIG (upper panels) and NAT series (lower panels). RMSD was measured of each 10<sup>th</sup> frame which equals 1 ps.

### Rotation of Met45 Side Chain

Another possibility to estimate the Met45 side chain's flexibility is to measure the dihedral angle  $\theta$  spanned by the Met45 backbone and side chain atoms ( $C$ ,  $C^\alpha$ ,  $C^\beta$  and  $C^\gamma$ , Figure 2.16) defining the rotation around the  $C^\alpha$ - $C^\beta$  bond at the base of the side chain. This angle has been selected because the comparison of crystal structures with and without ligands reveals a drastic variation of this dihedral angle ( $\theta_{\text{LIG}} = -166.93^\circ$  and  $\theta_{\text{NAT}} = 42.64^\circ$ , Figure 2.12). The dihedral angle  $\theta$  was measured within every frame without prior fitting to any reference structure and absolute values are monitored over 10-50 ns (Figure 2.20). Angles were measured in the range from  $-180$  to  $-180^\circ$  and were adjusted to  $0$ - $360^\circ$  for visualization purposes only. The picture found in all eight simulations reveals three clearly distinct states: I)  $50$  to  $100^\circ$ , II)  $150$  to  $200^\circ$  and III)  $275$  to  $325^\circ$ .

The full-ligand simulation (LIG +BSc2118) almost exclusively exhibits values belonging to state II. This fully corresponds to the angle observed in the BSc2189 cocrystal structure ( $\theta_{\text{LIG}} = -166.93^\circ$  corresponds to  $193.07^\circ$  in the diagram of Figure 2.20). In contrast, the angle of the shorter ethyl variant (LIG +BSc2118e) is predominantly localized in state I with minor state II occurrences. State III is preferred in the methyl variant (LIG +BSc2118m) simulation revealing a switch to the second and even third state within 25-35 ns. The ligand-free simulation (LIG -BSc2118) shows all three states. This is also observed in the native ligand-free simulation (NAT -BSc2118), but with a predominant residence in state I which



**Figure 2.20:** Variation of the dihedral angle  $\theta$  spanned by the four Met45 residue atoms (C, C $^\alpha$ , C $^\beta$  and C $^\gamma$ ) shown for the LIG (upper panels) and NAT simulation series (lower panels). Measured values originally spanning from  $-180$  to  $180^\circ$  were adjusted to  $0$ - $360^\circ$ . Every 10<sup>th</sup> frame (1 ps) was selected for angle calculation over 10-50 ns. For reference, the dihedral angles observed in the BSc2118 cocrystal (red,  $\theta_{\text{LIG}} = -166.93^\circ$  corresponds to  $193.07^\circ$ ) and native crystal structure (blue,  $\theta_{\text{NAT}} = 42.64^\circ$ ) are indicated by horizontal lines.

corresponds to the dihedral angle measured in the respective native crystal structure. The methyl variant (NAT +BSc2118m) reveals almost exclusively state II whereas the ethyl variant preferably state III (NAT +BSc2118e). The full-ligand simulation reveals state II and III (NAT +BSc2118).

The presence of all three states in the ligand-free simulations can be explained by the fact that Met45 side chain is spatially not constrained and can therefore adopt various conformations. State I and II are found in the native and full-ligand crystal structures suggesting that any medium-sized ligands lead to the same states I and II. The presence of state III suggests it to be an intermediate state that is necessary for the transition from state I into II and vice versa.

## 2.4.6 Conclusion

In order to shed light on the new proposed enzymatic mechanism, the so-called *mouse trap hypothesis* [165], it is necessary to approach the problem from different perspectives. Based on the assumption that the proteasome's substrate cleavage mechanism is triggered by the Met45 side chain, we performed MD simulations as an *in silico* approach to confirm the hydrophobic interaction of the P1 residue with the large side chain of Met45.

Within classical MD, no bond formation or cleavage events take place, i.e. the simulation of the inhibitor binding is not possible. Nevertheless, the model is sufficient to observe hydrophobic interactions. The aim of these computational experiments is to show if the Met45 side chain movement is related to the size of the ligand's P1 residue located in the S1 pocket of the 20S proteasome. A set of eight simulations was performed based on two different crystal structures leading to two distinct series, three of the four simulations within a series contained ligands with differently sized P1 residues (BSc2118, BSc2118e, BSc2118m) and one within a series was a ligand-free simulation.

Observing the overall RMSD, the NAT series shows the expected result that the average atomic displacement increases, if no ligand is present in the pocket. The RMSD progression of those simulations containing ligands is similar. In contrast to the LIG series, RMSD values are higher and do not reveal the same picture observed with NAT. It has to be taken into consideration whether it is sufficient to reproduce the mechanical behavior of the  $\beta 5$  active site by only taking two isolated subunits of the entire architecture of the huge 20S proteasome barrel. Further, the two protein subunits are fixed with three atoms each and the ligand is fixed to place it inside the pocket, which might interfere with the dynamics.

Nonetheless, the RMSF plots show a relation between the fluctability and the size of the ligand. The ligand seems to constrain the fluctuation of the binding pocket. Indeed, the RMSF values of that region increase compared to other parts of the protein. Considering the average displacement of the Met45 side chain only the number of different states adopted is related to the length of the P1 residue. The RMSD calculation accounts for the average displacement only but the dihedral angles give a more detailed picture of the actual conformation adopted by the side chain.

Observing the dynamics of the particular dihedral behavior is straightforward due to its obvious change observed in the crystal structures. The time series of the dihedral angles reveal three different states that correspond to the expected *gauche*(+), *trans* and *gauche*(-) conformations observed generally for the rotation around an amino acid side chain's  $C^\alpha$ - $C^\beta$  bond. The preference for one state is not necessarily related to the ligand size. It seems that the side chain is rather trapped in one conformation than being restrained by the P1 residue. Nevertheless, the two simulations of the original crystal structures (native and full-ligand) confirm the dihedral angles observed from the crystal structure being in a stable state.

A major drawback of MD simulations is the limited simulated time. Several molecular phenomena take place within a larger time scale and, thus, cannot be reproduced. This might be the reason for the missing coherence observed between the NAT and LIG simulation series: the native structure simulated with the inserted ligand did not resemble the cocrystal structure and vice versa, the simulated cocrystal structure with the ligand removed did not resemble the native state. In neither of the plots a cross-convergence could be detected between the LIG and NAT simulation series.

Based on the classical assumption that substrates in the  $\beta 5$  subunit are cleaved only after large hydrophobic residues according to the chymotrypsin-like activity, the Met45 side chain was supposed to act as a trigger. *In vitro* proteasome activity measurements further confirmed this hypothesis, revealing a dramatic increase in activity between ligands offering a the methyl and ethyl P1 residue [165].

Although there is a correlation between the P1 residue size and the inhibitory activity recent extended biochemical assays revealed that this is not true for substrates [165]. The cleavage activity is higher for substrates with small P1 residues. As a consequence, the activity of the  $\beta 5$  subunit is not chymotrypsin-like as generally assumed but rather exhibits a SNAAP activity. This activity of the 20S proteasome was observed before but could not be assigned to a distinct catalytic site. The  $\beta 5$  subunit still cleaves after large, hydrophobic P1 residues but in addition, several differently-sized derivatives can be accommodated [165].

This lies in sharp contrast to the initial inhibitor experiments and clearly contradicts the mouse trap mechanism and, thus, the Met45's role as a trigger. Due to the flexible nature of the Met45 side chain, MD simulation data rather suggests that it is responsible to shape the S1 substrate binding pocket according to any P1 residue size. In this way, the S1 pocket seems to perfectly adapt to various P1 residues. This leads to an ideal structural fit probably further stabilizing the ligand in the channel.

The fact if Met45 is not involved in kinetic acceleration of substrate cleavage cannot be directly confirmed or falsified by MD data. By focusing on the first step of the mechanism only, the MD simulations confirm that there is a relation between the P1 residue and the Met45 side chain with Met45 adapting to different P1 lengths. If this interaction leads to an increased or decreased cleavage cannot be answered by these *in silico* experiments.

### 3 The DNA-PK Complex

Many radiosensitizing agents interfere with DNA damage repair. Since non-homologous end joining (NHEJ) is the major DNA double-strand break (DSB) repair pathway in mammalian cells, one possible approach to selectively killing cancer cells is to focus on the NHEJ disruption [87, 100]. Targeting the NHEJ repair pathway turned out to be an effective strategy to enhance the efficiency of radiation therapy in the treatment of various cancers [122]. The DNA-PK complex (Ku70, Ku80 and DNA-PKcs) plays a major role in this repair pathway, in particular in mediating the first steps of NHEJ. It is therefore a promising target for NHEJ knockout [37].

Therapeutic DNA-PK inhibitors that are available at present exclusively target the ATP-binding pocket of the DNA-PKcs kinase domain [1]. Inhibitors of Ku70/Ku80 were not reported so far [100]. However, a broad clinical application of present DNA-PK inhibitors is limited by inadequate pharmacokinetics. Therefore, alternative more efficient targeting strategies are needed, e.g., small organic molecules targeting protein-protein interfaces to prevent complex formation or antibodies blocking phosphorylation sites [100].

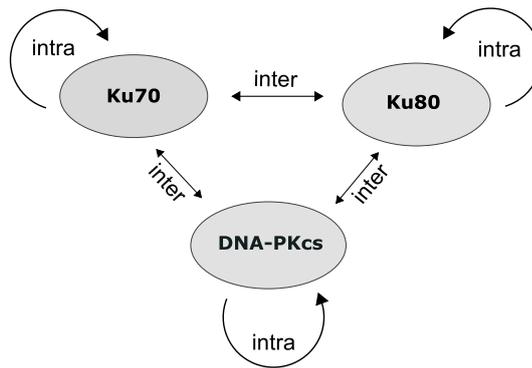
At present, we lack structural knowledge to follow those strategies with rational design. The structures of the single components, Ku70/Ku80 [192] and partially DNA-PKcs [171], have been elucidated giving us a rough idea about the three-dimensional, heterotrimeric topology of the DNA-PK complex. Nevertheless, the exact locations of interfaces and interacting residues remain unknown and thus preventing the development of rationally designed compounds that target the binding interface.

In protein complexes several amino acid residues located at the interface between the components structurally interact with each other. This is of particular importance to maintain complex formation. Like other structurally and functionally important residues, many of them are conserved throughout evolution. Sites of those residues are easily detectable in multiple sequence alignments (MSA) i.e. sets of homologous sequences of a protein. Here, an MSA column refers to a certain position in the protein sequence.

Nevertheless, mutations can occur at essential residues. In this case compensatory mutations can restore the original function [25], which leads to coevolution between the respective residues. The concept of molecular coevolution is analogous to Darwin's pioneering work on coevolution between orchids and fertilizing insects [36]. Analogously, we assume that residues at protein interfaces coevolve by experiencing a selective pressure on maintaining complex formation. Consequently, MSA columns associated with coevolving residues are not perfectly conserved and it is difficult to distinguish those sites from other non-conserved sites [56]. This work

aims at extracting coevolutionary signals out of MSAs to identify key residues for protein-protein interaction.

For this purpose, we apply mutual information (MI) as an information-theoretic measure to detect coevolutionary dependencies between MSA columns. MI measures the degree of correlation between residue substitutions in two alignment columns [56]. Hereby, it is possible to detect dependencies between residues 1) within a protein (Intra-MI) and 2) at interfaces between proteins (Inter-MI) (see Figure 3.1). Using the Inter-MI, the identification of key interacting residues will help to localize the interfaces between two proteins of unknown assembly to further predict the complex structure.



**Figure 3.1:** Evolutionary relations among the protein components of DNA-PK complex: within Ku70, Ku80 and DNA-PKcs (intra) and between the proteins Ku70/Ku80, Ku70/DNA-PKcs and Ku80/DNA-PKcs (inter) examined in this study.

When extracting coevolutionary signals from sequence data it is necessary to reduce undesired background signals. These can be attributed to the following effects:

- **Finite size effects:** In data sets with a limited number of sequences the small sample size causes random high MI signals [196]. Therefore, a minimum number of sequences is required to significantly distinguish real coevolving signals from random ones.
- **Phylogenetic effects:** Since protein sequences did not evolve independently, they have an inherent coevolutionary signal defined by their evolutionary relationship [60].
- **Degree of sequence conservation:** The entropy of two alignment positions correlates strongly with their MI [51, 125]. MI tends to show high signals at quickly evolving and, thus, less conserved sites whereas low signals occur at highly conserved sites.

In this work three different MI normalization procedures were combined with significance thresholds in order to eliminate the MI background signal. We evaluated these methods by their performance in terms of predicting interacting residues in the DNA-PK complex.

## 3.1 Background

### 3.1.1 Molecular Biology

The DNA-dependent protein kinase (DNA-PK) is the key complex formed at a DSB in the early steps of the NHEJ pathway [28]. It is a heterotrimeric complex composed of the proteins Ku (comprising the subunits Ku70 and Ku80) and DNA-dependent protein kinase catalytic subunit (DNA-PKcs) together with a double-stranded DNA (dsDNA) end. In this section, we will focus on the initial events in the first NHEJ phase, the so-called synapsis, where the formation of the DNA-PK complex takes place [195].

After a DSB occurred, each dsDNA end is recognized by a single Ku heterodimer. Subsequently, two Ku heterodimers self-associate to keep the broken ends in close proximity. Upon binding to DNA, this Ku complex shows an enhanced affinity for DNA-PKcs. Each Ku heterodimer recruits a DNA-PKcs protein by the Ku80 C-terminal region (Ku80CTR) [59]. As a result of this, each Ku heterodimer shifts approximately 10 bp inwards along the DNA in order to ensure that both DNA-PKcs molecules can associate with the free DNA ends [192]. Additionally, the two DNA-PKcs molecules interact with each other and this emerging so-called synaptic complex consists of two assembled DNA-PK complexes that keep the broken DNA ends together. This complex provides a stabilizing platform to recruit several other NHEJ enzymes.

### 3.1.2 Structure

**Ku Protein** Heterodimeric Ku consists of two subunits with size of 70 kDa and 83 kDa, named Ku70 and Ku80 (609 aa and 732 aa in *Homo sapiens*) [131]. Being a quasi-symmetrical molecule it is thus assumed that Ku70 and Ku80 are evolutionary related and diverged from a common ancestral homodimer [46].

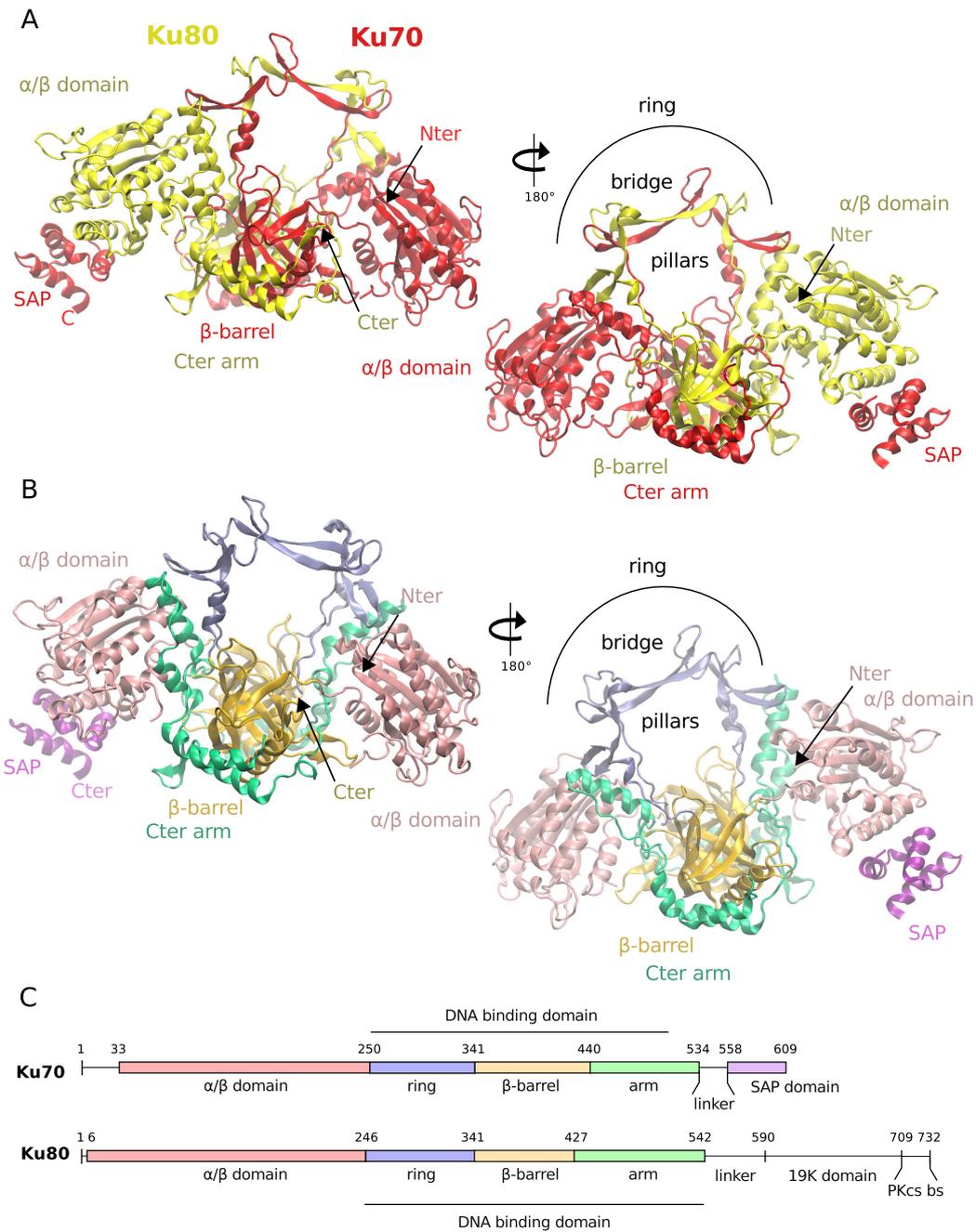
Ku is an abundant protein in the nucleus and binds with high affinity to duplex DNA ends. Remarkably, Ku can bind to several types of broken dsDNA ends (with both 3' and 5' overhangs) showing a great structural plasticity [42]. After a DSB has occurred, its primary task consists in protecting dsDNA ends from degradation and maintain the DNA ends in close proximity.

The Ku protein is a nice example of how a structure is apparently related to its function (Figure 3.2). Human Ku was crystallized together with and without a piece of dsDNA (PDB codes: 1JEY and 1JEQ, 2.5 Å and 2.7 Å [192]). The structural deviation between both Ku states is surprisingly low. Although the core domains of both Ku subunits could be elucidated, several conformationally disordered regions

are missing in the crystal structure due to weak or non-associated electron density. In this work the DNA-unbound structure of Ku70/Ku80 is used as it contains more residues (Figure 3.2). Those amino acids that are present in the structures are listed in appendix Table A.5.

Although sharing a low sequence identity ( $\sim 14\%$ ), the subunits Ku70 and Ku80 show quite a similar topology (An RMSD of  $2.3 \text{ \AA}$  was calculated with respect to the core  $C^\alpha$  atoms after superposing the structures). The crystal structure reveals a relatively large binding interface between the subunits ( $8688.5 \text{ \AA}^2$ ) which contributes to the fact that Ku70 and Ku80 form a very stable complex even in the absence of dsDNA. Describing the fold observed in the crystal structure according to Walker et. al [192], the common architecture of both subunits can be divided into five structural parts:

1. The N-terminal  **$\alpha/\beta$ -domains** have little contribution to make the dimer interface. The crystal structure indicates a six-stranded  $\beta$ -sheet of the Rossmann fold, a structural motif that recognizes nucleotides. Indeed, the Ku70  $\alpha/\beta$ -domain is shifted towards the DNA and is supposed to bind to DNA due to its acidic nature (containing many Asp and Glu residues). The carboxy edge of the sheets is not involved in DNA binding and might be involved in the interaction with other repair factors.
2. The centrally located and evolutionary conserved  **$\beta$ -barrels** are the core domains made of seven  $\beta$ -strands participating in the dimer interface. The domains of both subunits together form the cradle of the DNA-binding groove. This  $70 \text{ \AA}$  cradle can harbor approximately 20 bp of DNA.
3. Extensions of the  $\beta$ -barrels form an asymmetric **ring**. While the ring causes a stabilization of the DNA, it still keeps the DNA accessible to other DNA repair factors. Structurally, the ring can be further divided into a bridge and two pillars. One pillar exhibits three short  $\beta$ -hairpins and is stabilized further through a neighboring helix. Moreover, the bridge prevents Ku from binding to unbroken DNA.
4. Together, the ring and the cradle are positively charged and form the **DNA-binding channel**. Loop extensions of this channel perfectly fit the major and minor groove of dsDNA. The binding mode to DNA is not sequence-specific: there are no interactions with DNA bases but with the sugar-phosphate backbone only. Together, the positive charge and the preformed channel enable a strong binding of dsDNA molecules (dissociation constants  $K_d$  between 0.15-0.4 nM) [46]. Even in the absence of DNA, the ring is formed and only slight structural changes are observed upon DNA binding.
5. The **C-terminal parts** consist of a stretched arm, a linker and a  $\alpha$ -helical domain. Both linkers are highly disordered and therefore not visible in the crystal structures. The C-terminal arms embrace the  $\beta$ -barrel of the respective



**Figure 3.2:** Crystal structure of the human Ku protein (PDB code: 1JEY, at 2.5 Å resolution) [192] consisting of the two subunits Ku70 (chain A, 548 residues) and Ku80 (chain B, 520 residues.) Ku colored according to A) subunits Ku70 (red) and Ku80 (yellow) and B) structural domains. The structures are shown in new cartoon representation in front view (left) and back view (right). C) Positions of structural domains are indicated along the human Ku70 and Ku80 protein sequences (Table 3.1). Non-colored sections are missing in the crystal structure. The structures were rendered using VMD [93].

other subunit. The  $\alpha$ -helical domains clearly diverge between the subunits Ku70 and Ku80. The smaller 5K domain of Ku70 packs against the Ku80 N-terminal  $\alpha/\beta$ -domain. It contains an SAP domain, a helix-turn-helix motif responsible for DNA-binding. The Ku80CTR is the suggested site for DNA-PKcs recruiting, because small angle X-ray scattering (SAXS) data showed that the last 12 amino acids interact with DNA-PKcs [75]. Interestingly, like most regions of DNA-PKcs the Ku80CTR is made of  $\alpha$ -helical HEAT repeats. Moreover the SAXS data demonstrate that the Ku80 domain forms a flexible arm that extends up to 100 Å away from the DNA binding core [75].

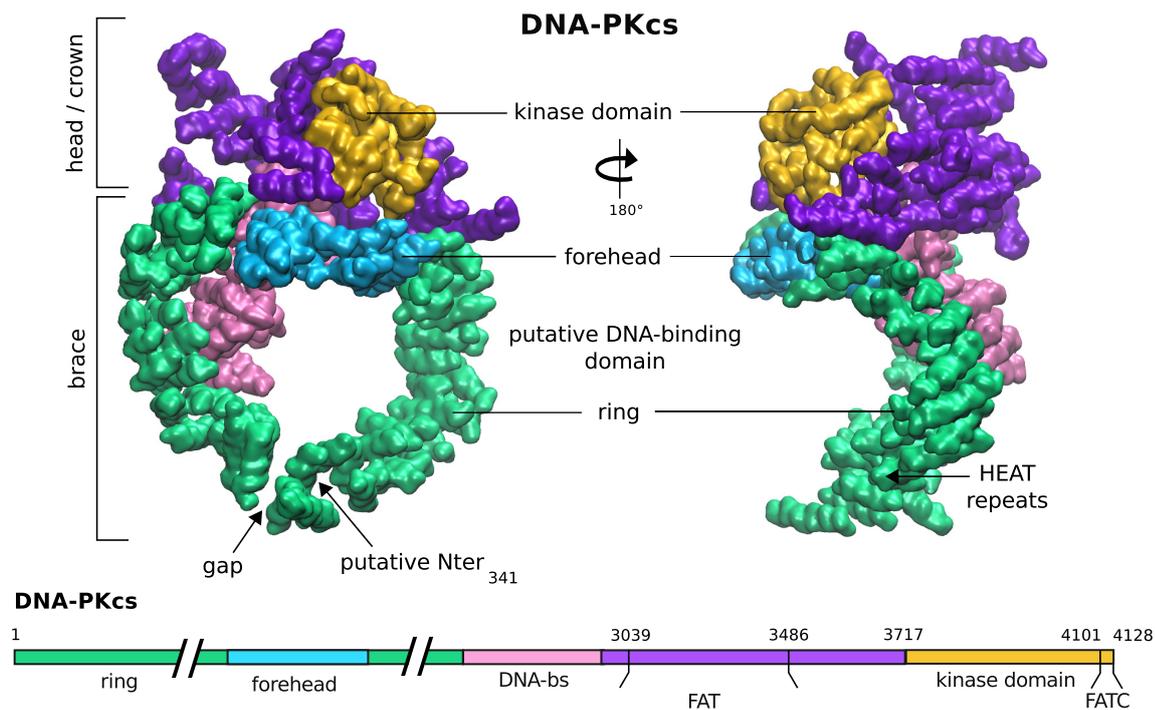
Due to the ring-like structure, two Ku heterodimers must consequently get threaded on the dsDNA strand after the DSB is repaired. It is suggested that Ku is removed from the DNA by degradation of the Ku80 subunit [154].

**DNA-PKcs** The DNA-dependent protein kinase catalytic subunit (DNA-PKcs) is the largest kinase known so far being composed of one single polypeptide chain (4218 aa in *Homo sapiens* and 450 kDa). It belongs to the phosphatidylinositol-3-OH kinase (PI3K)-related kinase (PIKK) family and exhibits a serine/threonine kinase activity that is stimulated upon DNA binding [15].

Due to its size and composition, the elucidation of its structure is extremely challenging. There has been several attempts using cryo-EM (cryo-electron microscopy) and X-ray crystallography, but the resulting medium-resolution structures yield a rough idea on the general topology only. Sibanda and coworkers managed to crystallize the human DNA-PKcs protein and obtained a crystal structure at a resolution of 6.6 Å (PDB code: 3KGV) [171]. Having such a low resolution density only  $\alpha$ -helical regions can be localized with certainty but not the remaining structural element revealing only 46% of the backbone. The identification of the correct residue order is not possible and thus the crystal structure lacks the primary amino acid sequence. The catalytic domain was built by superposition of the related kinase PI3K taken from another crystal structure (PDB code: 1E8X). Additionally, Lindert et al. started to reconstruct the N-terminal HEAT (Huntington, elongation factor 3, protein phosphatase 2A and yeast kinase TOR1)-repeats [121].

The majority of the DNA-PKcs structure is organized in  $\alpha$ -helical HEAT repeats, rod-like structures consisting of helix-turn-helix motifs. These repeats are connected through conformationally flexible loops, thus conferring a substantial flexibility to the entire protein. However, the overall architecture of DNA-PKcs is roughly visible (Figure 3.3) and following putative domains have been defined [171]:

1. The large **brace-like domain** probably contains the N-terminus shaping an extensive circular base with a gap in-between. This region exclusively contains HEAT repeats showing an inner and outer layer of  $\alpha$ -helices. The polypeptide chain appears to form the entire brace first and then continues to the **forehead**. It is suggested that a conformational change might widens the gap for DNA-release [172]. Additionally, a further small globular HEAT repeat domain is attached to one arm of the brace proposed to be a DNA-binding domain [171].



**Figure 3.3:** Fragmented crystal structure of the human DNA-PKcs protein with putative domains: N-terminal ring structure (green), forehead (lime), DNA-binding domain (purple), crown region (red) and the C-terminal kinase domain (yellow). The structure is composed of  $\alpha$ -helices only, therefore lacking the primary acid sequence. The  $\alpha$ -helical fragments are shown in surface representation for visualization purposes. The overall dimensions are of  $160 \text{ \AA} \times 120 \text{ \AA} \times 50 \text{ \AA}$  (PDB code: 3KGV) at a resolution of  $6.6 \text{ \AA}$  [171]. The structure was rendered in VMD.

2. The **head or crown region** is located on top of the brace. Here, the focal adhesion targeting (FAT) domain, the kinase domain and the FATC (C-terminal FAT) domain are located in the order specified. The kinase domain also found in PI3K-related kinases is located on top of the crown, therefore well-positioned for substrate access. The crown region is supposed to contain the C-terminus.

Certainly, the evolution of a large conserved PI3K-related catalytic domain connected to a large brace structure enables DNA-PKcs to function both as an enzyme involved in DNA damage signaling and as a platform for other DNA-repair proteins [144].

Interestingly, the structure of DNA-PKcs was crystallized in complex with the short C-terminal domain of Ku80 (Ku80CTR) which is known to be responsible for DNA-PKcs recruiting. However, it was not possible to locate the Ku fragment in the crystal. As a consequence, the location of the interaction between Ku and DNA-PKcs is still unknown [144].

**DNA-PK Complex** This protein complex consists of the components Ku70, Ku80 and DNA-PKcs [28]. DNA-PK is considered as the regulatory unit in NHEJ, controlling the accessibility of the DNA ends to other repair factors. Due to its size and complexity the structural elucidation of this multicomponent complex is challenging. The organization of the structure is poorly understood, and it is not clear how the individual proteins assemble together with the two broken DNA ends.

Spagnolo et al. presented a cryo-EM structure using single-particle electron microscopy at a resolution of 25 Å [172]. Here, a detailed description of the interactions between DNA-PKcs and Ku is difficult. Anyhow, a rough localization of the DNA-PK regions within the electron microscopy (EM) density was possible.

Another structure of DNA-loaded DNA-PKcs/Ku70/Ku80 complexes obtained from SAXS confirmed the presence of the globular head domain connected to a palm region [75]. The data clearly show a dimerization of two DNA-PK complexes each harboring a DNA end (DNA-PK dimer). Ku is not involved in dimer formation, only two DNA-PKcs molecules interact with each other.

Still, there is uncertainty about the interaction interfaces of two DNA-PKcs molecules. Whereas Spagnolo et al. suggest an interaction of two DNA-PKcs molecules at the palm-to-palm region interacting at the braces [172], Hammel et al. favors the head-to-head interaction with the palm regions pointed outward [75]. The latter propose this to be a suitable physiological arrangement, where the two kinases located in the head domain are in close proximity to the broken DNA ends. Due to the flexible and elongated nature of the Ku80CTR, Ku might even bind to the DNA-PKcs at the opposite site of the DNA break (trans-binding) [75].

Last but not least it has been proposed that a structural rearrangement upon trans-autophosphorylation of two DNA-PK molecules across the DSB leads to a conformational change and subsequent DNA-PK dissociation [41].

## 3.2 Theory: Mutual Information

### Shannon Entropy

In information theory, Shannon entropy [168] quantifies the uncertainty of the information content of a random variable  $X$

$$H_X = - \sum_{\sigma_x \in X} p(\sigma_x) \cdot \log_2(p(\sigma_x)) \quad (3.1)$$

where  $p(\sigma_x)$  is the probability of realization  $\sigma_x$  of the discrete random variable  $X$ . If the entropy is zero, the outcome of the realization of the variable is eventually constant. The maximal entropy corresponds to a uniform distribution of realizations. If the binary logarithm is used, the entropy is measured in bit. In this study, we measure the entropy of different columns  $X$  in protein alignments. Here, the realizations  $\sigma_x$  are taken from a discrete set of symbols comprising the one-

letter code of the 20 canonical amino acid residues (see *List of Abbreviations*). The alignment gap-character (-) is treated in the same way as amino acids and thus regarded as a 21<sup>st</sup> character. The column entropy indicates the degree of conservation of a certain residue position.

Equation 3.1 can be extended to measure the uncertainty that two random variables  $X$  and  $Y$  share about each other. The joint entropy  $H_{X,Y}$  is then defined as

$$H_{X,Y} = - \sum_{\sigma_x \in X} \sum_{\sigma_y \in Y} p(\sigma_x, \sigma_y) \cdot \log_2(p(\sigma_x, \sigma_y)) \quad (3.2)$$

where  $(\sigma_x, \sigma_y)$  is a pair of symbols  $\sigma_x$  and  $\sigma_y$  at positions  $X$  and  $Y$ , respectively. The symbol pairs can be considered as unique elements in an extended alphabet.

### Kullback-Leibler Divergence

In order to calculate the difference between two distributions the Kullback-Leibler divergence  $D_{KL}$  can be used [116]. It measures the information needed to get from a distribution  $X$  to a distribution  $Y$  by

$$D_{KL}(X||Y) = \sum_{\sigma_x \in X} p(\sigma_x) \cdot \log_2 \left( \frac{p(\sigma_x)}{p(\sigma_y)} \right) \quad (3.3)$$

where  $\sigma_x$  and  $\sigma_y$  are realizations of the random variables  $X$  and  $Y$ . It requires that  $X$  and  $Y$  have the exact same realizations  $\sigma$ ; the  $D_{KL}(X||Y)$  is not defined in case of one entry  $p(\sigma_y)$  equals zero.  $D_{KL}$  is a non-symmetric measure applied to quantify the deviation of two distributions. If two distributions are identical, the  $D_{KL}$  is zero.

### Mutual Information

Here, we relate the actual joint probability distribution  $p(\sigma_x, \sigma_y)$  observed in two MSA columns  $X$  and  $Y$  to the distribution of statistical independence being the product of the marginal distributions  $p(\sigma_x)$  and  $p(\sigma_y)$ . As a special case to  $D_{KL}$  of Equation 3.3 the mutual information (MI) quantifies the difference between those distributions

$$MI_{X,Y} = \sum_{\sigma_x \in X} \sum_{\sigma_y \in Y} p(\sigma_x, \sigma_y) \cdot \log_2 \left( \frac{p(\sigma_x, \sigma_y)}{p(\sigma_x) \cdot p(\sigma_y)} \right) \quad (3.4)$$

where the probabilities  $p(\sigma_x)$  and  $p(\sigma_y)$  are estimated by their observed frequencies of symbols  $\sigma_x$  and  $\sigma_y$  in columns  $X$  and  $Y$ , respectively. MI measures the amount of information that one random variable contains about another random variable. It is non-negative and symmetric  $MI_{X,Y} = MI_{Y,X}$ . An MI value of zero indicates that the two variables are independent. In contrast to other commonly used correlation measures MI can also quantify non-linear dependencies as well as linear ones.

In this study, the MI is applied to detect co-evolutionary signals between amino acid residues within a single protein (Intra-MI) and between different proteins (Inter-MI) leading to Intra-MI and Inter-MI matrices, respectively (see Figure 3.5). Precisely, we measure correlations between two columns  $X$  and  $Y$  of a single protein alignment or a joint protein-protein alignment as it was described before in the Subsection *Shannon Entropy*. The MI was calculated with the BioPhysConnectoR package [85] in R [158].

### Z-Scores

In order to test the calculated MI values for statistical significance, a null model has been applied [73, 196]. The idea is that the single column entropies  $H_X$  and  $H_Y$  (Equation 3.1) remain constant whereas the joint entropy  $H_{X,Y}$  in Equation 3.2 is perturbed. For this purpose, the symbols within each alignment column were shuffled independently 10,000 times with the MI matrix being recomputed. For each column pair, a background distribution of 10,000 MI values is obtained with an average MI and the standard deviation. Hence, the Z-score  $Z_{X,Y}$  is calculated by dividing the difference between the average background MI matrix  $\langle \tilde{\mathbf{MI}} \rangle$  and the actually observed raw MI matrix  $\mathbf{MI}$  by the standard deviation of the background matrices

$$Z_{X,Y} = \frac{\mathbf{MI} - \langle \tilde{\mathbf{MI}} \rangle}{\sqrt{\text{Var}(\tilde{\mathbf{MI}})}}. \quad (3.5)$$

The Z-score represents as a measure of statistical significance accounting for background noise in MI signals. Only those MI values are considered, that exhibit a Z-score above a threshold. Within this thesis, the Z-score threshold was defined at the  $q(75\%)$  quantile of each Z-score distribution (see appendix Table A.7). The application of a Z-score cutoff to a measure  $M$  is denoted by the suffix Z ( $M.Z$ ).

### Mutual Information Normalization

The MI is influenced by several effects e.g. sampling bias and phylogenetic effects. Thereby, background signals are estimated in order to correct the raw MI. Various normalization techniques are to be found in literature. The following three different variants are employed in this work:

- **RCW:** The row column weighting (RCW) normalization specifically addresses the problem that high MI values may arise due to the phylogenetic history of the sequences. This effect was observed in random alignments derived by a simulated tree-like evolution [60]. Some conserved amino acid patterns are more common in the MSA than others. Those sites score high against each other and thus leading to the so-called *row-and-column effect*. The undesired

phylogenetic signal is to be eliminated by weighting every entry  $MI_{X,Y}$  of the raw  $n \times n$  matrix  $\mathbf{MI}$  by the average MI of both columns X and Y

$$RCW_{X,Y} = \frac{MI_{X,Y}}{(\overline{MI}_{.X} + \overline{MI}_{.Y} - 2MI_{X,Y})/(2n - 2)} \quad (3.6)$$

where the denominator is the average over the sum  $\overline{MI}_{.X}$  over all entries in column X and the sum  $\overline{MI}_{.Y}$  over all entries in column Y with the actual  $MI_{X,Y}$  value excluded.

- **APC:** The average product correction (APC) accounts for several sources for background MI arising from random and phylogenetic signals [44]. Precisely, the APC is an estimate for background MI shared by positions X and Y. It is defined as the product of the average MI of columns X and Y divided by the overall mean  $\langle MI \rangle$  of the raw  $n \times n$  matrix  $\mathbf{MI}$ . This correction term is subtracted from each raw matrix entry  $MI_{X,Y}$

$$APC_{X,Y} = MI_{X,Y} - \frac{\overline{MI}_{.X} \cdot \overline{MI}_{.Y}}{\langle MI \rangle}. \quad (3.7)$$

- **MNE:** The background MI was observed to highly correlate with the minimum column entropy (MNE) [125]. Thus, it was suggested to divide raw MI values  $MI_{X,Y}$  by the minimum entropy of their two respective alignment columns

$$MNE_{X,Y} = \frac{MI_{X,Y}}{\min\{H_X, H_Y\}}. \quad (3.8)$$

The minimum entropy  $\min\{H_X, H_Y\}$  of column X and Y can be considered as an upper bound to a  $MI_{X,Y}$  value at a certain position.

The normalization variants RCW, APC and MNE were additionally combined with the Z-score significance approach (RCW.Z, APC.Z and MNE.Z). For this purpose, the normalized matrices were constrained by applying the MI-derived Z-score thresholds. For discrimination purposes, the non-normalized MI will be referred to raw MI.

## 3.3 Methods

### Sequence Retrieval

Amino acid sequences of the three polypeptide chains of Ku70, Ku80 and DNA-PKcs of *Homo sapiens* were retrieved from the National Center for Biotechnology Information (NCBI) database comprising non-redundant (nr) protein sequences (see Table 3.1) [156]. Those sequences were used as query strings to search the nr database for similar sequences of other organisms using the basic local alignment search tool (BLAST) with the `blastp` algorithm [3]. The expect threshold (*E*-value) cutoff was set to  $10^{-5}$ , the maximum target sequences to 20,000 and the BLOSUM62 (block substitution matrix) was used [82]. The gap opening penalty was set to 10 and the gap extension penalty to 0.2. A total of 458 (for Ku70), 412 (for Ku80), 2070 (for DNA-PKcs) sequences were returned. Only eukaryotic sequences were found.

**Table 3.1:** Details of the human amino acid protein sequences of Ku70, Ku80 and DNA-PKcs. Sequences can be retrieved from the NCBI protein database via the NCBI accession number. XRCC: X-ray repair cross-complementing protein, aa: amino acids.

protein	alternative name	NCBI accession no.	length [aa]
Ku70	XRCC6	AAH08343.1	609
Ku80	XRCC5	NP_066964	732
DNA-PKcs	XRCC7	P78527	4128

### Sequence Dataset Refinement

Several filter criteria were applied to the sequences obtained from the BLAST search to improve the signal-to-noise ratio. First, sequences with nonspecific organism names (e.g. *NA* and *synthetic constructs*) were removed. For remaining sequences organism names were truncated: genus and species names were kept whereas additional information like strain specifications were removed. Furthermore, those sequences with keywords e.g. *hypothetical*, *predicted* and *putative* indicating unreliable information were removed. Those sequences exhibiting at least one non-canonical letter from the amino acid alphabet (B,J,O,U,Z,X) or a gap-character (–) were discarded. In order to avoid redundancy, duplicated sequences as well as shorter ones were removed, if they are contained within another longer sequence (so-called subsequences). Depending on the sequence length distributions, adequate length cutoffs  $d$  were applied, to discard exceptionally short and long sequences (for Ku70  $d_{\min} = 450$  and  $d_{\max} = 800$ ; for Ku80  $d_{\min} = 500$  and  $d_{\max} = 1000$ ; for DNA-PKcs  $d_{\min} = 1500$  and  $d_{\max} = 5000$ ). Due to the close sequence similarity of Ku70 and Ku80, false positive sequences of the corresponding

partner were detected and eliminated. During the refinement steps described, the query sequence might have been deleted, as it appears as a subsequence within a longer sequence. In this case, the longer sequence is taken as the new query.

### Multiple Sequence Alignments

The refined sequences were aligned separately resulting in three single alignments of Ku70, Ku80 and DNA-PKcs. In addition, the full-length alignment was reduced to those columns where the query sequence does not contain a gap character. For the alignment calculation, ClustalW 2.1 [117] was used with the BLOSUM62 [82] as substitution matrix. The alignments were visually inspected in JalView 2.0 [194].

### Alignment Combination Approaches

In order to calculate the MI between two proteins, their individual alignments must be concatenated organism-wise. This is not a straightforward task, because the organisms found in both alignments will differ in species and number of representatives. This requires a combination scheme, where sequences within the same organism subset are concatenated to each other. Organisms being present in both alignments are considered only and are combined within one organism subset.

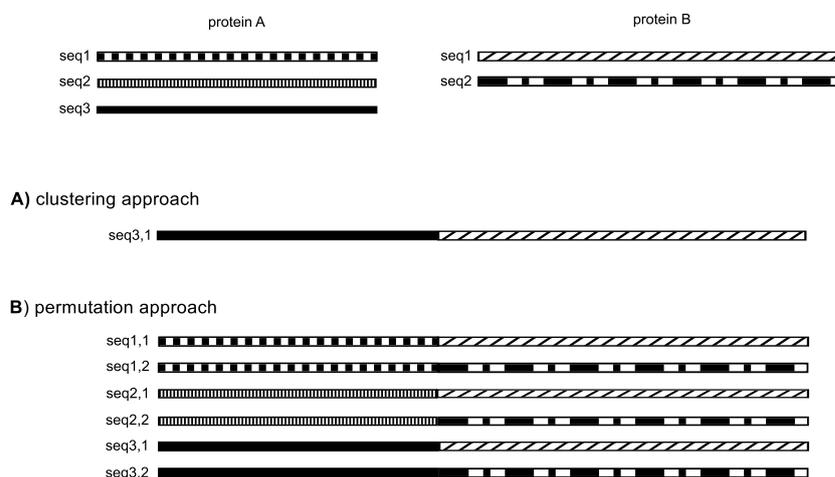
In order to investigate co-evolution between proteins, the concatenation of two single alignments (protein A and B) to a larger combined alignment is necessary (see Figure 3.5). Three combined alignments are produced:

$$\begin{array}{lclcl} Ku70 & + & Ku80 & = & Ku70/Ku80 \\ Ku70 & + & DNA-PKcs & = & Ku70/DNA-PKcs \\ Ku80 & + & DNA-PKcs & = & Ku80/DNA-PKcs \end{array}$$

It is required to divide the sequences of alignment A and B according to the organism names into subgroups. If an organism group is present in one alignment only, it will be discarded. There are two types of alignment combinations termed the clustering and the permutation approach:

- *Clustering*: All sequences within one organism subgroup are compared with each other by computing the Levenshtein distance [120]. This distance measure allows the comparison of strings of different sizes by indicating the minimum number of edit operations required to transform one sequence into the other. The distance increases by 1 if a substitution, gap insertion or deletion event was necessary to equalize the strings. That sequence with the lowest mean distance to the other sequences in the subgroup is selected as the best representative for a certain organism. The mean sequence of alignment B is concatenated to that of alignment A. The total number of sequences is reduced since only one single sequence per organism will be kept.
- *Permutation*: All  $n$  sequences within one organism subgroup found in alignment A are combined to all  $m$  sequences of the same organism subgroup

found in alignment B. The sequences of alignment B are concatenated to those of alignment A realizing all possible  $n \cdot m$  combinations. The total number of sequences increases and certain organism groups drastically get overrepresented (see appendix Table A.6).



**Figure 3.4:** Schematic view of the approaches used to combine two alignments of single proteins A and B to a larger alignment. Sequences are divided into subsets for each organism that is found in both alignments. A) In the clustering approach, the mean sequences of each subset is determined and concatenated to a single large sequence. B) In the permutation approach, all sequences in subset for protein A are combined to those in subset for protein B. All combinations are realized.

### Validating the Inter-MI Matrix (Approach I)

In order to compare the entries of the Inter-MI matrix to those of the contact map from Ku70/Ku80 crystal structure (PDB code: 1JEQ). The contact map is derived from the interaction definition in appendix Table A.10 and visualized in appendix Figure A.8). The rows and columns of the Inter-MI matrices must be adjusted to fit exactly the residues in the crystal structure (548 residues for Ku70, 520 residues for Ku80). Considering the MI methods that are combined with Z-score constraining (MI.Z, MNE.Z, APC.Z and RCW.Z), matrix entries with associated Z-scores above the significance threshold were simply excluded. The assumption is that high Inter-MI values predict interacting residues. The predictive performance of the different MI methods is evaluated using the area under curve (AUC) with respect to receiver operating characteristic (ROC) curves.

## Separating Intra- from Inter-MI signals (Approach II)

The attempt to separate Intra- from Inter-MI signals involves the comparison of both Intra'-MI matrices and Inter-MI ones. Taking their eigenvectors for comparison offers the possibility to transfer the information from the two-dimensional into one-dimensional space as long as the spectral decomposition of the matrices is dominated by a few, or even only one eigenvalue. Eigenvectors and eigenvalues are obtained from singular value decomposition (SVD) where each  $m \times n$  matrix  $M$  is decomposed as

$$M = U \cdot \Lambda \cdot V^T \quad (3.9)$$

where  $U$  is an  $m \times n$  matrix of column-wise left eigenvectors,  $\Lambda$  is a  $n \times n$  diagonal matrix of eigenvalues  $\lambda_i$  and  $V$  is an  $m \times n$  matrix of column-wise right eigenvectors. The original matrix  $M$  can be calculated back by summing up the partial matrices obtained from the full set of  $n$  eigenvectors and  $n$  eigenvalues

$$M = \sum_{i=1}^n \mathbf{u}_i \cdot \lambda_i \cdot \mathbf{v}_i^T \quad (3.10)$$

with  $i \in \{1..n\}$ . The eigenvector  $\mathbf{v}_1$  or  $\mathbf{u}_1$  that is associated with the highest eigenvalue  $\lambda_1$  is termed *first eigenvector* and contains most information of the original matrix  $M$ . The partial matrix  $M_1$

$$M_1 = \mathbf{u}_1 \cdot \lambda_1 \cdot \mathbf{v}_1^T \quad (3.11)$$

where  $\mathbf{u}_1$  being the first left eigenvector,  $\mathbf{v}_1^T$  the transposed right eigenvector  $\mathbf{v}_1$  and  $\lambda_1$  their associated eigenvalue. In order to quantify the information amount of the first eigenvector, the correlation of the partial matrix  $M_1$  and the original matrix  $M$  is calculated (see Section 2.2.3). The Pearson correlation coefficients  $r_{xy}$  for several decomposed Inter, Intra and Intra' matrices are given in appendix Table A.8.

The decomposition of the  $m \times n$  Inter-MI matrix results in first eigenvectors of different dimensions that correspond to the number of residues in protein A and B. As a consequence, the proteins are considered separately. For protein A, the first left eigenvector  $\mathbf{v}_1$  of the Inter-Matrix is compared to the first left eigenvector  $\mathbf{v}_1$  of the Intra-MI matrix of protein A. For protein B, the first right eigenvector  $\mathbf{u}_1$  of the Inter-MI matrix is compared to the first right eigenvector  $\mathbf{u}_1$  of the Intra-MI matrix of protein B.

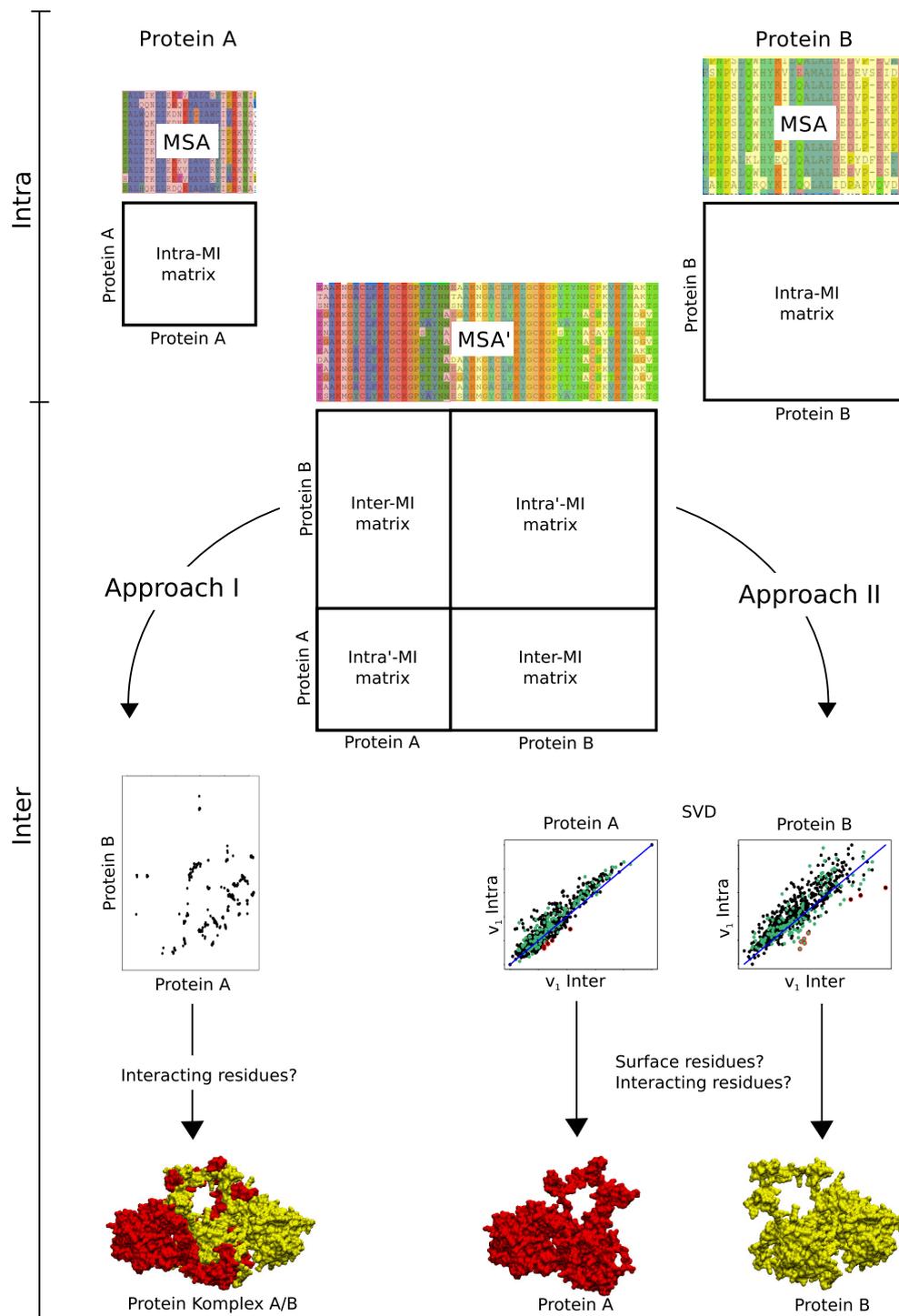
The two eigenvectors for each protein are normalized to the range between 0 and 1 and plotted against each other. Each position in the eigenvector corresponds to a certain protein residue. In order to select those positions which show a stronger Inter-MI signal than Intra-MI signal, the data points located in the lower triangle of the plot are focused and their geometric distance to the diagonal (slope  $m = 1$ , intercept  $b = 0$ ) is measured. Here, we assume that positions with high distances predict residues 1) located at the surface and 2) participating in a protein-protein

interaction. Additionally, the predictive power of the alignment column entropies is assessed. Here, we assume that high entropy values are associated with surface residues and low entropy values with interacting residues.

Selected residues of protein A and B are visualized in VMD 1.9 [93] (see Figure 3.14). In the correlation plot, data points that correspond to residues forming an interaction that is observed in the crystal structure (PDB code: 1JEQ) are marked in green (see Figure 3.11). An interaction is defined according to the interaction criterion in appendix Table A.10.

The assumption is that residues are likely located at the surface or involved in an interaction, are associated with data points distantly located from the diagonal. The predictive performance of the different MI methods is evaluated using the area under curve (AUC) with respect to ROC curves.

In Figure 3.5, an overview is given of the research strategy pursued in this chapter.



**Figure 3.5:** Schematic overview of the research strategy pursued in this chapter. MI of proteins A and B is computed (Intra-MI) from single protein MSA. The alignments are further combined (MSA') and MI is computed again (Inter-MI). Two different approaches were pursued to predict interacting (and surface) residues by I) the entire Inter-MI matrix and II) separating Intra- from Inter-MI using SVD.

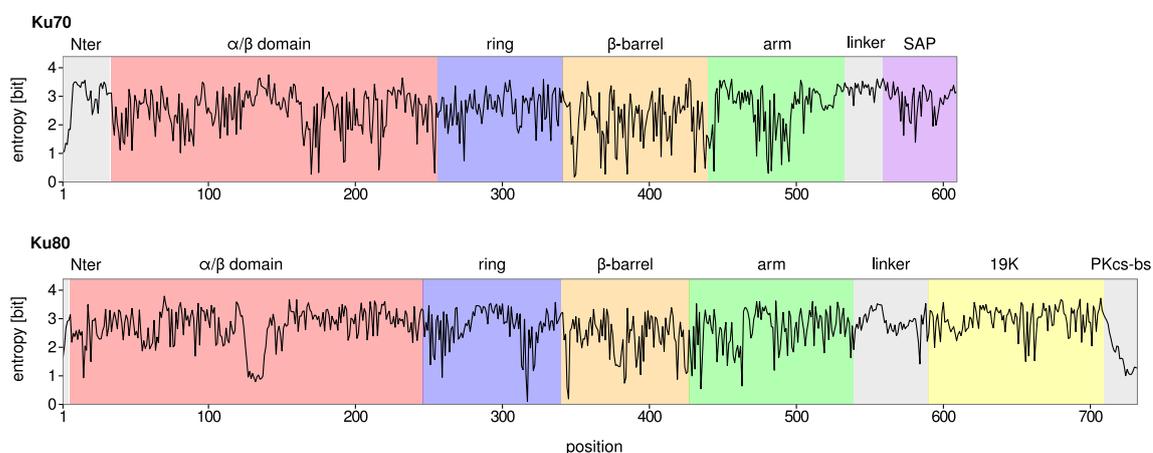
## 3.4 Results & Discussion

The results are presented separately for the investigation of single protein components Ku70, Ku80 and DNA-PKcs (Intra-MI) and inter-protein relations of Ku70 and Ku80 (Inter-MI).

### 3.4.1 Intra Mutual Information

#### Alignment Column Entropy

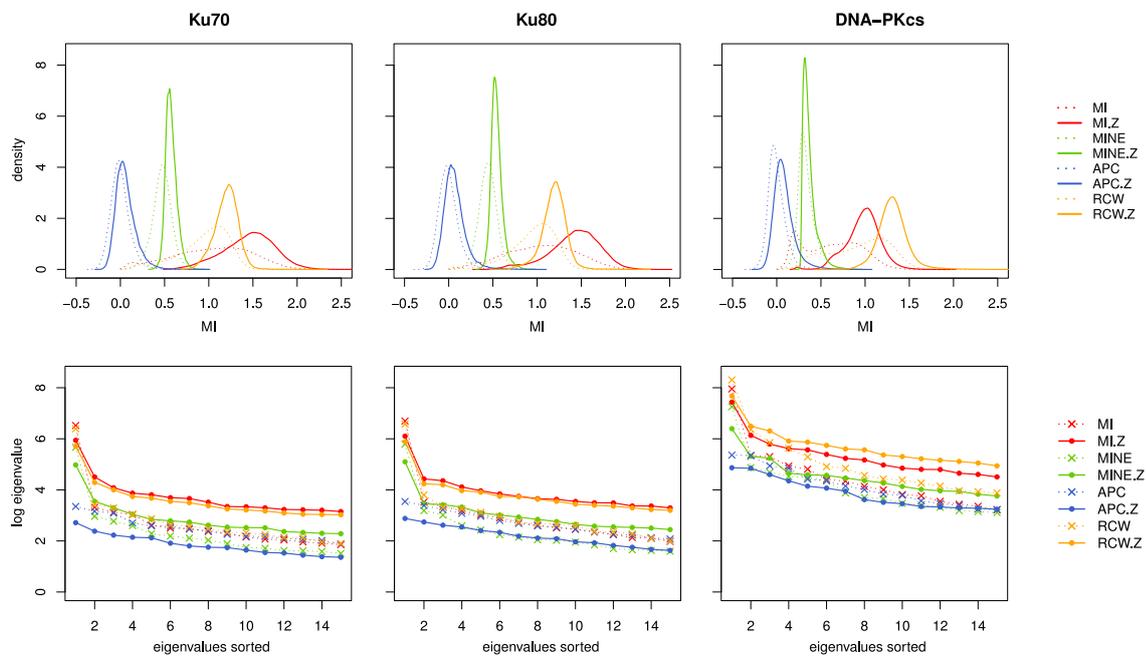
Low entropy values indicate a higher conservation whereas higher ones show a higher variability of the occurrence of amino acids at a certain alignment position. The entropy values of each alignment column of Ku70 (sequence positions 1-609) and Ku80 (sequence positions 1-732) are shown in Figure 3.6. The positions are annotated by the structural domains observed in the crystal structure Figure 3.2. The N-terminal regions and the C-terminal linker that are absent from the crystal structure due to high flexibility exhibit high entropy values. The  $\alpha/\beta$ -domain of Ku80 exhibits a low-entropy site at positions 120-130. The two low-entropy sites in the Ku80 ring domain correspond to the pillars. The  $\beta$ -barrels of both subunits show several low entropy peaks. Interestingly, the distal end (last 12aa) of the Ku80CTR is expected to be highly conserved due to its known function in DNA-PKcs recruitment and conservation. Indeed, a decrease of entropy is observed but this is rather due to a higher gap content commonly found at the alignment ends.



**Figure 3.6:** Alignment column entropies of Ku70 and Ku80 shown along their residue positions. Colors indicate domains and structural motifs observed in the crystal structure (PDB code: 1JEQ, Figure 3.2). The C-terminal domains differ between Ku70 and Ku80 with a DNA binding domain (SAP) and a 19K domain (19K) with the putative DNA-PKcs binding site (PKcs-bs). Regions missing in the crystal structure are colored gray. The structure of the 19K domain (yellow) was determined separately (PDB code: 1Q2Z) [78].

## MI Normalization Variants

**MI Matrix Distribution** Three different normalization variants of the raw MI were analyzed in this study: MNE, APC and RCW (Section 3.2). In addition, the four normalization variants were further analyzed in terms of significance Z-score (MI.Z, MNE.Z, APC.Z and RCW.Z). MI matrix positions showing a Z-score value below a certain threshold defined by the  $q(75\%)$  quantile of the respective Z-score distribution are not considered as significant and therefore these positions are excluded. The respective Z-score thresholds are listed in appendix Table A.7. In order to compare the different normalization variants, the distributions of MI values found in the upper triangles of the Ku70, Ku80 and DNA-PKcs Intra-MI matrices are shown in Figure 3.7. The distributions of all three molecules are similar with respect to the mean values. Apparently, the Z-score criterion affects the filtered distributions in such a way that they become more narrow and approach normal distributions.



**Figure 3.7:** Distributions of Ku70, Ku80 and DNA-PKcs Intra-MI matrices (upper panels). The lower panels show the first 15 entries from the logarithm of sorted eigenvalues obtained from singular value decomposition (SVD) of the MI matrices. The raw MI (MI) and the three different normalization variants (MNE, APC, RCW) are indicated by dotted lines whereas Z-score constrained variants MI.Z, MNE.Z, APC.Z and RCW.Z by straight lines. Z-score thresholds are listed in appendix Table A.7.

**MI Matrix Decomposition** The different matrices are compared by the highest eigenvalues obtained from singular value decomposition (lower panels of Figure 3.7). For all normalization variants except for APC and APC.Z, the first eigenvalue is dominant. Applying a Z-score cutoff to the normalized matrices leads to elevated eigenvalues except for APC.

High eigenvalues indicate that their corresponding left- and right eigenvectors significantly have a large contribution to the respective MI matrix. Partial matrices were reconstructed using the first eigenvector according to Equation 3.11. In order to analyze the correlation between the entries of the partial matrices and the original matrices the Pearson correlation coefficient  $r$  was computed (appendix Table 3.2).

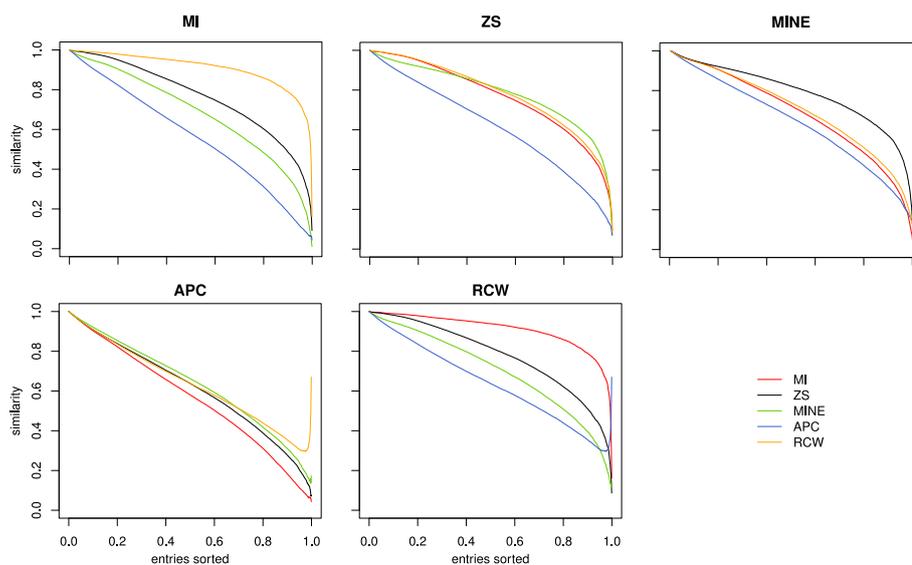
For all three molecules Ku70, Ku80 and DNA-PKcs the same trend is observed: in case of MI and MNE the original matrix was reproduced to a very high degree ( $r > 0.9$ ), whereas with APC, APC.Z, MI.Z and RCW.Z a medium reproduction capacity was observed ( $0.75 < r < 0.8$ ). The correlation was low for MNE and RCW ( $r > 0.5$ ). Comparing the normalization variants it becomes evident that this relation cannot be deduced from the eigenvector dominance seen in Figure A.8 where dominant first eigenvectors are shown with MI.Z, RCW, RCW.Z and MNE.

**Table 3.2:** Relevance of first eigenvectors of MI matrices for different normalization variants obtained by SVD. Pearson's correlation coefficient  $r$  was computed to get the correlation between entries of the partial and those of the original MI matrix.

	Intra-MI							
	MI	MI.Z	MNE	MNE.Z	APC	APC.Z	RCW	RCW.Z
Ku70	0.96	0.75	0.43	0.90	0.79	0.76	0.33	0.75
Ku80	0.95	0.75	0.42	0.86	0.79	0.75	0.26	0.75
DNA-PKcs	0.95	0.76	0.51	0.80	0.76	0.68	0.36	0.71

**Ranking Behavior** The normalization variants are additionally compared based on their ranking behavior of the MI values according to their size. We assume that matrix positions with high MI values indicate important residue pairs being involved in intra-protein or inter-protein interaction.

In order to compare the ranking behavior the similarity measures the overlap between the sets of predicted top residue pairs with a stepwise increasing MI value. Figure 3.8 shows the results for Ku70 only as those for Ku80 and DNA-PKcs exhibit a similar trend (data not shown). In general, a decrease in similarity is observed which indicates that each normalization variant predicts different top residue pairs. With respect to the raw MI all normalization variants exhibit a different behavior. Compared to the Z-score significance matrix (ZS) the APC shows the greatest distance. All normalization variants display a similar distance to APC, and RCW is the variant most similar to the raw MI.



**Figure 3.8:** Similarity of Ku70 residues found in predicted sets of top residue pairs according to different MI thresholds. MI matrices of raw MI, normalization variants MNE, APC, RCW and the Z-score significance matrix (ZS) are compared to each other.

The fact that low similarities are observed between sets of predicted top residue pairs, especially at high MI thresholds, indicates that the three normalization variants as well as the Z-scores effects the ranking behaviors of matrices drastically. Thus, the prediction of residue pairs highly depends on the normalization variant used. These findings encouraged us to continue taking all normalization variants in the following study.

### 3.4.2 Inter Mutual Information

Using the Inter-MI approach aims at predicting interacting residues between two proteins. To this end, the methodology from the Intra-MI section is extended. For the validation of predicted residues available structural information is inevitable and, therefore, the Ku70/Ku80 complex is the subject of investigation.

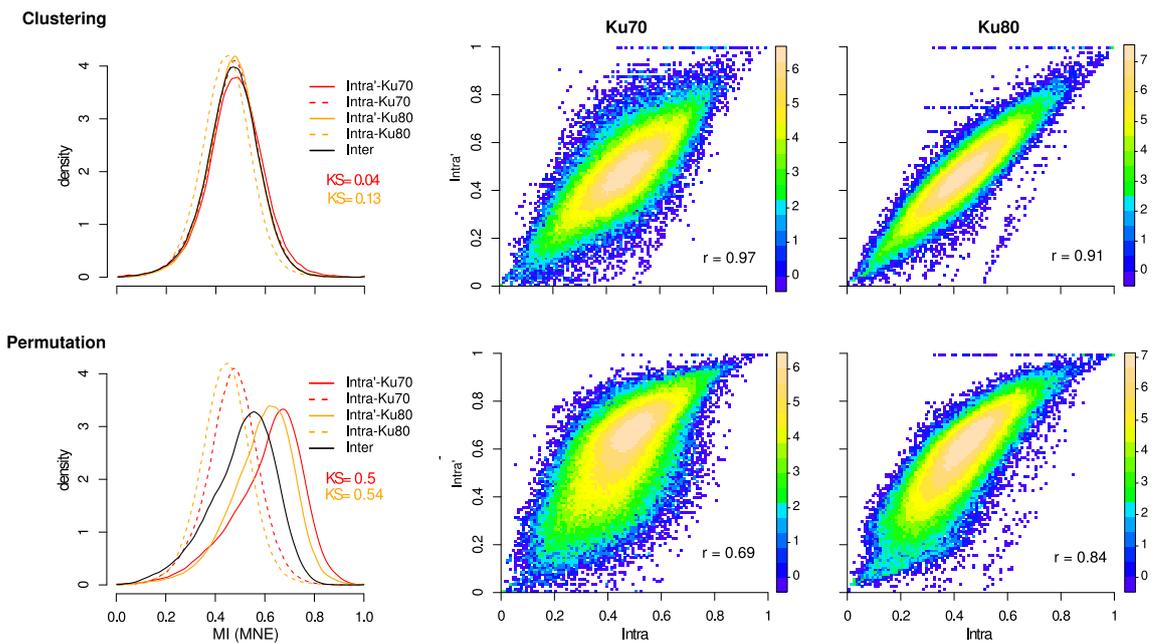
After the combination of the single alignments of Ku70 and Ku80 and subsequent MI recalculation, two approaches are pursued: I) validating the entire Inter-MI matrix II) separating Intra- from Inter-MI signals. An overview of the strategy is given in Figure 3.5.

### Comparison of Alignment Combination Approaches

Two different alignment combination approaches are introduced: the permutation and the clustering approach. In the permutation approach, all sequences within an organism subset are combined to each other, whereas in the clustering approach, the

consensus sequence is considered to be the best representative of the species subgroup and is used for concatenation.

In order to inspect the possible influence of these approaches on the MI signals, the Intra<sup>1</sup>-MI matrix obtained after alignment combination is compared to the original Intra-MI matrix before alignment combination (Figure 3.5). The results for the MNE-normalized data is shown because the influence on MI distribution are strongest (Figure 3.9). For comparison of normalization variants MI, APC and RCW, see appendix Figure A.3 and A.3. For the Z-score constrained variants, the same effects were observed (data not shown).



**Figure 3.9:** Evaluation of the influence of the two alignment combination approaches clustering (upper panels) and permutation (lower panels) on MI. The Intra<sup>1</sup>-MI matrices after alignment combination are compared to the Intra-MI matrices before alignment combination. The effect is demonstrated for the normalization variant MNE for the proteins Ku70 and Ku80, for other variants see appendix Figure A.3 and A.3. Left: MI value distributions from different matrices for Ku70 (red), Ku80 (yellow) and Ku70/Ku80 (black) are shown. The Kolmogorov-Smirnov coefficient KS indicates the distance between two distributions (all p-values  $p = 2.2 \cdot 10^{-16}$ ). Middle and right: Two-dimensional density plots with a color-scale (blue to yellow) exhibit the correlation between all matrix entries of Ku70 (middle) and Ku80 (right). The Pearson correlation coefficients  $r$  are indicated.

Apart from the normalization variant used and from Z-score constraining, it can be observed that the clustering approach affects the MI distributions to a minor degree which is reflected by the low Kolmogorov-Smirnoff distances ( $D_{KS} < 0.15$ ). The correlation between the full matrix entries is relatively high, in particular for

Ku80 ( $p > 0.91$ ). If on the other hand the permutation approach is applied, a drastic shift towards higher MI values is to be seen ( $D_{KS} > 0.5$  and  $p < 0.84$ ).

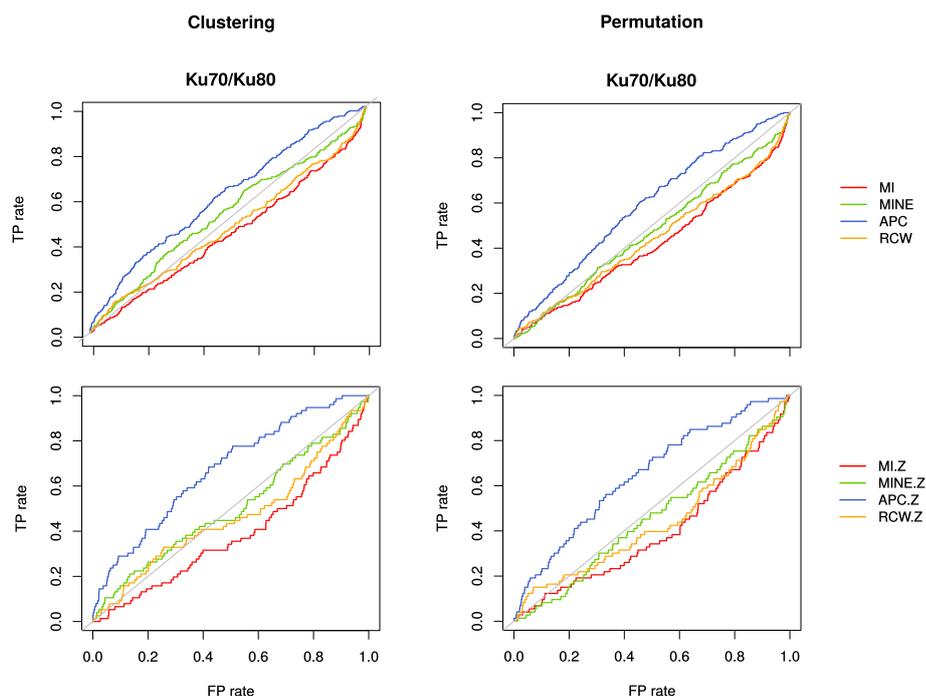
Since sequences are amplified using the permutation approach certain organisms such as *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens* are considerably over-represented (appendix Table A.6). Interestingly, these results tend to show higher MI values. This is most likely due to the fact that the entropy in several alignment columns considerably decreases. Based on these observations the clustering approach was identified to be the more reasonable alignment combination approach since original MI signals are preserved.

### **Approach I: Validating the Inter-MI Matrix**

The performance of the different Inter-MI methods to predict interacting residues is evaluated assuming that high Inter-MI matrix values indicate interacting residue pairs between protein A and B. Predictions are validated by using the contact map of the crystal structure where a contact is defined according to the interaction criterion in appendix Table A.10. In order to compare the Inter-MI method, ROC curves were drawn (Figure 3.10) with the true positive (TP) rate plotted against the false positive (FP) rate. By comparing the AUC, the predictive quality of each method is assessed.

In general, there is no significant difference observed in performance between alignment combination approaches clustering and permutation. For MI, MI.Z and RCW, the AUC is smaller than 0.5 and thus even tend to a predictive power against non-interacting residues. Higher AUC around 0.6 are obtained using APC and its Z-score constrained variant APC.Z. In combination with the clustering approach APC.Z reaches the best performance (AUC = 0.674). Its Z-score constrained variant APC.Z shows the highest performance towards predicting interacting residues (AUC = 0.674).

Nevertheless, the performance of most MI methods towards predicting interacting residues is low. We conclude that coevolutionary signals indicating interacting residues might be covered in the pure Inter-MI matrix. Hence, a second approach is proposed to extract relevant signals out of the Inter-MI matrix (see approach II).



**Figure 3.10:** ROC curves for performance evaluation of Ku70/Ku80 Inter-MI matrices to predict interacting residues. High Inter-MI values are supposed to predict interacting residues. Inter-MI matrices of the uncorrected MI (MI), the three normalization variants (MNE, APC, RCW) and their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z) are evaluated for both alignment approaches clustering and permutation. Predictions are verified using the crystal structure contact map (PDB code: 1JEQ) with the interaction definitions described in appendix Table A.10). The true positive (TP) rate is plotted against the false positive (FP) rate.

**Table 3.3:** AUC calculated from ROC curves shown in Figure 3.10 to assess the performance of interacting residue prediction. Inter-MI matrices of the raw MI (MI), the three normalization variants (MNE, APC, RCW) and their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z) are evaluated for both alignment approaches (clustering and permutation). Associated  $p$ -values are considered to be significant for  $p < 0.05$  and are highlighted in bold type.

	clustering		permutation	
	AUC	$p$	AUC	$p$
MI	<b>0.435</b>	<b><math>2.86 \times 10^{-4}</math></b>	<b>0.420</b>	<b><math>7.57 \times 10^{-6}</math></b>
MNE	0.521	$2.30 \times 10^{-1}$	0.474	$1.51 \times 10^{-1}$
APC	<b>0.600</b>	<b><math>2.32 \times 10^{-8}</math></b>	<b>0.592</b>	<b><math>2.63 \times 10^{-7}</math></b>
RCW	<b>0.464</b>	<b><math>4.26 \times 10^{-2}</math></b>	<b>0.443</b>	<b><math>1.33 \times 10^{-3}</math></b>
MI.Z	<b>0.384</b>	<b><math>4.78 \times 10^{-4}</math></b>	<b>0.396</b>	<b><math>2.04 \times 10^{-3}</math></b>
MNE.Z	0.501	$9.85 \times 10^{-1}$	0.456	$1.94 \times 10^{-1}$
APC.Z	<b>0.674</b>	<b><math>1.61 \times 10^{-7}</math></b>	<b>0.648</b>	<b><math>1.27 \times 10^{-5}</math></b>
RCW.Z	0.464	$2.77 \times 10^{-1}$	0.438	$6.59 \times 10^{-2}$

## Approach II: Separating Intra- from Inter-MI signals

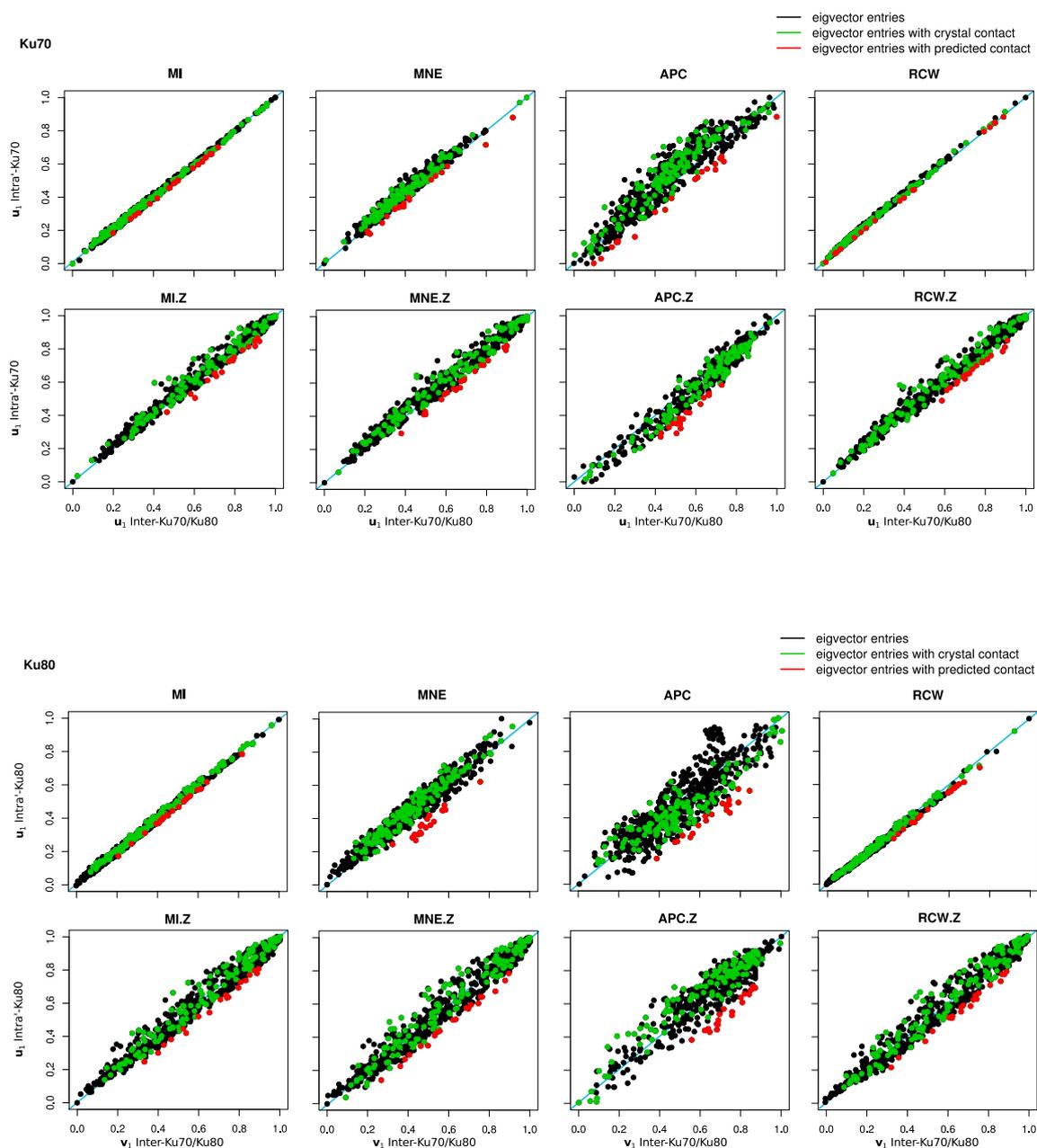
Results of approach I exhibited low predictive qualities. Thus, a more complex procedure is developed. Here, the underlying consideration is that Inter-MI signals might be covered by Intra-MI ones as it was shown that their first eigenvectors are highly related (Figure 3.11). It is therefore planned to extract those Inter-MI values that are composed of a high Inter-MI signal and furthermore of a low Intra-MI signal (see overall strategy in Figure 3.5).

**Eigenvector Correlation** The correlation of the first eigenvectors for the clustering alignment combination approach is displayed in Figure 3.11 for Ku70 and Ku80, separately. The relative dominance of the first eigenvectors is shown in appendix Table A.8. Slight differences between the various MI methods are observed. For raw MI and RCW compared to MNE and APC, high correlations are obtained. Z-score constraining (MI.Z, MNE.Z, APC.Z and RCW.Z) leads to lower correlations with minor differences between the normalization variants. In general, the permutation approach (appendix Figure A.5) exhibits lower correlations, in particular for the Z-score constrained variants. In case of APC and APC.Z for Ku80 no correlation is observed.

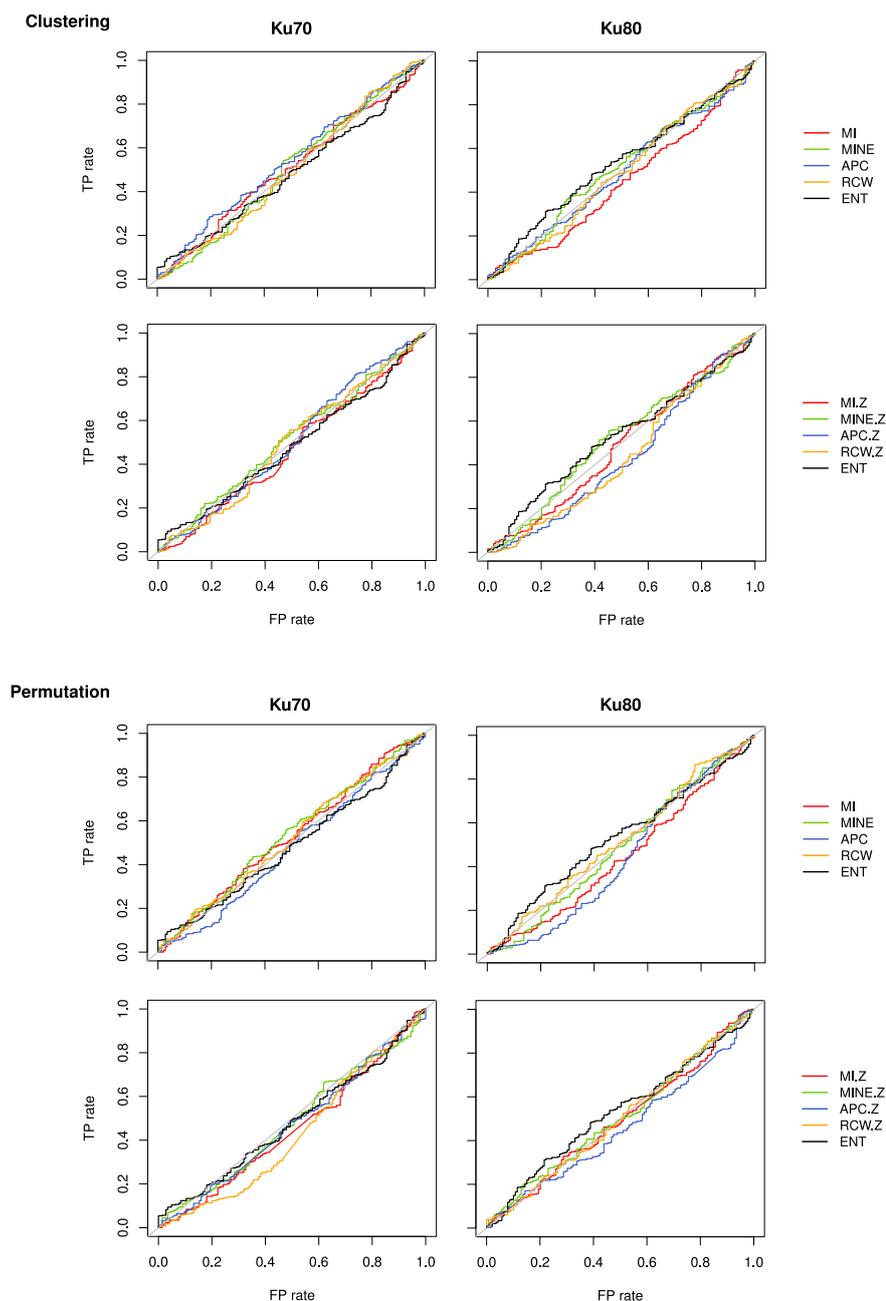
In the following we focus on data points in the lower triangle of the plot being most distant to the diagonal are focused as they exhibit a higher Inter-MI signal than Intra-MI signal. The top 20 most distant data points are selected (red) and their associated residues are proposed to participate in protein-protein interactions and in surface residue prediction. Those residues that were observed to form an interaction in the crystal structure (appendix Table A.10) are highlighted (green) to indicate the TP fraction. Apparently, it is not possible to predict all interactions using this approach because the true positives are quite uniformly distributed over the entire scatter plot without any visible accumulations in the lower triangle.

**Surface Residue Prediction** The different MI methods are evaluated according to their ability to predict surface residues. Assuming that residues with high distances to the diagonal observed in the first eigenvector correlation (Figure 3.11) indicate surface residues. Additionally, the predictive power of the column entropy is assessed by assuming that high entropy values indicate surface residues. Predicted surface residues are validated using the Ku70/Ku80 crystal structure (PDB code: 1JEQ) where a residue is defined to be located at the surface if its solvent-accessible surface area (SASA) value is above a certain cutoff value (appendix Table A.9).

In order to compare the different Inter-MI methods, ROC curves were drawn (Figure 3.12) and AUC values were calculated. ROC curves of several Inter-MI based methods are located near the diagonal (AUC values around 0.5). This observation is consistent for both proteins, Ku70 and Ku80. In conclusion, neither of the observed Inter-MI methods nor the alignment column entropy shows the ability to predict surface residues.



**Figure 3.11:** Comparison of Inter- vs. Intra-MI signals for the **clustering approach**. The scatter plot shows the correlation of first eigenvectors. Upper panels: For Ku70, first left eigenvector  $u_1$  of the Ku70 Intra'-MI matrix and first left eigenvector  $u_1$  of the Ku70/Ku80 Inter-MI matrix. Lower panels: For Ku80, first left eigenvector  $u_1$  of the Ku80 Intra'-MI matrix and first right eigenvector  $v_1$  of the Ku70/Ku80 Inter-MI matrix. Eigenvectors were normalized to the range from 0 to 1. Red dots represent the top 20 data points being most distant to the diagonal (blue). Green dots represent the entirety of the remaining 295 contacts found in the Ku70/Ku80 crystal structure (PDB code: 1JEQ, appendix Table A.10).



**Figure 3.12:** ROC curves for performance evaluation of **surface residue prediction** for Ku70 (left) and Ku80 (right) separately using the approach II. Different MI methods are evaluated for both alignment combination approaches (clustering and permutation): raw MI (MI), the three normalization variants (MNE, APC, RCW), their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z) and, for comparison, the alignment column entropies (ENT). For MI methods and entropy values, it is assumed that high distances to diagonal (Figure 3.11) indicate surface residues. Predicted residues are validated using the crystal structure (appendix Table A.9). TP: true positives, FP: false positives.

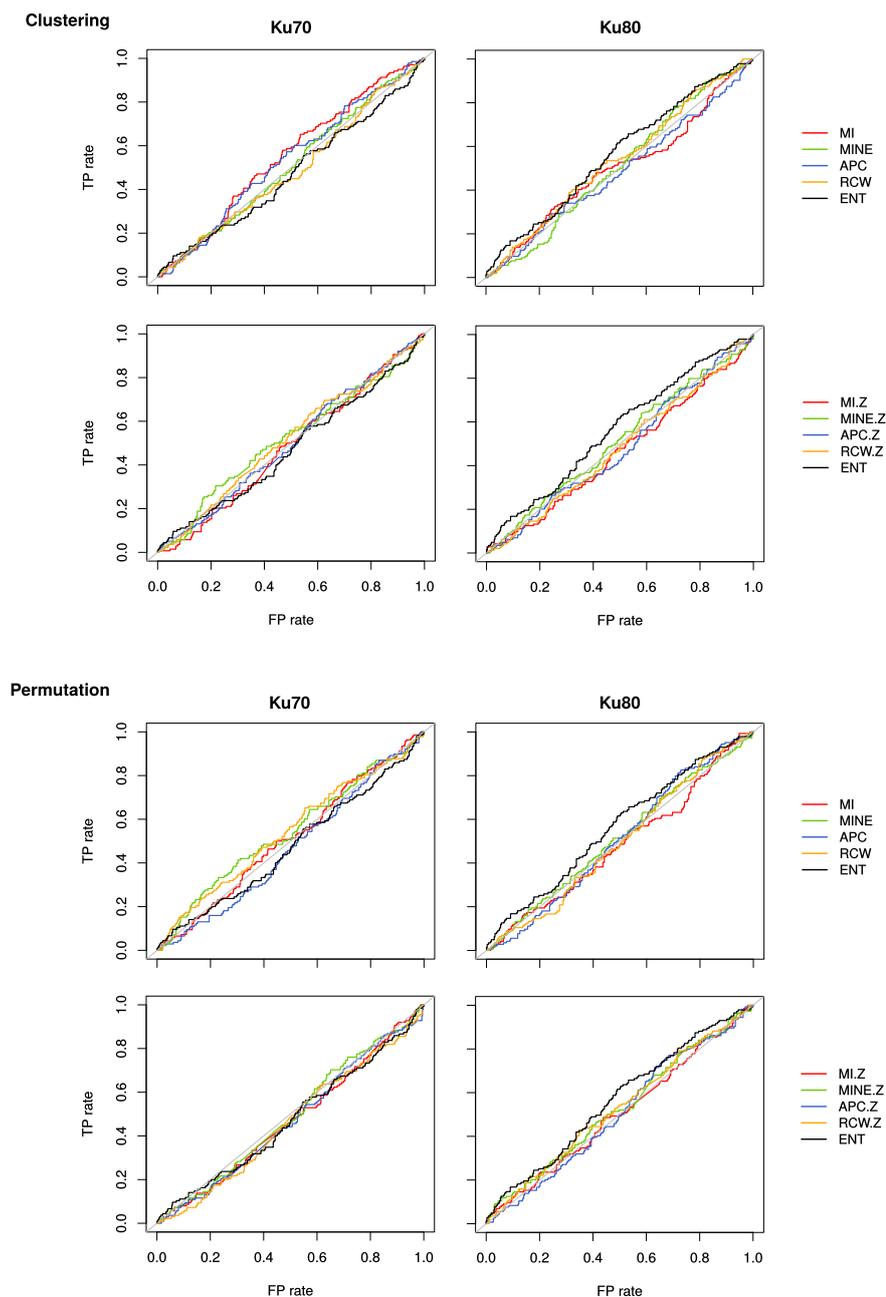
**Table 3.4:** AUC calculated from ROC curves shown in Figure 3.12 with respect to **surface residue prediction**. Different Inter-MI matrices are evaluated for both alignment approaches clustering and permutation: the raw MI (MI), the three normalization variants (MNE, APC, RCW), their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z). Additionally, the alignment column entropy (ENT) is applied. Associated  $p$ -values are considered to be significant for  $p < 0.05$  and are highlighted in bold type.

	Ku70				Ku80			
	clustering		permutation		clustering		permutation	
	AUC	$p$	AUC	$p$	AUC	$p$	AUC	$p$
MI	0.505	$8.59 \times 10^{-1}$	0.526	$3.53 \times 10^{-1}$	0.446	$5.80 \times 10^{-2}$	0.448	$6.82 \times 10^{-2}$
MNE	0.502	$9.52 \times 10^{-1}$	0.533	$2.40 \times 10^{-1}$	0.508	$7.73 \times 10^{-1}$	0.490	$7.38 \times 10^{-1}$
APC	0.539	$1.76 \times 10^{-1}$	0.473	$3.34 \times 10^{-1}$	0.491	$7.49 \times 10^{-1}$	0.443	<b><math>4.78 \times 10^{-2}</math></b>
RCW	0.494	$8.28 \times 10^{-1}$	0.519	$5.00 \times 10^{-1}$	0.488	$6.79 \times 10^{-1}$	0.516	$5.74 \times 10^{-1}$
MI.Z	0.471	$2.96 \times 10^{-1}$	0.451	$8.05 \times 10^{-2}$	0.488	$6.80 \times 10^{-1}$	0.493	$7.95 \times 10^{-1}$
MNE.Z	0.512	$6.72 \times 10^{-1}$	0.481	$4.96 \times 10^{-1}$	0.516	$5.81 \times 10^{-1}$	0.507	$8.14 \times 10^{-1}$
APC.Z	0.503	$9.17 \times 10^{-1}$	0.47	$2.91 \times 10^{-1}$	<b>0.436</b>	<b><math>2.60 \times 10^{-2}</math></b>	0.457	$1.34 \times 10^{-1}$
RCW.Z	0.497	$9.12 \times 10^{-1}$	<b>0.428</b>	<b><math>1.00 \times 10^{-2}</math></b>	<b>0.442</b>	<b><math>4.28 \times 10^{-2}</math></b>	0.500	$9.94 \times 10^{-1}$
ENT	0.483	$5.54 \times 10^{-1}$	0.483	$5.54 \times 10^{-1}$	0.529	$3.18 \times 10^{-1}$	0.529	$3.18 \times 10^{-1}$

**Interacting Residue Prediction** Even though the performance with respect to surface residue prediction was shown to be low, the different MI methods are further evaluated according to their ability to predict interacting residues. In analogy to the surface residue prediction, the underlying assumption is that residues with high distances to the diagonal observed in the first eigenvector correlation (Figure 3.11) indicate interacting residues. Again, the predictive power of the column entropy is assessed by assuming that low entropy values predict interacting residues. This is based on our hypothesis that interacting residues are evolutionary conserved reflected by low entropy values.

Predicted interacting residues are validated using the Ku70/Ku80 crystal structure (PDB code: 1JEQ). Here, a protein-protein interaction is defined if a residue of Ku70 lies within a certain distance cutoff range to another residue in Ku80 and vice versa. The different types of biochemical interactions taken into consideration are listed in appendix Table A.10. Since this interaction criterion is highly dependent on the side chain conformation observed in the crystal structure the side chain geometries were further optimized using rotamer-dependent libraries.

In order to compare the different MI methods, ROC curves were drawn (Figure 3.13) similar to the previous subsection. The results have turned out to be very much alike those observed with the surface residue prediction. All ROC curves are located around the diagonal with corresponding AUC values close to 0.5, consistently for both proteins Ku70 and Ku80. Thus, none of the observed MI methods



**Figure 3.13:** ROC curves for performance evaluation of **interacting residue prediction** for Ku70 (left) and Ku80 (right) separately using the approach II. Different MI methods are evaluated for both alignment combination approaches (clustering and permutation): raw MI (MI), the three normalization variants (MNE, APC, RCW), their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z) and, for comparison, the alignment column entropies (ENT). For MI methods, it is assumed that high distances to diagonal (Figure 3.11) indicate interacting residues and low values do for entropy. Predicted residues are validated using the crystal structure (appendix Table A.10). TP: true positives, FP: false positives.

nor the alignment column entropies show a high performance in predicting residues involved in protein-protein interaction of the Ku70/Ku80 complex.

Although the overall performance of the considered MI methods is relatively low we took a closer look at the lower left part of the ROC curves (appendix Figure A.6). Here, minor differences in the predictive performance between the MI methods are observed.

**Table 3.5:** Area under curve (AUC) calculated from receiver operating characteristics (ROC) curves shown in Figure 3.13 with respect to **interacting residue prediction**. Different Inter-MI matrices are evaluated for both alignment approaches clustering and permutation: the uncorrected MI (MI), the three normalization variants (MNE, APC, RCW) and their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z). Additionally, the alignment column entropy (ENT) is applied. Associated  $p$ -values are considered to be significant for  $p < 0.05$  and are highlighted in bold type.

	Ku70				Ku80			
	clustering		permutation		clustering		permutation	
	AUC	$p$	AUC	$p$	AUC	$p$	AUC	$p$
MI	<b>0.551</b>	<b><math>7.48 \times 10^{-2}</math></b>	0.509	$7.51 \times 10^{-1}$	0.499	$9.69 \times 10^{-1}$	0.484	$5.69 \times 10^{-1}$
MNE	0.505	$8.49 \times 10^{-1}$	0.541	$1.46 \times 10^{-1}$	0.509	$7.56 \times 10^{-1}$	0.513	$6.43 \times 10^{-1}$
APC	0.531	$2.79 \times 10^{-1}$	<b>0.467</b>	$2.49 \times 10^{-1}$	<b>0.484</b>	$5.71 \times 10^{-1}$	<b>0.507</b>	$8.18 \times 10^{-1}$
RCW	0.486	$6.26 \times 10^{-1}$	0.54	$1.58 \times 10^{-1}$	0.535	$2.21 \times 10^{-1}$	0.497	$9.19 \times 10^{-1}$
MI.Z	0.483	$5.58 \times 10^{-1}$	0.471	$3.04 \times 10^{-1}$	0.464	$2.07 \times 10^{-1}$	0.503	$9.07 \times 10^{-1}$
MNE.Z	0.515	$5.88 \times 10^{-1}$	0.49	$7.33 \times 10^{-1}$	0.511	$6.89 \times 10^{-1}$	0.525	$3.80 \times 10^{-1}$
APC.Z	0.499	$9.82 \times 10^{-1}$	0.47	$2.98 \times 10^{-1}$	<b>0.482</b>	$5.20 \times 10^{-1}$	0.504	$8.96 \times 10^{-1}$
RCW.Z	0.515	$6.10 \times 10^{-1}$	0.461	$1.72 \times 10^{-1}$	0.481	$5.03 \times 10^{-1}$	0.527	$3.32 \times 10^{-1}$
ENT	0.473	$3.55 \times 10^{-1}$	0.473	$3.55 \times 10^{-1}$	<b>0.564</b>	<b><math>2.45 \times 10^{-2}</math></b>	<b>0.564</b>	<b><math>2.45 \times 10^{-2}</math></b>

**Structural Details for Top 20 Predicted Interacting Residues** In order to search for differences in prediction quality we observed close-up ROC curves (appendix Figure A.6). Here, a set of the top 20 residues (of Ku70 and Ku80, respectively) with highest MI values are selected. The predicted residues are inspected through visualizing their location in the crystal structure (PDB code: 1JEQ). Comparing the different methods, we found similar top predicted residues in Z-score constrained normalization variants MI.Z, MNE.Z or RCW.Z of the permutation alignment approach. The residues are located within three different sites I, II, III that participate in the Ku70 and Ku80 interface (Figure 3.14). Interestingly, those residues were detected in the top 20 residue sets of all three normalization variants. The sites are defined as follows:

- *site I*: Three residues of the Ku70  $\alpha/\beta$ -domain (Ile75, Ala113 and Ile116) interact with a  $\beta$ -hairpin located at the Ku80 DNA-binding ring (Tyr316), thereby stabilizing the ring structure.
- *site II*: The observed residues participate in an interaction between the arms of the two subunits: Asp441 located at the beginning of the Ku70 arm interacts with the Ku80 arm (Asn484) and further with the Ku80  $\alpha/\beta$ -domain (Arg44). Moreover, our results predict Asp441 to be in contact with the Ku80 Pro485, instead the crystal structure reveals the neighboring Asn484 to establish the interaction.
- *site III*: In the crystal structure, the Ku70  $\beta$ -barrel (Pro429) forms a direct interaction with Phe435 of the Ku80 arm. However, Phe435 is not the predicted partner but instead the neighboring Tyr433. Similarly, Ku70 Phe382 is in contact with Leu438 in the crystal structure but instead the predicted residue partner Lys439 is.

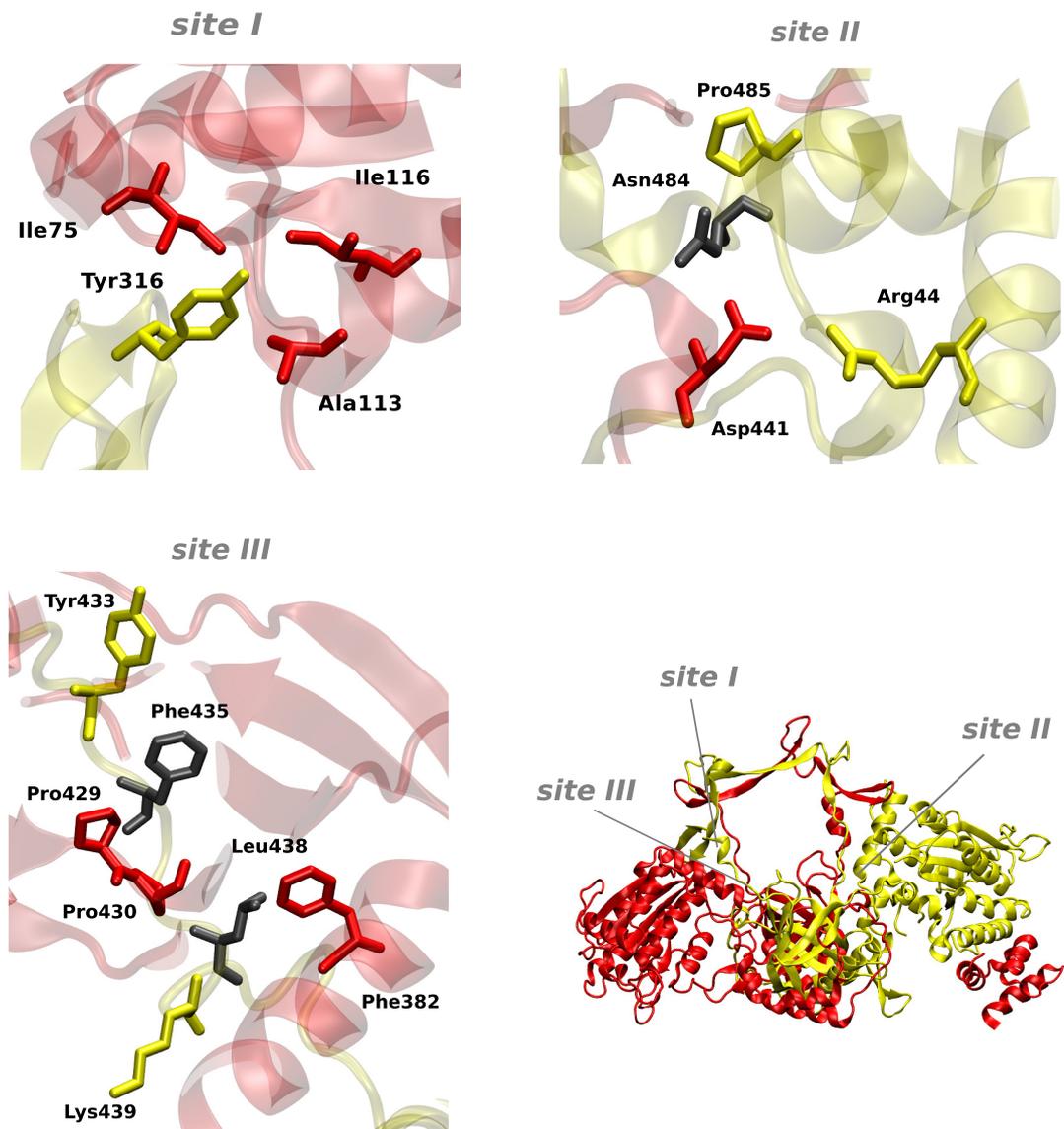
Visualization of the predicted residues show that those found within the top 20 MI values of MI.Z, MNE.Z und RCW.Z Inter-MI matrices indeed form several interaction clusters. It is remarkable that in three cases (in site II and III) the direct interacting residue partner could not be predicted but instead closely neighboring residues were suggested to form the interaction. This might be attributed to the fact that MI-derived methods do not distinguish between direct and indirect relations, a major drawback previously described [24].

Taken together, this leads to the conclusion that the eigenvector-based comparison is insufficient to separate of Intra-MI signals from Inter ones. More sophisticated methods are still needed to extract the relevant coevolutionary signals out of the Inter-MI matrix.

### 3.5 Conclusion

The purpose of this chapter has been a detailed investigation on the DNA-PK complex using a sequence-based study to analyze molecular coevolution. In particular, we focused on the interaction of the Ku70/Ku80 complex because crystal structure information is available to validate the results. Knowledge of the DNA-PK's spatial organization with the interacting interfaces of each component protein is crucial to promote the development of specific protein-protein interaction inhibitors to block the NHEJ repair pathway.

In addition to the amino acid residues that maintain a protein's structure and function those residues mediating protein-protein interactions are also suggested to be constrained due to underlying evolutionary pressure. The majority of mutations observed in proteins are responsible for compensating the deleterious substitutions that occurred at other sites [149]. Those coevolutionary mutations can be detected in sets of homologous sequences of a certain protein belonging to different species.



**Figure 3.14:** Visualization of the three interface sites I, II and III of the Ku70/Ku80 complex crystal structure (PDB code: 1JEQ). The predicted residues are derived from the top 20 residues of Ku70 and that of Ku80 separately using the approach II following the Z-score constrained normalization variant MI.Z, MNE.Z or RCW.Z of the permutation alignment combination approach. The proteins Ku70 (red) and Ku80 (yellow) are shown in new cartoon representation with predicted residues as sticks. A false negative residue is colored gray, if a direct neighboring residue was predicted. The locations of the sites are indicated at the Ku70/Ku80 complex (bottom right).

MI is a parameter-free measure to identify sites of correlated mutations in a multiple sequence alignment. A variety of normalization procedures has been

proposed to normalize the MI to further improve the coevolutionary signal. MI has been successfully applied in several studies to detect intra-molecular coevolution [73, 197, 16]. In this chapter we extend the established method towards analyzing inter-molecular coevolution which already has been described before [58].

This work shows the attempt to compare several MI-derived methods with respect to their ability of predicting protein-protein interacting residues. Hereby, two main methodological approaches were pursued. In the first approach high MI values found in the Inter-MI matrices were tested. Although the predictive quality of most methods are low, the APC turned out to give the best prediction results with a medium prediction. The second and more sophisticated approach consists in separating the Intra-MI signals from Inter-MI ones to extract relevant coevolutionary signals. By taking a detailed look at the top 20 three distinct sites of interacting residues were identified. However, based on the low overall predictive performance we came to the conclusion that the results of approach II arose by coincidence. Therefore, Inter-MI methods need further enhancement.

Molecular coevolution constitutes a complex phenomenon with selective pressure acting on different scales. According to Codoner et al. [30], coevolutionary signals arise from several components:

- stability (secondary and tertiary structures)
- function (catalytically active sites)
- stabilizing interactions with other protein chains (quaternary structures)
- transient interaction with other proteins (formation of complexes)
- folding (dynamical interaction of certain protein parts during the folding process and interaction with chaperones)

Indeed, intra-protein events (stability, catalytic activities, folding) and inter-protein ones (interaction with other protein chains or proteins) are highly related as it has been shown in approach II. This raises the fundamental question if it is possible to disentangle these two phenomena. Certainly, coevolution acts beyond artificial borders as defined in our "inter" and "intra"-distinction. Considering protein-protein interaction as only one subphenomenon out of many others it is not straightforward to simply extract the corresponding signals by measuring MI. Moreover, keeping in mind that most of the proteins exhibit a bunch of interaction partners it is even more challenging to sample the specific signals for one particular interaction. More sophisticated methods seem to be necessary to tackle that problem.

Due to the fact that the Ku together with DNA-PKcs provides a stabilizing platform for several other DNA repair factors it is difficult to extract signals specific for Ku70/Ku80 binding. In addition, this complex exhibits a large heterodimerization

interface [192]. Furthermore, it was shown that Ku70/Ku80 heterodimers can self-associate during NHEJ forming a complex of two heterodimers [75] which requires further interactions contributing to the complexity of the coevolutionary signals.

All in all, the initial machinery of the NHEJ pathway, namely the DNA-PK complex, is marked by a great inherent flexibility. The elongated Ku80 C-terminal arms, the DNA-PKcs autophosphorylation-induced conformational changes and the formation of DNA-PKcs dimers mark the structural plasticity and form a dynamical interaction pattern in space and time dependent on the particular NHEJ step [41]. This further increases the number of interactions and complicates the task of predicting distinct protein-protein interaction interfaces. However, this structural plasticity is exactly the feature that makes the DNA-PK an extremely flexible complex to adapt to the different situations in DSB DNA repair [144].

## 4 The Rad54 Protein

Non-homologous end joining (NHEJ) is the predominant pathway for DNA double strand break (DSB) repair that is available throughout the cell cycle. In contrast, homologous recombination (HR) is often considered as being the minor pathway as it is only available during the late S-phase and the G2-phase. Nonetheless, HR provides high-fidelity DSB repair by employing the homologous sister chromatid as template [170]. Particularly in cells with high proliferative capacity and increased population of S- and G2-phase cells HR is critical for the maintenance of genomic stability [99]. Tumor cells are proliferating.

Since proliferation levels are highly increased in tumor cells, modulating the HR pathway for radiosensitization purposes is a promising approach to selectively inactivate tumor cells. Indeed, a considerable increase in radioresistance is observed in irradiated cells during late S-phase and the G2-phase [178]. Moreover, it was shown that in tumor therapy with heavy ions the induced complex DSBs are primarily repaired via the HR pathway [99].

Paradoxically, it has often been observed that the sensitivity against IR-induced DNA damage is not increased in mammalian HR mutants [103, 161, 113]. This may have its roots in the existence of diverse alternative HR subpathways that complement each other [127]. In cancer cells genomic instability is commonly observed and mutations often accumulate in HR pathways [35] as it is the case in breast cancer associated BRCA1 and 2 defects [188]. Thus, HR modulation is expected to be particularly efficient if alternative pathways have been impaired [99].

Rad54 is one of the key players in HR that fulfills a variety of functions [89] such as stimulating Rad51 strand exchange activity [151], promoting the branch migration of Holliday junctions [23] and remodeling nucleoprotein complexes [98]. In yeast, Rad54 knockout mutants showed an increased sensitivity to ionizing radiation [53]. Surprisingly, murine Rad54 is not essential for viability and mutants only show radiosensitivity during early development [48]. Nonetheless, due to the fact that Rad54 cooperates with the major player Rad51 at several HR stages, Rad54 can be considered a promising target to modulate the HR pathway. One example of a clinically relevant Rad54 inhibitor is Streptonigrin, an aminoquinone antibiotic that targets the ATPase activity of Rad54 [130] by generating reactive oxygen species [38]. It presents a broad antitumor activity against a variety of cancers [77].

Beside the possibility to target individual functions of Rad54 a more effective approach might be the prevention of Rad54 activation. Recent experimental findings from our collaborators of the Löbrich lab revealed that human Rad54 is phosphorylated by NEK1 (never in mitosis gene A-related protein kinase 1) at serine residue

572 (Ser572) [173]. This phosphorylation reaction was identified to be crucial for the activation of ATP-driven Rad54 translocation along dsDNA and thus generally for promotion of HR.

In order to interfere with Rad54 activation a broad understanding of the underlying structural mechanisms is necessary. The aim of this work is to investigate putative structural changes that are induced upon Ser572 phosphorylation and therefore facilitate Rad54 translocation on dsDNA. Two effects are suggested that might 1) trigger ATP-hydrolysis or 2) enable dsDNA-binding.

For analysis of phosphorylation-induced conformational changes we applied a molecular mechanics based method where the linear response theory (LRT) [94] is combined with the anisotropic network model (ANM) [7]. In this coarse-grained approach, the protein is reduced to its C<sup>α</sup> atoms and inter-residue interactions are modeled via harmonic springs. The aim is to predict phosphorylation-induced conformational changes upon perturbation of the single residue Ser572.

## 4.1 Background

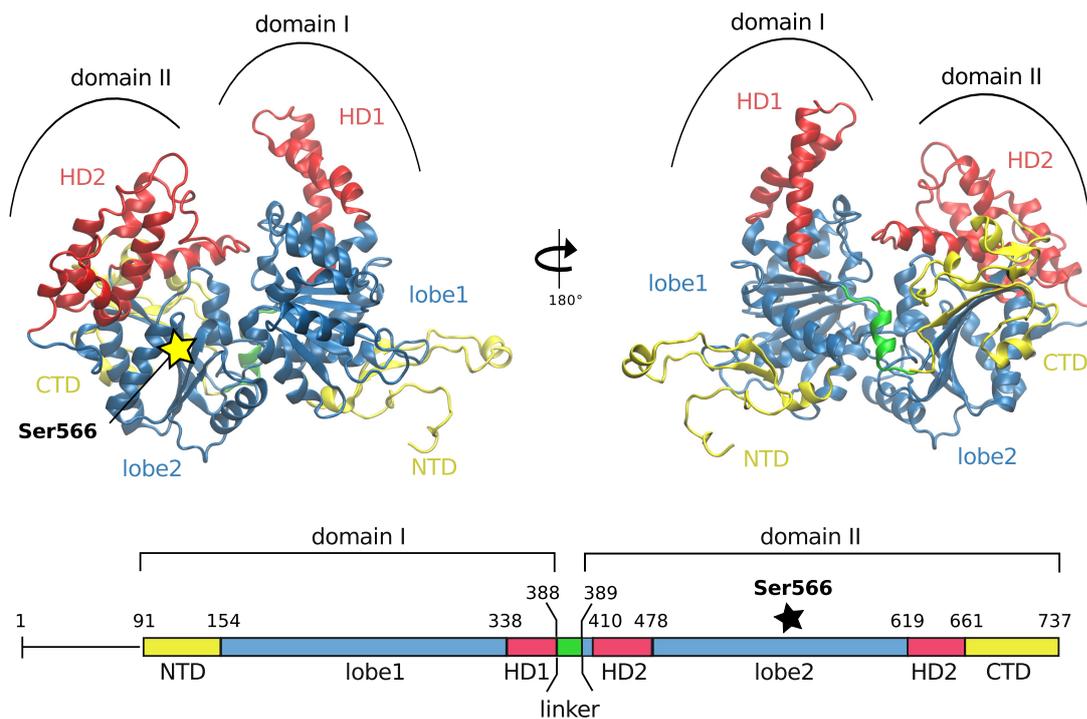
### 4.1.1 Molecular Biology

Rad54 belongs to the switch/sucrose non-fermentable 2 (SWI/SNF2) family of SF2 (superfamily 2) helicases of ATP-dependent DNA translocases. Unlike classical helicases, SWI/SNF2 enzymes act as ATP-dependent chromatin remodeling enzymes: they translocate on dsDNA but do not catalyze DNA strand separation [98]. Instead, they use the energy derived from ATP hydrolysis to remodel chromatin structure by generating negative superhelical torsions in dsDNA [79]. The induction of those topological changes results in the displacement of DNA-bound proteins or even in the disruption of DNA-protein complexes. This enhances the accessibility to nucleosomal DNA. In HR this function is responsible to clear the template chromatid from proteins or nucleosomes during homology search facilitating DNA repair [179].

These biochemical activities 1) ATPase activity 2) dsDNA translocation and 3) remodeling activity render Rad54 as a multifunctional tool that fulfills diverse functions in HR pathway [126]. Rad54 is involved in diverse stages of HR acting tightly together with Rad51. The main functions are: stimulation of Rad51-mediated strand exchange, promoting branch migration of Holliday junctions and further the stimulation of endonuclease activity.

### 4.1.2 Structure

Rad54 is a highly conserved protein in eukaryotes [126] with an amino size of 740 aa and a molecular weight of approximately 90 kDa (Table 4.1). The crystal structure of the Rad54 core domain (ATPase domain) isolated from the zebrafish *Danio rerio* was elucidated at a resolution of 3 Å (Figure 4.1, PDB code: 1Z3I) [179]. In general the Rad54 structure consisting of two major structural domains (domain 1 and 2) is

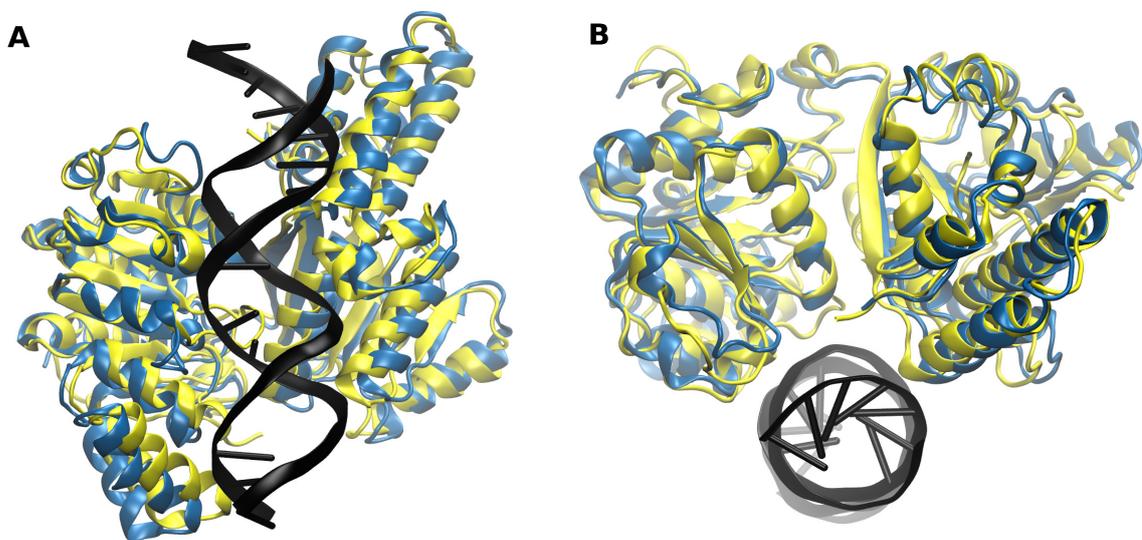


**Figure 4.1:** Rad54 crystal structure of the zebrafish *Danio rerio* at a resolution of 3 Å (PDB code: 1Z3I) [179], front view (left) and back view (right). Two domains 1 and 2 are connected by an  $\alpha$ -helical linker (green). Domain I is shown with the N-terminal domain (NTD, yellow), the  $\alpha$ -helical domain insertion (HD1, red) and the central lobe (lobe1, blue); domain 2 with the  $\alpha$ -helical domain insertion (HD2, red), the central lobe (lobe2, blue) and the C-terminal domain (CTD, yellow). Positions of structural domains are indicated along the zebrafish Rad54 protein sequence (Table 4.1). Serine residue 566 (Ser566) is highlighted that is equivalent to Ser572 in human Rad54. The structures were rendered using VMD [93].

very similar to the helicases of superfamily 2 (SF2). Nevertheless, the presence of specific insertions found exclusively in SWI/SNF2 enzymes distinguish Rad54 from classical helicases. Several structural features are defined:

1. The **N-terminal domain (NTD)** consists of a three-stranded  $\beta$ -sheet stabilized by three small  $\alpha$ -helices. By packing to the first lobe via extensive hydrophobic interactions this domain contributes to the overall structural stability of the protein. The presence of several solvent-exposed hydrophobic residues suggests a protein-protein binding interface.
2. The central structure consists of 2 RecA-like  $\alpha/\beta$ -domain **lobes** commonly found in helicases that show the seven typical SF2 helicase signature motifs of an ATP-dependent motor protein. Each lobe consists of a  $\beta$ -sheet flanked by  $\alpha$ -helices. The two lobes are connected via an  $\alpha$ -helical **linker**.

3. Each RecA-like domain exhibits  **$\alpha$ -helical domains (HD)**. In contrast to other helicase structures these insertions are characteristic for SWI/SNF2 family proteins and are most probably involved in chromatin remodeling activity. In the first lobe the 55-residue insertion (**HD1**) shows three  $\alpha$ -helices in a kinked 'V'-like arrangement. In the second lobe the 125-residue domain (**HD2**) is constituted by two insertions that fold into a six  $\alpha$ -helical structure with a hydrophobic core and solvent-exposed positively charged residues. HD2 is highly-conserved and probably involved in DNA-binding. Mutations in HD2 are attributed to non-Hodgkin's lymphomas and X-linked mental retardation [179].
4. The inter-domain region flanked by HD1 and HD2 is referred to as the DNA-binding **cleft**. It can harbor approximately 17 bp of dsDNA. It is supposed that opening and closing of the cleft result in Rad54 translocation along a dsDNA strand [187].
5. The **C-terminal domain (CTD)** has a zinc-stabilized  $\alpha/\beta$ -structure extending the  $\beta$ -sheet of the second lobe by two additional  $\beta$ -strands. The CTD exhibits a positively-charged surface patch and thereby elongates the inter-lobe cleft.



**Figure 4.2:** Rad54 crystal structure of the archaea *Sulfolobus solfataricus* in complex with dsDNA (blue) and without dsDNA (yellow) shown in new cartoon representation. The dsDNA is colored black. Both crystal structures were analyzed at a resolution of 3 Å (PDB codes: 1Z6A and 1Z63, respectively) [179]. A) top view and B) side view along the DNA axis. The structures were rendered using VMD.

In addition, two crystal structures of the putative Rad54 homolog of the archaea *Sulfolobus solfataricus* in complex with and without dsDNA are available at a resolution of 3 Å (PDB codes: 1Z6A and 1Z63, respectively) [45]. The structures reveal the same architecture of the core domain as it is observed in the zebrafish Rad54 structure. Surprisingly, superposition of DNA bound (blue) and DNA-unbound (yellow) crystal structures show only minor differences (Figure 4.2).

## 4.2 Theory: Linear Response Theory

### Anisotropic Network Model

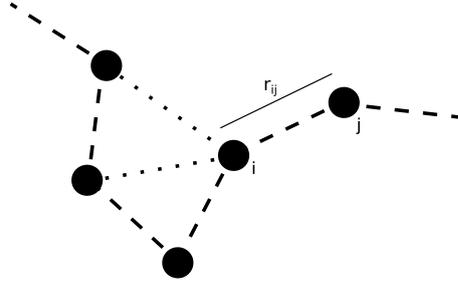
Proteins constantly undergo dynamical changes that are related to their biological functions. Global structural changes can be derived from normal mode analysis (NMA) [22] where normal modes can be computed analytically using all-atomic force fields (Section 2.4.3). Each normal mode characterizes a path on the free energy surface that represents collective conformational changes. It was shown that the lowest frequency modes often refer to biological functions as their associated conformational change require only a small amount of energy for the transition [8]. Since a rigorous energy-minimized structure is required for NMA, this technique is computationally very intensive.

As an approximation, Tirion suggested to model the atom-atom interactions by single parameter Hookean harmonic potentials [180]. Since results from this approach and NMA are highly related, this simplified model was shown to sufficiently capture the fluctuation of the slow modes. In order to further reduce computational costs, coarse-grained elastic network models were introduced [9]. Here, an amino acid residue is typically represented by its C<sup>α</sup> atom only. If two beads are located within a certain distance range defined by the cutoff  $r_c$ , they are assumed to be in contact and connected by harmonic springs (Figure 4.3).

Anisotropic network models (ANMs) are a special type of elastic network models that in addition to magnitudes, take the directional information of the residue fluctuations into account [7]. In an extended ANM, different harmonic potentials are assigned for bonded and non-bonded interactions [74]. Here, the overall potential  $V$  of the protein system is defined as

$$V = \alpha a^{-2} \left[ \frac{a^2 K}{2} \sum_i (r_{i,i+1} - r_{i,i+1}^o)^2 + \sum_{(i,j) \in I} \kappa (s_{ij} - s_{ij}^o)^2 \right] \quad (4.1)$$

where  $K$  is the spring constant for covalent interactions,  $r_{i,i+1}$  being the distance between a C<sup>α</sup> atom  $i$  and its covalently bound partner atom  $i + 1$  in the current state and  $r_{i,i+1}^o$  the respective distance in the native state. Here,  $I$  is the set of non-covalent interactions defined by those atoms whose distances are smaller than a certain distance cutoff  $r_c$ . For those non-covalent interactions,  $\kappa$  is the spring constant,  $s_{ij}$  is the distance between two neighbored C<sup>α</sup> atoms  $i$  and  $j$  that are not



**Figure 4.3:** Reduced protein system used in elastic network model. Amino acid residues are centered to their  $C^\alpha$  atoms and represented as beads. Two beads  $i$  and  $j$  are in contact if their distance  $r_{ij}$  is below a certain distance cutoff  $r_c$ . Their interaction is modeled by a harmonic spring where different Hookean potentials are assigned for covalent (dashed line) and non-covalent interactions (dotted line) [74].

linked via a covalent bond in the current state, and  $s_{ij}^o$  the respective distance in the native state. The term  $\alpha a^{-2}$  is an overall scaling factor.

Assuming that a given structure is in equilibrium, the  $3N \times 3N$  Hessian matrix  $\mathbf{H}$  can be derived analytically where  $N$  is the number of residues in the protein. The matrix  $\mathbf{H}$  is then composed of  $3 \times 3$  super-elements

$$\mathbf{H}_{ij} = \begin{bmatrix} \partial^2 V / \partial x_i \partial x_j & \partial^2 V / \partial x_i \partial y_j & \partial^2 V / \partial x_i \partial z_j \\ \partial^2 V / \partial y_i \partial x_j & \partial^2 V / \partial y_i \partial y_j & \partial^2 V / \partial y_i \partial z_j \\ \partial^2 V / \partial z_i \partial x_j & \partial^2 V / \partial z_i \partial y_j & \partial^2 V / \partial z_i \partial z_j \end{bmatrix} \quad (4.2)$$

where  $x_i$ ,  $y_i$  and  $z_i$  are the Cartesian coordinates that define the position of atom  $i$ . Super-elements  $\mathbf{H}_{ij}$  are calculated by the second order partial derivation of the potential  $V$  with respect to the  $C^\alpha$  atom positions. The Hessian matrix  $\mathbf{H}$  has a sixfold symmetry with six eigenvalues being zero and thus six degenerated zero eigenvalues. Therefore, the matrix is singular and not invertible. Instead, a Moore-Penrose pseudo-inverse  $\mathbf{H}^{-1}$  of  $\mathbf{H}$  can be computed [136]. Assuming an equilibrated structure at a constant temperature  $T$ , the  $3N \times 3N$  mechanical covariance matrix  $\mathbf{C}$  is defined as

$$\mathbf{C} = k_B T \cdot \mathbf{H}^{-1} \quad (4.3)$$

where  $k_B$  is the Boltzmann constant. The off-diagonal super-elements  $\mathbf{C}_{ij}(i \neq j)$  describe the cross-correlations between the  $x$ ,  $y$  and  $z$  components of atoms  $i$  and  $j$  whereas the diagonal super-elements  $\mathbf{C}_{ij}(i = j)$  describe the self-correlations between the components of atom  $i$  [7].

### Linear Response Theory

In linear response theory (LRT) a protein's conformational change can be predicted by assuming a linear relationship between the equilibrium fluctuations in the unper-

turbed state and its perturbed state [94]. The perturbation is modeled by a force that either acts on one  $C^\alpha$  atom or on a set of multiple  $C^\alpha$  atoms. Thus, attractive or repulsive forces can be defined towards a certain direction in the Cartesian coordinate space. The  $3N$  external force vector  $\mathbf{f}_j$  acting on atom  $j$  is defined by the perturbation point  $Q$  and the external point  $P$

$$\mathbf{f}_j = Q - P \quad (4.4)$$

where the sign defines whether the force is attractive (positive sign) or repulsive (negative sign). For the perturbation of a number of  $j$   $C^\alpha$  atoms,  $F$  different force vectors  $\mathbf{f}_j$  are created. The LRT relation is given by

$$\Delta \mathbf{r} \simeq \beta \cdot \sum_{j \in F} \mathbf{C} \cdot \mathbf{f}_j \quad (4.5)$$

where  $\Delta \mathbf{r}$  is the expected coordinate shift of  $N$  atoms and  $\mathbf{C}$  is the covariance matrix of the unperturbed state. Here, the covariance matrix  $\mathbf{C}$  can be derived either from MD simulations or ANMs.

The expected coordinate shift  $\Delta \mathbf{r}$  is added to the initial atoms' position of the unperturbed state to calculate the effective atom displacement. This predicted displacement is proportional to the force strength applied, so scaling factor  $\beta$  can be adjusted to obtain different intermediate steps of the linear displacement for dynamic visualization purposes. Since MD simulations are computationally expensive, the use of coarse-grained models was preferred and thus the covariance matrix was derived from ANMs.

## 4.3 Methods

### Multiple Sequence Alignment

The amino acid sequences of human, zebrafish and archaea Rad54 protein were retrieved from the NCBI database of non-redundant (nr) protein sequences (Table 4.1) [156]. Multiple sequence alignments were created using ClustalW 2.1 [117] with the BLOSUM62 [82] as substitution matrix. The alignments were visually inspected in JalView 2.0 [194].

**Table 4.1:** Details of the Rad54 protein sequences of various organisms. Sequences can be retrieved from the NCBI protein database via the NCBI accession number. hyp: hypothetical protein.

organism name	NCBI accession no.	length [aa]
Human <i>Homo sapiens</i>	CAA66379.1	747
Zebrafish <i>Danio rerio</i>	NP_957438.1	738
Archea <i>Sulfolobus solfataricus P2</i>	NP_343078.1	906

### Anisotropic Network Model

Cartesian coordinates from the Rad54 crystal structures of zebrafish *Danio rerio* (PDB code: 1Z3I, chain X) [179] and archaea *Sulfolobus solfataricus* (PDB code: 1Z6A, chain A, dsDNA-unbound) [45] are used for computing the contact map. This binary matrix specifies the contacts between two amino acid residues  $i$  and  $j$  with respect to the spatial distance of their C $^\alpha$  atoms observed in the crystal structure. Various distance cutoffs (6-15 Å) were tested and the same overall displacement pattern was observed with a difference in scaling only (data not shown). Therefore, a general distance cutoff of 13 Å was used in all ANM experiments.

A homogeneous spring potential was applied where no distinction is made between the 20 canonical amino acid residues. It has been shown previously that no significant difference exists between heterogeneous and homogeneous inter-residue potentials [84]. For all non-covalent residue-residue contacts, the average of the Miyazawa-Jernigan (MJ) [132] interaction potential  $\bar{\gamma}_{MJ} = 3.166 RT/\text{Å}^2$  was taken. For covalent interactions between neighbored residues, an interaction strength of  $82 RT/\text{Å}^2$  is used as it was suggested before [74]. The covariance matrix was computed using the BioPhysConnectoR package [85] in R [158].

### Linear Response Theory – Basic Setup

In the zebrafish Rad54 crystal structure, Ser566 is equivalent to Ser572 in human Rad54. In LRT I three different perturbation scenarios are conducted to mimic the Ser566 phosphorylation reaction: 1) an attractive force acting upon Ser566 towards an external point outside the protein, 2) attractive forces acting on the two positively-charged side chains Arg536, Lys568 located in close proximity of the Ser566 (Figure 4.5) towards its O $^\gamma$  atom and 3) a combination of attractive and repulsive forces acting on seven neighbored residues found within a distance of 8 Å of Ser566 towards Ser566 O $^\gamma$  atom. In LRT III, an attractive force acting upon Ser806 in the archaea structure towards an external point outside the protein was applied. An overview of LRT setups is given in Table 4.2.

The linear shift is visualized in VMD by drawing arrows pointing from the initial atom location to the final state after LRT.

### Linear Response Theory – Null Model

We developed a LRT null model to investigate the influence of the force direction and the ANM spring constants on the observed displacement. For this purpose, the perturbation of one single C $^\alpha$  atom (Ser566 in zebrafish and Ser806 in archaea) is repeated in 1000 independent LRT runs with multiple force vectors pointing from the C $^\alpha$  atom to 1000 different external points. They were obtained by randomly placing points on a surface of a sphere surrounding the C $^\alpha$  atom. In order to ensure an isotropic distribution points, we take spherical coordinates  $\theta$  and  $\phi$  being uniformly distributed with  $\theta \in [0, 2\pi]$  and  $\phi \in [-1, 1]$  to generate a point  $P = (P_x, P_y, P_z)$  with

**Table 4.2:** LRT setups applied to the Rad54 zebrafish (PDB code: 1Z3I) and archaea (PDB code: 1Z6A) structure by perturbing different atoms. On each atom, forces are attractive (A) or repulsive (R) directing to an arbitrarily chosen point  $P$  being located externally (ext) or at the  $O^{\gamma}$  atom (og) of the Ser566 side chain in the full-atomic crystal structure.

run	PDB code	residues perturbed	force type	direction
LRT I	1Z3I	Ser566	A	ext
LRT I	1Z3I	Asn536, Lys568	A, A	og
LRT I	1Z3I	Asn536, Lys537, Asp538 (upper $\beta$ -sheet)	A, R, R	og
		Ser514, Asn515, Thr517 (side loop)	R, R, R	og
		Lys568	A	og
LRT III	1Z6A	Ser806	A	ext

$$P_x = \sqrt{(1 - \phi^2) \cdot \cos(\theta) \cdot r} \quad (4.6a)$$

$$P_y = \sqrt{(1 - \phi^2) \cdot \sin(\theta) \cdot r} \quad (4.6b)$$

$$P_z = \phi \cdot r \quad (4.6c)$$

where  $P_x$ ,  $P_y$  and  $P_z$  are the three-dimensional coordinates of point  $P$  and  $r$  the radius of the sphere. The sphere radius was set to 1 as scales are gauged by the parameter  $\beta$  in Equation 4.5. Additionally, for each run a new covariance matrix is obtained with random spring constants  $\gamma$  for non-covalent interactions. The spring constants  $\gamma$  are uniformly distributed around the mean MJ potential  $\bar{\gamma}_{MJ} = 3.166 RT/\text{\AA}^2$  with  $\gamma \in [0, 2 \cdot \bar{\gamma}_{MJ}]$ . For each  $C^\alpha$  atom a set of 1000 expected coordinate shifts is yielded. For shift visualization, a structure file was produced with the average shifted coordinates and colored according to the standard deviation.

### Linear Response Theory – Extended Null Model

In order to check if a displacement observed from LRT perturbation is characteristic for the perturbed atom  $j$  exclusively (in the following referred to as the reference shift of the reference atom), it is compared to the expected displacement resulting from perturbation of all other residues  $i$ . Here, the reference shift  $\langle \Delta \mathbf{r}^{\text{ref}} \rangle$  is defined as the mean expected coordinate shift

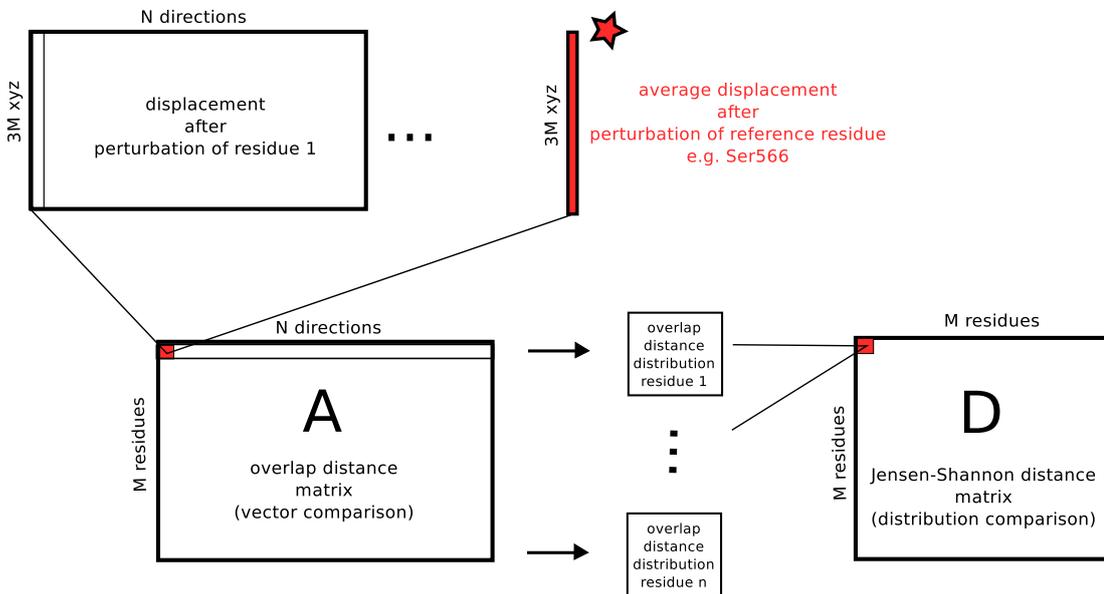
$$\langle \Delta \mathbf{r}^{\text{ref}} \rangle = \frac{1}{N} \sum_{j \in N} \Delta \mathbf{r}_j^{\text{ref}} \quad (4.7)$$

after applying the null model previously described to the reference atoms Ser566 and Ser806 for the zebrafish and archaea structure, respectively.

In addition, the null model is applied to every  $C^\alpha$  atom  $i \in \{1 \dots M\}$  found in the structure and perturbed from 1000 different directions  $j \in \{1 \dots N\}$  leading to the  $M \times N$  matrix  $\mathbf{A}$  (Figure 4.4). Each of the  $N$  shifts were compared to the reference shift given by

$$\mathbf{A}_{ij} = \arccos \frac{\Delta \mathbf{r}_j^i \cdot \langle \Delta \mathbf{r}^{\text{ref}} \rangle}{|\Delta \mathbf{r}_j^i| \cdot |\langle \Delta \mathbf{r}^{\text{ref}} \rangle|} \quad (4.8)$$

that is similar to the overlap distance [84]. Prior to the calculation, those positions belonging to the Cartesian coordinates of the current  $C^\alpha$  atom  $i$  and the reference atom are set to zero in both, the coordinate shift vector  $\Delta \mathbf{r}_j^i$  and the reference shift vector  $\langle \Delta \mathbf{r}^{\text{ref}} \rangle$  to neglect the displacement bias of the perturbed atoms. Thus, quasi  $3N - 6$  vectors are obtained.



**Figure 4.4:** Overview of the extended LRT null model: For each  $C^\alpha$  atom, 1000 independent runs with different force directions are performed. Each of the 1000 resulting  $3M$  displacement vectors is compared to the reference displacement vector by using the overlap distance yielding a  $M \times N$  matrix  $\mathbf{A}$ . For each  $C^\alpha$  atom a distribution of 1000 overlap distances is observed. The distributions of all  $C^\alpha$  atoms are compared to each other by using the Jensen-Shannon distance yielding a symmetric  $M \times M$  matrix  $\mathbf{D}$ .

For each  $C^\alpha$  atom  $i$ , a unique distribution of 1000 overlap distances is obtained that describe the similarity to the reference shift. In a subsequent step, all  $M$  distributions are compared to each other by using the Jensen-Shannon divergence

$$D_{\text{JS}}(P, Q) = \frac{1}{2}(D_{\text{KL}}(P, R) + D_{\text{KL}}(Q, R)) \quad (4.9a)$$

$$R = \frac{1}{2}(P + Q) \quad (4.9b)$$

where  $D_{\text{JS}}$  measures the divergence between two distributions  $P$  and  $Q$  by calculating the Kullback-Leibler divergence  $D_{\text{KL}}$  with respect to their joint distribution  $R$  ( $D_{\text{KL}}$  was described earlier in Section 3.2). Here, the square root of  $D_{\text{JS}}$  is taken to obtain a true metric referred to as Jensen-Shannon distance  $\widetilde{D}_{\text{JS}}$  [47]. A symmetric  $M \times M$  matrix  $\mathbf{D}$  of  $\widetilde{D}_{\text{JS}}$  values is obtained.

A hierarchical clustering using the distance information was applied with the complete-linkage clustering method. The matrix rows and columns were reordered and  $\text{C}^\alpha$  atoms were clustered into groups and visualized by a dendrogram. Applying a cutoff to the dendrogram at  $\widetilde{D}_{\text{JS}} = 1.5$  resulted in formation of six distinct residue groups.  $\text{C}^\alpha$  atoms were colored according to their groups and visualized as colored surface patches in VMD [93].

### Linear Response Theory – Inverse Relation

According to Equation 4.5 it is also possible to calculate the expected forces acting upon all the  $\text{C}^\alpha$  atoms from a coordinate shift observed e.g. in two crystal structures.

A structural fit of Rad54 structures without dsDNA (PDB code: 1Z6A, chain A) and in complex with dsDNA (PDB code: 1Z63, chain A) was done over a set of 466  $\text{C}^\alpha$  atoms present in both structures (residue 432-809, 811-829 and 835-903) and the coordinate shift  $\Delta\mathbf{r}$  was calculated by subtracting the coordinates of the dsDNA-bound structure from those of the dsDNA-unbound structure. The expected force vector  $\mathbf{f}$  is calculated as follows

$$\mathbf{f} \simeq \frac{1}{\beta} \cdot \mathbf{H} \cdot \Delta\mathbf{r} \quad (4.10)$$

where the Hessian  $\mathbf{H}$  is derived as the Moore-Penrose pseudoinverse from the covariance matrix  $\mathbf{C}$  computed from an ANM of the dsDNA-unbound structure, homogeneous parameterization and a distance cutoff of 13 Å. The scaling factor  $\beta$  was set to 1. Those residues showing a force above a threshold of 200 were selected.

### Structural Modeling of the Open-State Zebrafish Structure

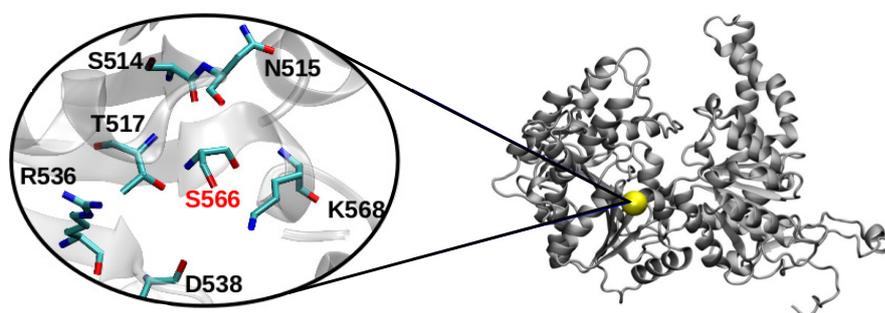
The putative *open-state* zebrafish Rad54 structure was manually prepared based on the template of the *open-state* archaea structure (PDB code: 1Z6A, chain A). A structural fit was done by superposing the structures over the  $\text{C}^\alpha$  atoms belonging to the central  $\beta$ -sheets of lobe 1 and lobe 2 separately. The domains were separated and reconnected by maintaining the 11-residue linker from 1Z6A structure (residue

652-662). A short energy minimization was done in GROMACS 4.5.1 [155] using the Amber03 force field [43] to relax the structure and eliminate possible clashes.

## 4.4 Results & Discussion

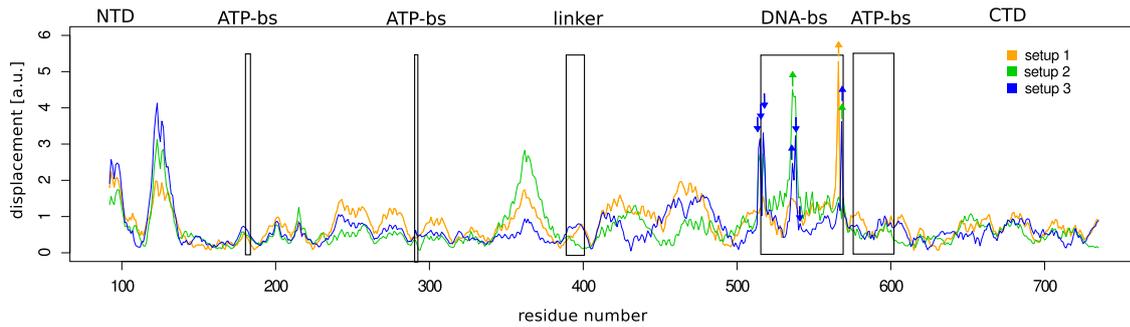
### LRT I: Perturbing Ser566 in Zebrafish Rad54

Due to the close relation to human Rad54, the first step is to investigate the LRT structural response of zebrafish Rad54. Multiple sequence alignments revealed Ser566 to be equivalent to Ser572 in human Rad54 (appendix Figure A.8).



**Figure 4.5:** Location of Ser566 in zebrafish Rad54 structure (PDB code: 1Z3I) visualized in new cartoon representation (gray). Neighbored amino acid residue side chains located in a  $\beta$ -sheet (Asn536, Lys537, Asp538), loop (Ser514, Asn515, Thr517) and an  $\alpha$ -helical region (Lys568) are shown as sticks.

Three different perturbation scenarios were employed and the expected atom displacements are visualized in Figure 4.6. As expected, the extended N-terminal region shows a large displacement in contrast to the C-terminal region that is tightly packed against domain 2 (Figure 4.1). An additional peak appears at residues 350-380 being part of the HD1 domain, the remodeling-specific insertion in domain 1. In general, most regions show a similar displacement in all three setups. A stronger variation is observed around residue positions 400-500. This is the region of the HD2  $\alpha$ -helical insertion of domain 2 where the displacement patterns are setup-dependent. The flexibility of the insertions HD1 and HD2 might be attributed due to their suggested role in temporal dsDNA-binding during the translocation cycle [179]. None of the three regions involved in ATP-binding reveal an extraordinary shift. The dominant peaks visible in the dsDNA-binding cleft around positions 500-600 result from the forces applied to Ser566 which effect the displacement of this region dramatically.



**Figure 4.6:** LRT I: Expected displacement of zebrafish Rad54 C $\alpha$  atoms of one single LRT run using different perturbation scenarios: setup 1) Ser566 (orange), setup 2) Arg536, Lys568 (green) and setup 3) Ser514, Asn515, Thr517, Asn536, Lys537, Asp538, Lys568 (blue). For details, see Table 4.2. Perturbed atoms are marked by an arrow in corresponding colors. Several structural regions are indicated: ATP-binding site (ATP-bs), DNA-binding site (DNA-bs) and the linker region connecting the two domains. Note that the absolute scaling is arbitrary.

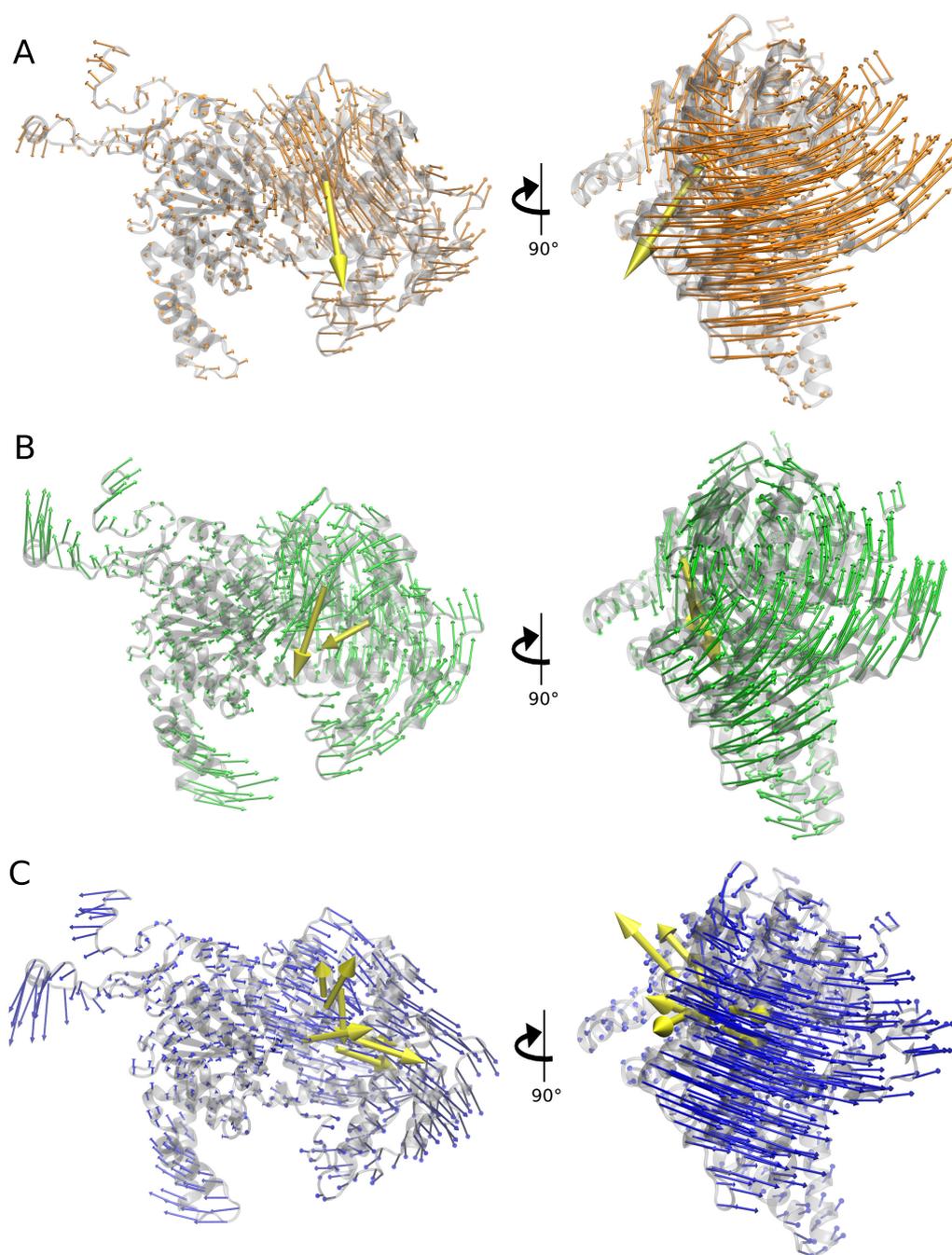
Direct visualization of the collective motions on the structure (Figure 4.7) reveal a significant shift of domain 2 in comparison to domain 1 observed for all three setups. Note that for visualization purposes the displaced structure was fitted to the initial structure by the first domain. A rotational movement of the entire domain 2 is indicated. Still, the orientation of the rotational axes varies among the three setups.

### Null Model of LRT I

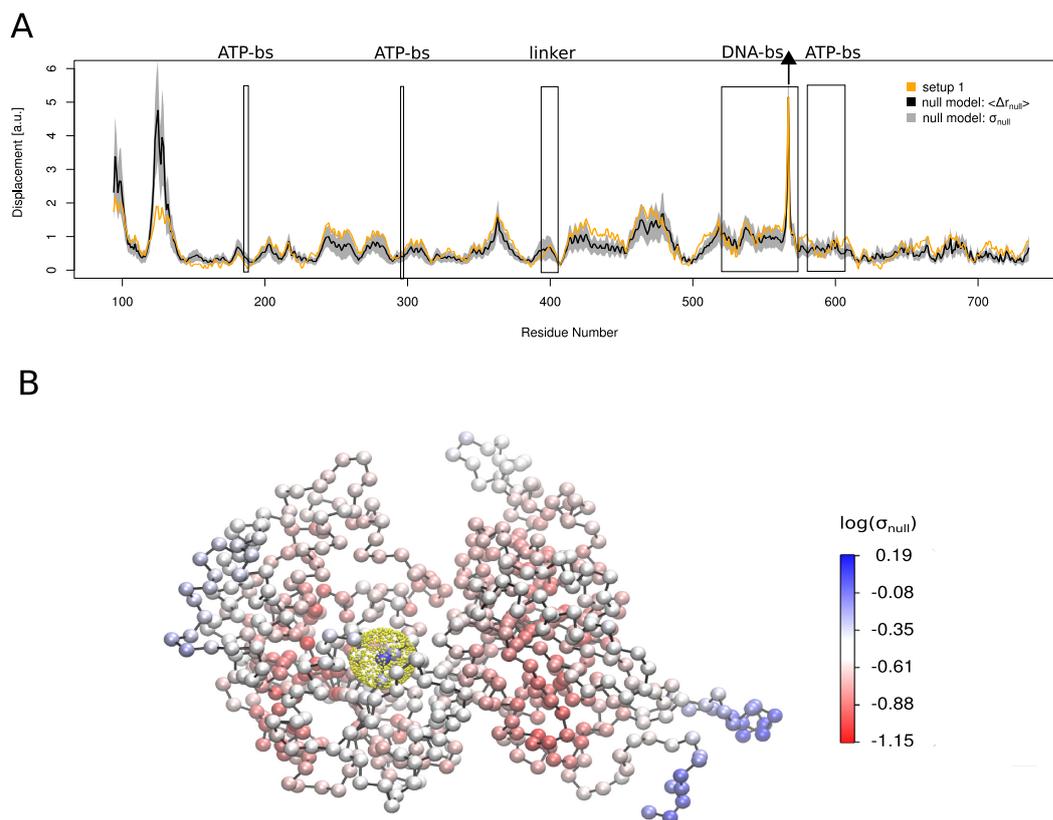
In order to inspect the influence of the force direction on the resulting displacement, we conducted 1000 independent LRT runs perturbing the Ser566 C $\alpha$  atom from random directions as described in Section 4.3. Additionally, random ANM spring constants were applied to check the possible influence.

The average displacement of 1000 runs together with the corresponding standard variation is shown for each C $\alpha$  atom (Figure 4.8A). Compared to the single run for Ser566 perturbation using an arbitrary force direction, results show that this is sufficient to capture the displacement pattern observed from 1000 runs. As a consequence, this indicates that the predicted displacement is rather independent of the force direction. The correlation between the average shift of 1000 runs with additionally varied spring constants to that of non-varied ones is relatively high (Pearson correlation coefficient  $r = 0.926$  with  $p < 2.2 \cdot 10^{-16}$ , Section 2.2.3). Correspondingly, this demonstrates that the spring constants do not have a relevant influence on the displacement either.

With respect to the structure, the average displacement together with the colored standard deviation obtained from 1000 random runs are shown in Figure 4.8B. Note that the displaced structure was not fitted to domain 1. As expected, the central



**Figure 4.7:** LRT I: Visualization of expected displacement for A) setup 1 (orange), B) setup 2 (green) and C) setup 3 (blue) according to Figure 4.6. Coordinate shifts are visualized by small arrows pointing from the C $\alpha$  atom's initial atom positions to the final after LRT using force scaling factor of 100. Force vectors of perturbed atoms are indicated by big yellow arrows. The structure of zebrafish Rad54 (PDB code: 1Z3I) is shown in new cartoon representation (gray) in the background, front view (left panels) and side view (right panels). Note that the shifted structure was fitted to the initial by superimposing domain 1 for visualization purposes. Structures were rendered using VMD.



**Figure 4.8:** LRT null model. A) Expected displacement of the zebrafish Rad54  $C^\alpha$  atoms after perturbing atom Ser566 (black arrow). One single LRT run applying an arbitrary force direction using homogeneous ANM spring parameterization (orange). Average (black) displacement with standard deviation range (gray) of 1000 independent LRT runs from random directions and with random ANM spring constants. Note that the scaling of the displacement is arbitrary. B) Average coordinates of  $C^\alpha$  atoms are shown from the independent 1000 runs, colored by the logarithm of standard deviation. Force directions are indicated by yellow arrows. Plot was done using the R-package ggplot2 [199] and structure was rendered in VMD.

$\beta$ -sheets of the two lobes show a relatively low deviation whereas the N-terminus show a high deviation.

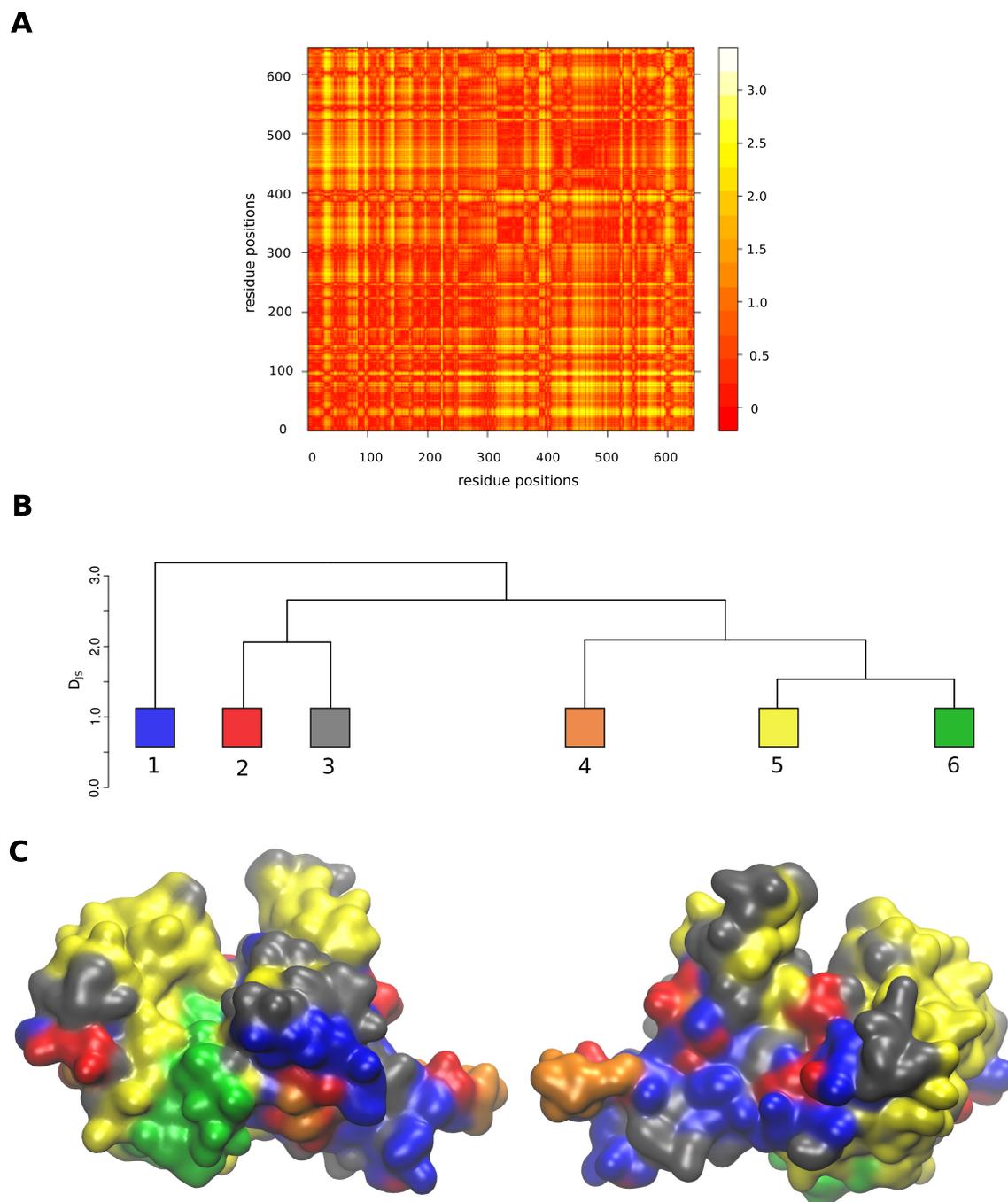
### Extended Null Model of LRT I

The displacement pattern observed upon Ser566 perturbation appears to be robust. In order to verify if this pattern is specific for Ser566, an extended null model approach is applied as described in Section 4.3. Figure 4.9A shows the Jensen-Shannon distance matrix between all overlap distributions. Based on this distance information the hierarchical clustering for residues was computed in Figure 4.9B.

The cutoff  $D_{JS} = 1.5$  was chosen to yield around six distinct residue groups for visualization at the zebrafish Rad54 structure (Figure 4.9C).

Considering the fact that the groups are not arbitrarily distributed over the structure but instead define localized regions (the so-called *response patches*) indicates that response patterns are characteristic for certain structural regions. Residues being clustered together in one group exhibit a response pattern similar to that of the Ser566 reference shift. Ser566 itself is found in group 6 (green patch). Residues belonging to that group are located in the lower  $\beta$ -sheet of domain 2. Group 5 residues show also a similar response pattern and are located within the helical insertions HD1 and HD2. Interestingly, residues of the N-terminal region (group 4) are clustered together with group 5/6 residues but nevertheless respond differently. The response patterns of group 2/3 residues are highly related to each other but more distant to the reference shift. Further, group 2/3 residues are localized in various smaller patches (red) and are distributed over the entire molecule. Group 1 residues are located in domain 1 and show the response pattern being most distant to all other groups.

Taken together, overall results of LRT I experiments indicate that the response patterns are robust and relatively independent from the atom sets perturbed, force directions and ANM-spring constants. Still, results do not resolve the question whether structural changes upon Ser566 phosphorylation facilitate dsDNA or ATP-binding. Biochemical experiments suggest that ATP-hydrolysis is induced upon dsDNA-binding [45]. Therefore, the next question we address is, if Ser566 phosphorylation might activate dsDNA-binding.



**Figure 4.9:** Results of the LRT extended null model. A) Jensen-Shannon distance matrix  $D$  comparing the overlap distance histograms of each residue. Values range from  $D_{JS}$  0-51 (red to yellow). B) Dendrogram obtained after hierarchical clustering of rows and columns of matrix  $D$ . Applying a similarity cutoff  $D_{JS} = 1.5$  leads to six distinct residue groups: 1 (blue), 2 (red), 3 (gray), 4 (orange), 5 (yellow) and 6 (green). C) Surface visualization of residue groups at the zebrafish Rad54 structure, front view (bottom left) and back view (bottom right) reveal distinct response patches. Structures were rendered in VMD.

## LRT II: Comparing the Archaea Rad54 Structures

Comparing zebrafish and archaea Rad54 crystal structures by superpositioning the domains 2 reveals the zebrafish Ser566 to be in close proximity of Ser806 in the archaea structure. This finding confirms the results from a previous multiple sequence alignment (appendix Figure A.8). In the conformational state of the archaea crystal structure, Ser806 is exposed on the protein surface (see red sphere in Figure 4.10). This suggests that this might be the state in which Ser806 could be phosphorylated and thereby might enable dsDNA-binding. This step is also supposed to occur before ATP-binding.

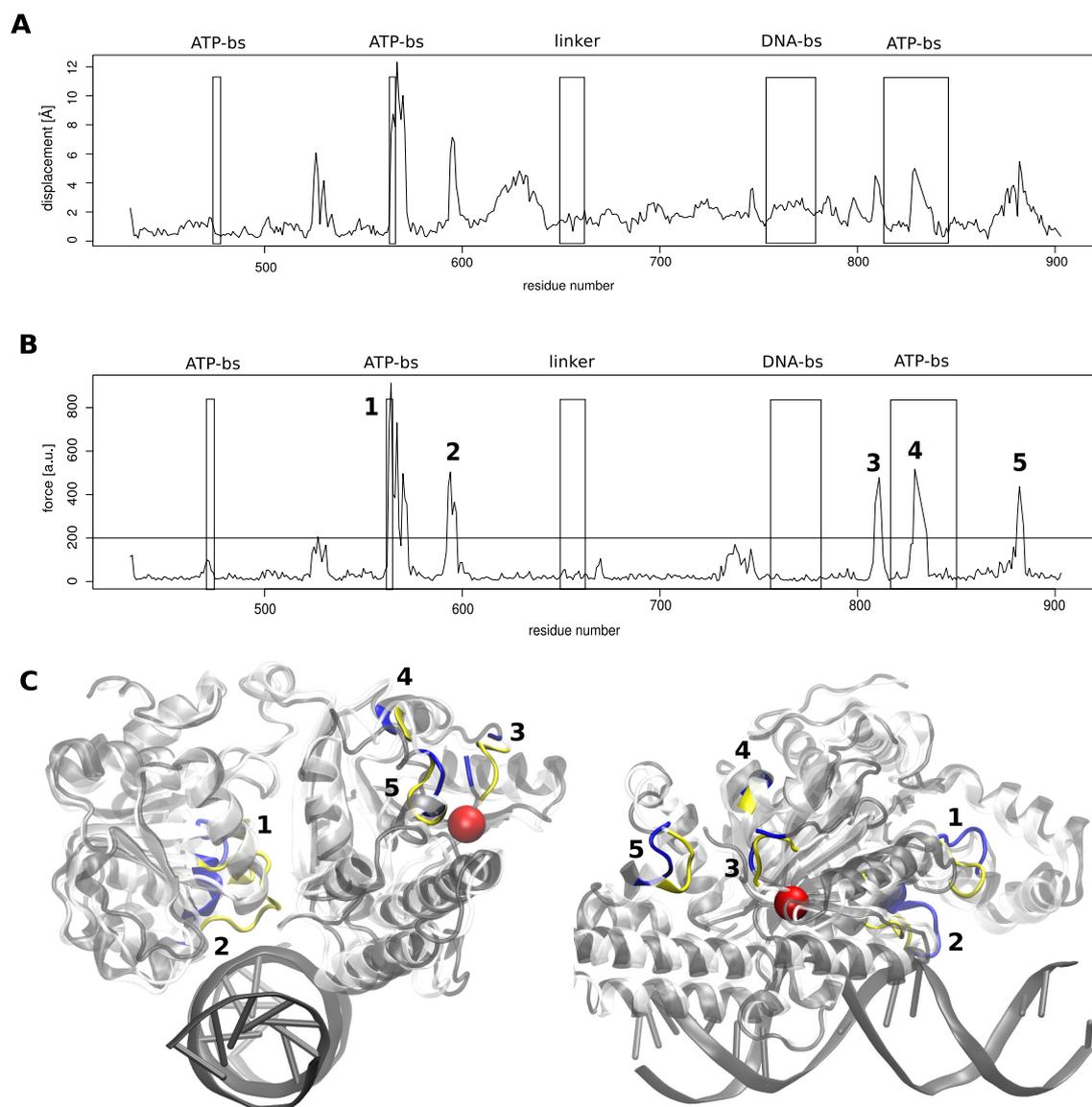
To address this question, the focus is shifted away from the zebrafish structure (ATP-bound state) towards the two archaea structures (ATP-unbound state). The purpose is to investigate the mechanisms of possible dsDNA-binding induction upon Ser806 phosphorylation. By using the inverse LRT equation, the expected forces  $f$  are derived from the coordinate shifts  $\Delta r$  calculated by comparing the archaea dsDNA-unbound and the dsDNA-bound state (Figure 4.10A). The expected forces which are supposed to be responsible to transform the dsDNA-unbound into the dsDNA-bound state are shown in Figure 4.10B.

Structural regions with high resulting forces are selected and visualized on the archaea DNA-unbound crystal structure (Figure 4.10C). Five dominant regions 1-5 are distinguished (Table 4.3). Apparently, all residues with a predicted high force simultaneously show a large displacement. In contrast, regions with high observed displacements not necessarily reveal a high predicted force.

LRT I results have shown that regions with applied forces are also associated with high shifts (Figure 4.6). Most notably, selected regions exhibit loops connecting domains and, not surprisingly, they are the most flexible regions. All in all, from these results it is not possible to conclude that Ser806 might be implicated in dsDNA-binding activation.

**Table 4.3:** Structural regions where expected high forces act upon supposed to transform the dsDNA-unbound into the dsDNA-bound state. Forces are calculated using inverse LRT in Figure 4.10B and C.

region	residues	description
1	563-572	loop region, located in DNA minor groove binding
2	593-597	loop region, connecting the N-terminus to the first lobe of domain 1
3	809-812	loop region, in the vicinity of the predicted phosphorylation site Ser806
4	835-836	an $\alpha$ -helix
5	882-885	loop region, connecting the C-terminus to the lobe of domain 2



**Figure 4.10:** LRT II: A) Structural shift calculated between Rad54 conformations observed in dsDNA-bound and dsDNA-unbound crystal structures. B) Expected forces acting upon several  $C^\alpha$  atoms of the archaea dsDNA-unbound structure using the inverse LRT relation of Equation 4.10. A cutoff of 200 is applied to select residue regions with high predicted forces. Structural overlay of the archaea dsDNA-bound (dark gray, PDB code: 1Z63) and dsDNA-unbound crystal structure (light gray, PDB code: 1Z6A) showing residues where high expected forces act upon are colored blue (dsDNA-bound) and yellow (dsDNA-unbound). The location of Ser806 is indicated by a red sphere. Structures were rendered in VMD.

### LRT III: Perturbing Ser806 in Archaea Rad54

In order to further investigate a possible Ser806 implication in dsDNA-binding, the LRT response in the dsDNA-unbound archaea structure is analyzed (Table 4.2). Figure 4.11 shows the expected displacement after perturbing Ser806. Several peaks in the second domain are observed. The visualization of the expected displacement at the structure reveals a drastic rotational movement of domain 2. However, this displacement does not resemble a transition towards the dsDNA-bound form. It rather appears to be a global general dynamics that is observed upon perturbation of several different atoms similar to the general dynamics observed in the zebrafish in LRT I (4.7). However, this 180° turn of domain 2 has been proposed in the literature as being part of the translocation cycle [45].

The question arises how the dsDNA-bound state would respond to Ser806 perturbation in contrast to the dsDNA-unbound form. The results are expected to be similar, because the structural deviation is low (RMSD = 2.27 Å) and the C<sup>α</sup> atom contact configuration is very similar: Pearson's correlation coefficient between the contact maps is  $r = 0.923$  with  $p < 2.2 \cdot 10^{-16}$ .

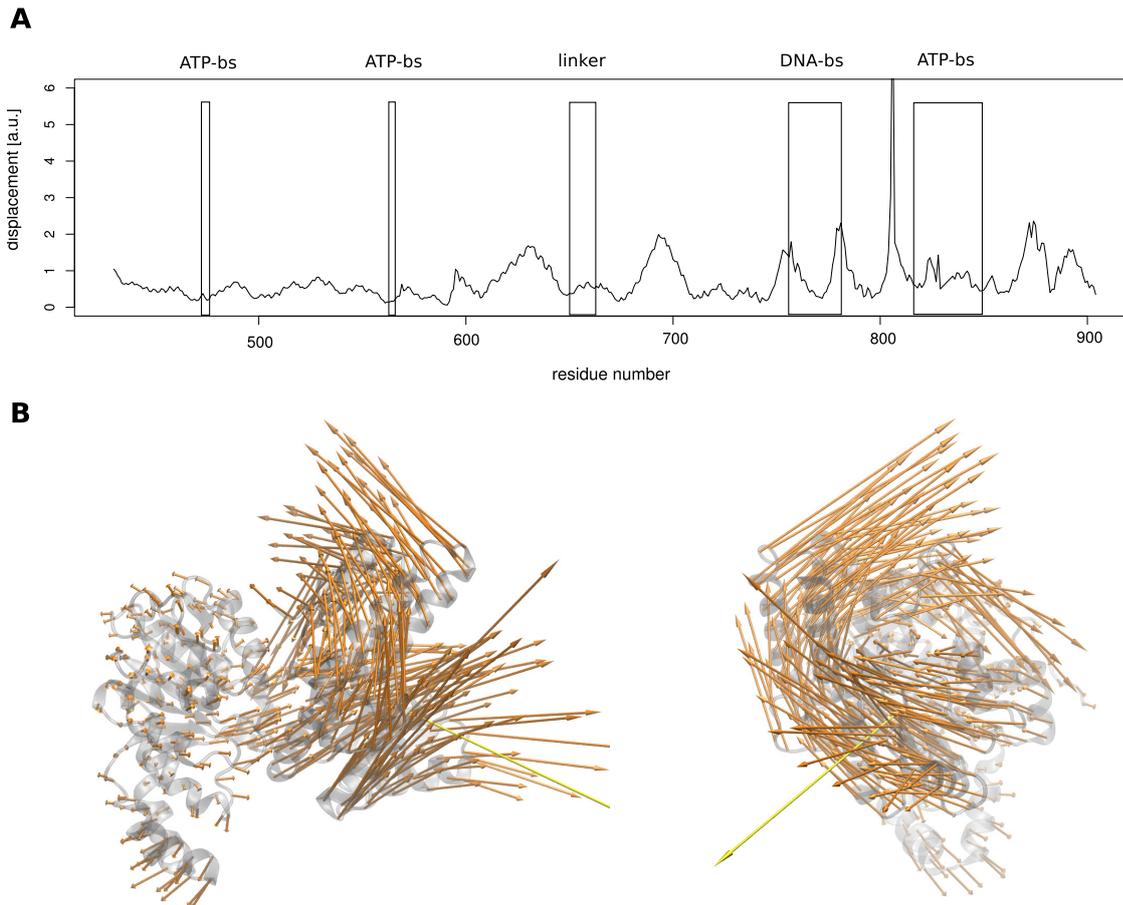
### Results in Context of the Rad54 Translocation Cycle

Although detailed insights of the Ser566 phosphorylation induced structural changes could not be obtained in this study, the LRT results contribute to the elucidation of the Rad54 translocation cycle.

Perturbing the zebrafish structure (PDB code: 1Z3I) leads to a rather homogeneous LRT response: the domain 2 reveals a drastic conformational change relative to the first domain, namely a 180° turn as it was hypothesized before [45]. Since the change is robust and rather independent of the perturbation site, it cannot be attributed to the phosphorylation reaction. The change rather indicates a switch towards the conformation observed in the archaea dsDNA-bound state. Reverse LRT results of the archaea dsDNA-bound and dsDNA-unbound state indicate several regions, where forces might act upon to induce the switch from one state to the other. The serine of interest (Ser806) is in close proximity to a region of high forces. Finally, the perturbation of the archaea dsDNA-unbound state reveals a similar structural response observed in LRT I. This additionally accounts for a conserved, robust dynamics intrinsic to the Rad54 structure.

These LRT results together with general considerations from the literature, biochemical findings and structural insights from the three crystal structures lead to a formulation of the Rad54 translocation cycle (Figure 4.12) [187, 45, 179]. The following mechanistic steps might involved:

1. NEK1 phosphorylates Rad54 at a single position (Ser572 in *Homo sapiens*, Ser566 in *Danio rerio* and Ser806 in *Sulfolobus solfataricus*). Experiments confirm this to be necessary for Rad54 activation whereas other trigger mechanism cannot be excluded [173]. The phosphorylation takes place most likely in the dsDNA-unbound state, that is represented by PDB structure 1Z6A.



**Figure 4.11:** LRT III: A) Expected displacement of archaea Rad54 C $\alpha$  atoms of one single LRT run after perturbing atom Ser806 (marked by a black arrow). For details, see Table 4.2. Several structural regions are indicated: ATP-binding site (ATP-bs), DNA-binding site (DNA-bs) and the linker region connecting the two domains. Note that the absolute scaling is arbitrary. B) Visualization of expected displacement. Coordinate shifts are visualized by small arrows pointing from the C $\alpha$  atom's initial atom positions to the final after LRT using force scaling factor of 100. The force vector of the perturbed atom is indicated by a big yellow arrow. The structure of archaea Rad54 (PDB code: 1Z6A) is shown in new cartoon representation (gray) in the background, front view (left panels) and side view (right panels). Note that the shifted structure was fitted to the initial by superimposing domain 1 for visualization purposes. Structures were rendered using VMD.

- As a structural consequence, this phosphorylation might lead to a widening of the cleft flanked by the domains 1 and 2. dsDNA can now interact with the positively-charged patch and binds to the high affinity domain 1 whereas domain 2 is only loosely bound [45]. This state is represented by PDB structure 1Z63.

3. Upon dsDNA-binding, the DExx motif interacts with the dsDNA and changes from an unusual  $\beta$ -conformation into a typical active  $\alpha$ -conformation (as it is commonly observed in the functional active site of Walker box ATPases). Thereby, Glu563 flips by  $180^\circ$  and is able to polarize a water molecule required for ATP hydrolysis. This fact was observed in the archaea crystal structures [45].
4. ATP binds to the pocket and induces a  $180^\circ$  rotation of the loosely-bound domain 2 with the entire molecule changing from the *open-state* to the *closed-state* (represented by 1Z3I). The domain 2 pushes the minor groove located upstream. This results in a forward translocation and an increased tension of the complex [187].
5. The structure is relaxed upon ATP hydrolysis and the entire molecule returns from the *closed-state* back to the *open-state*. The released force causes the high-affinity domain 1 to be pulled upstream before it reattaches again tightly to dsDNA. The *open state* (represented again by 1Z63) is restored. The pocket is active and can bind further ATP-molecules (see step 4).

Biochemical results have shown that the Rad54 dsDNA recognition is not influenced by ATP but vice versa the ATPase activity is stimulated by dsDNA [45]. This leads us to conclude that the Ser566 phosphorylation reaction occurs once in the dsDNA-unbound state and is responsible for the activation of dsDNA-binding.

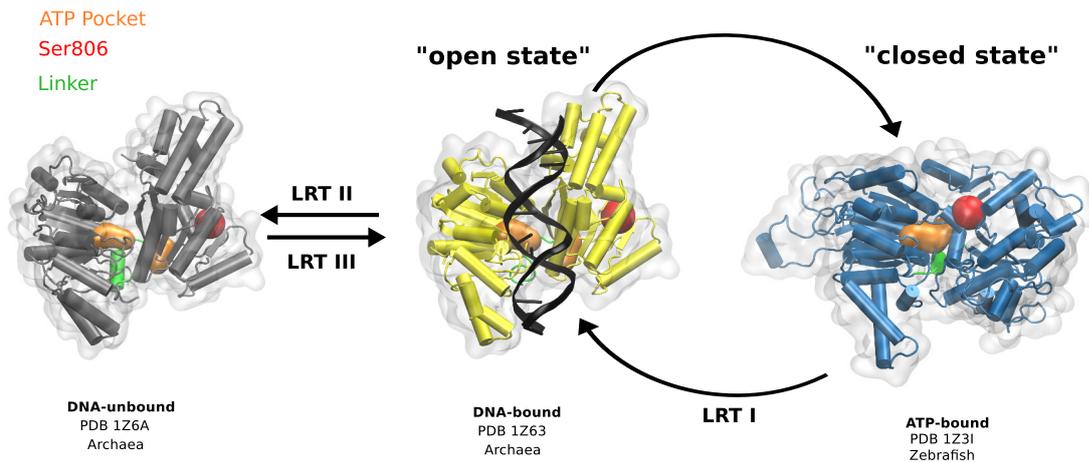
### Modeling the Zebrafish Open-State Structure

In order to obtain the supposed correct starting structure for Ser566 phosphorylation (outside the cycle) of zebrafish Rad54, we modeled the *closed-state* (Figure 4.12) based on the template of the dsDNA-unbound archaea structure.

The two domains were separated and superposed to the template structure over the  $\beta$ -sheets of the central lobes. Both domains 1 and 2, revealed a very low deviation to the template indicating minor conformational changes within the domains. The most drastic structural change is observed in the  $\alpha$ -helical linker connecting the two domains showing two different conformations in the crystal structures. The entire 11-residue linker from the archaea structure is adopted.

After reconnecting the rotated domain 2 to the domain 1, the large C-terminal region of the second one showed a severe clash with the first domain. The prediction of the arrangement and location of this C-terminal region in the modeled state is not clear and therefore the C-terminal region was omitted.

The modeled zebrafish structure is supposed to exhibit the same LRT response upon perturbation of Ser566 as observed with its template structure in LRT III. Because of the high congruence between the domains the  $C^\alpha$  atom network topology will be highly redundant leading to similar LRT responses. Nevertheless, this structure provides us with a closer look into the dynamic Rad54 translocation cycle and serves as an initial structure for several further studies.



**Figure 4.12:** Crystal structures of zebrafish and archaea Rad54 used in this study show putative conformational states within the Rad54 translocation cycle on dsDNA. Certain structural transitions were investigated by means of ANM-based LRT experiments I-III.

## 4.5 Conclusion

The original idea of this study was to compute Rad54 structural changes upon phosphorylation of the single residue Ser572 that was experimentally shown to be crucial for human Rad54 activation [173]. This phosphorylation reaction is thought to trigger the activation by a conformational change. Here, the linear response theory in combination with anisotropic network models (ANM-based LRT) is chosen as a straightforward, mechanical approach to reveal those structural changes by mimicking the phosphorylation reaction.

Three crystal structures of the Rad54 core ATPase domain of zebrafish *Danio rerio* and dsDNA-bound and -unbound archaea *Sulfolobus solfataricus* are available revealing the Rad54 protein in different conformational states. Indeed, the protein sequences of those species exhibit a conserved serine residue being equivalent to the human Ser572 (for zebrafish Ser566 and archaea Ser806) and superpositioning of the structures shows the serine residue to be in the exact same location in domain 2. Based on these crystal structures, three different LRT experiments I-III were performed.

Because of the close relation of zebrafish Rad54 to the human analog, we investigated the LRT response of this structure after Ser566 perturbation (LRT I). A drastic conformational change of domain 2 (180° turn) relative to domain 1 is observed. This displacement is shown to be rather independent from perturbation force strength and direction as is clear from our null model approach.



the dsDNA-unbound archaea structure shows the serine residue to be exposed at the surface at a highly accessible position. That motivates the dsDNA-unbound state to be the correct state for our phosphorylation investigations. Therefore, the LRT of the dsDNA-unbound archaea structure is conducted upon perturbation of Ser806 (LRT III). The resulting structural response reveals a similar rotational shift of domain 2 relative to domain 1. This was also observed in LRT I.

By analyzing the LRT results, it appears that there is a general dynamics inherent to the Rad54 structure which is independent from perturbing one or more atoms. The global dynamics revealed seem to be highly related to the protein's role as a remodeling enzyme. During the complex cycle of Rad54 translocation on dsDNA, the protein undergoes drastic conformational changes as they are observed in related helicases [187]. Based on the literature [187, 45, 179], it was possible to assign the conformations observed in the three different crystal structures to distinct stages of the cycle. Based on these findings, we created a structural model of the zebrafish Rad54 in the putative initial state or *open-state* based on the ATP- and dsDNA-unbound archaea template.

Although our results did not reveal a structural response specific to the phosphorylation reaction, this does not mean the ANM-based LRT to be an inadequate method to model such conformational changes. One might argue that the approach is not sufficient for modeling a phosphorylation reaction, because on the one hand this biochemical phenomena is highly side-chain dependent and on the other hand provokes a charge transfer. Although in the ANM-based LRT both physicochemical characteristics are not treated explicitly, the application of attractive and repulsive harmonic potentials should be sufficient to account for general charge transfer. A phosphorylation reaction can therefore thoroughly be mimicked by applying a combination of attractive and repulsive forces to the atom set in the environment of the phosphorylated residue.

In this work the linear response is mediated by the ANM-derived covariance matrix. Despite the very simplistic nature of the ANMs, in several studies the results showed good agreement to those based on computationally intensive and accurate force-field-based NMA [22]. The normal modes obtained by NMA can be divided into high-frequency modes responsible for local changes accounting for structural stability of the protein and low-frequency modes showing global structural changes [8]. The latter were often found to represent those collective motions being relevant for biological function and ANMs were shown to predict these modes [7]. Most probably, the phosphorylation reaction leads to a transitional change between two conformational states constituting a path on the potential energy surface.

Instead, the LRT results reveal a robust global structural change being relatively independent from the applied force strengths and direction, the perturbation site and ANM spring constants. This is not surprising, since for highly-dynamic ATP-driven motor proteins like helicases or topoisomerases a drastic conformational change is expected [54]. Indeed, the structural LRT response obtained in this work confirms the rotational changes necessary for the transition between *open-* and *closed-state* that was proposed in a previous work [45].

Most structural changes that proteins undergo are too complex to be accurately calculated using LRT. This drawback is due to 1) the constant contact topology between  $C^\alpha$  atoms and 2) linear character of LRT. As an extension, alternative network models could be applied that consider non-linear dynamics [50].

A further limitation of ANM-based LRT is the inability to account for possible influences of point mutations, as they are commonly introduced within biochemical mutagenesis studies. On the one hand, a uniform spring constant was applied for the ANMs that neglect differences in amino acid residues and, on the other hand, even with heterogeneous parameterization schemes the differences in LRT response would be minor.

A follow-up experiment would be a protein-protein docking approach with Rad54 and NEK1 that would help to determine the adequate structural state in which the phosphorylation reaction occurs. For this purpose, the modeled *open-state* dsDNA-unbound zebrafish Rad54 structure obtained in this work could be used. In addition, a model of the *open-state* dsDNA-bound state could be created based on the archaea template.

## 5 Conclusion

The main objective of this work is the investigation of DNA damage response in radio-stressed tumor cells by computational means at a molecular level. Such detailed insights would provide new opportunities in modulating DNA damage response pathways in order to enhance radiotherapy efficiency in cancer treatment. Complex molecular machineries involving numerous proteins have evolved to overcome various types of DNA damages. Hence, several strategies exist to sensitize tumor cells to ionizing radiation, e.g., the enhancement of the programmed cell death or the interference with DNA double strand break repair pathways. As a consequence, proteins and protein complexes playing substantial roles in mediating those pathways were chosen as subjects of investigation in this work: the 20S proteasome, the DNA-PK complex and the Rad54 protein.

A protein's function is primarily determined by its three-dimensional fold. In recent years, full-atomic crystal structures of many proteins have been solved and led to detailed insights into the overall topology. However, X-ray crystallography data does not account for dynamical changes within the protein which are important to fully capture the protein's behavior. Computational (*in silico*) models offer the possibility to investigate molecular mechanisms at spatial and temporal scales that are impossible to observe in wet-lab (*in vitro* and *in vivo*) experiments. The purpose of biophysical simulations is to describe protein motions as accurate as possible. However, various assumptions are made to reduce computational complexity. A challenging aspect of this work was to select models of appropriate reduction to address the specific research questions.

Molecular docking simulations are based on a full-atomic representation of ligands inside a protein pocket to predict the correct binding pose. In this work, different aspects of covalently bound ligand docking were investigated to guide structure-based, rationale drug design of proteasome inhibitors. In a collaboration project it was possible to develop new highly potent and selective inhibitors [190]. Further, we established a protocol for the docking of covalently bound ligands to accelerate future development of proteasome inhibitors.

Additionally, full-atomic and thus computationally more intensive molecular dynamics simulations were performed to capture the dynamics of an amino acid side chain located in one of the proteasome's active sites. Based on these results we can confirm a hypothesized induced-fit mechanism where the binding pocket adapts to ligands of different sizes. These findings successfully contributed to the elucidation of the proteasome's natural substrate cleavage mechanism [165].

In order to study dynamics of the Rad54 protein, elastic network models were applied to predict structural responses upon residue perturbation. Instead of local

changes expected to be induced upon a certain phosphorylation reaction, global dynamics were revealed. In our developed null model, we could show them to be rather independent from the perturbation site. This major structural change might be attributed to Rad54's inherent role of a dsDNA translocase. With detailed structural studies of the translocation cycle we were able to transform the zebrafish Rad54 structure into another conformation relevant to these dynamics. This structural model can serve as subject for further investigations. Still, the phosphorylation-induced local changes need to be analyzed using alternative models.

Apart from physicochemical techniques to study mechanical dynamics of biomolecules, information-theoretical concepts were applied to describe coevolutionary dynamics in homologous sequence sets of proteins. In this work, a mutual information (MI)-based workflow was developed to predict protein-protein interactions in the yet unknown structure of the DNA-PK complex. Here, the elimination of background MI signals and the separation of intra-protein signals from inter-protein ones presented major challenges. The predictive power of different MI correction variants was assessed of which most of them achieving a rather low prediction quality. These evaluation results confirm the hypothesis of Codoner et al. [30] stating that coevolution can be considered as a complex phenomenon influenced by diverse molecular events. Thus, the extraction of relevant signals associated with protein-protein interactions is challenging. A promising strategy consists in the application of alternative methodologies such as the direct coupling analysis (DCA). Here, a global statistical model can be estimated to account for position-specific amino acid bias. Regardless of the statistical method used a wider number of protein complexes must be taken into account for general and accurate validation of any results.

Concerning the investigation of the DNA-PK complex, a major obstacle consisted in the lack of the full-atomic DNA-PKcs structure. Thereby, analyses were restricted to the Ku70/Ku80 complex. Sequence-based approaches using MI-based networks might be appropriate to establish a more detailed model of DNA-PKcs.

Taken together, adequate computational strategies were chosen to explore the structure and function of three different proteins and protein complexes involved in the DNA damage response. The computational models applied were proven to be successful in elucidating underlying molecular mechanisms. In conclusion, it has been shown that the results obtained in this work contributed to a broader understanding of DNA damage response modulation [190, 165] and might lead to additional insights in the ever-growing field of radiation biology.

# A Appendix

## PDBBind2007 Core Database: Selected Structures

**Table A.1:** Crystal structures of protein-ligand complexes used for redocking study in AutoDock. 50 structures were taken from the refined PDBBind2007 database A.1.

structure			protein		ligand			affinity		
PDB code	resolution [Å]	release [year]	name	EC number	name	mass	PSA	type	constant [nM]	pKd
1A1B	2.20	1998	c-Src tyrosine kinase	2.7.1.112	4MER	569	201	Kd	400	6.40
1A69	2.10	1998	purine nucleoside phosphorylase	2.4.2.1	FMB	268	136	Ki	5000	5.30
1A15	2.36	1997	penicillin amidohydrolase	3.5.1.11	MNP	180	86	Ki	189000	3.72
1APW	1.80	1994	penicillopepsin	3.4.23.20	5MER	507	137	Ki	10	8.00
1AXZ	1.95	1998	lectin		GAL	180	110	Kd	637000	3.20
1C84	2.35	2000	protein-tyrosine phosphatase 1b	3.1.3.48	761	257	109	Ki	9900	5.00
1DET	1.80	1996	ribonuclease t1	3.1.27.3	2GP	361	217	Kd	50000	4.30
1DF8	1.51	2000	streptavidin		BTN	243	107	Kd	0.2	9.70
1E5A	1.80	2000	transthyretin		TBP	331	20	Kd	23	7.64
1F4F	2.00	2000	thymidylate synthase	2.1.1.45	TP3	425	195	Ki	24000	4.62
1FD0	1.38	2002	retinoic acid receptor gamma-1		254	400	73	Kd	4	8.40
1FKI	2.20	1994	FK506 binding protein		SB1	438	90	Ki	100	7.00
1HA2	2.50	2001	serum albumin		SWF	308	64	Kd	2900	5.54
1HI4	1.80	2001	eosinophil-derived neurotoxin	3.1.27.5	A3P	423	264	Ki	32000	4.50
1J16	1.60	2002	trypsin-2, anionic	3.4.21.4	BEN	121	52	Ki	143000	3.84
1JQE	1.91	2002	histamine n-methyltransferase	2.1.1.8	QUN	400	37	Ki	360	6.44
1K4G	1.70	2002	tRNA-guanine transglycosylase	2.4.2.29	AIQ	288	143	Ki	1400	5.85
1KV5	1.65	2002	triosephosphate isomerase, glycosomal	5.3.1.1	PGA	153	122	Ki	60000	4.22
1L2S	1.94	2002	$\beta$ -lactamase	3.5.2.6	STC	317	120	Ki	26000	4.58
1L16	2.00	2002	lysozyme	3.2.1.17	SMP	81	12	Kd	160000	3.80
1L0L	1.90	2002	orotidine 5'-monophosphate decarboxylase	4.1.1.23	XMP	362	208	Ki	410	6.39
1O0H	1.20	2003	ribonuclease, pancreatic	3.1.27.5	ADP	424	258	Ki	1200	5.92
1O3P	1.81	2003	urokinase-type plasminogen activator	3.4.21.73	655	337	106	Ki	220	6.66
1OM1	1.68	2004	casein kinase-2	2.7.1.37	IQA	291	72	Ki	170	6.77
1PB9	1.60	2003	n-methyl-d-aspartate receptor		4AX	103	66	Ki	241000	3.62
1Q7A	1.60	2004	phospholipase A2 VRV-PL-VIIIa	3.1.1.4	OPB	324	61	Kd	64	7.19
1RDI	1.80	1996	mannose-binding protein-C		MFU	178	79	Ki	880000	2.06
1SYH	1.80	2005	glutamate receptor 2		CPW	239	117	Ki	487	6.31
1TOJ	1.90	2004	aspartate aminotransferase	2.6.1.1	HCI	149	40	Kd	410000	3.39
1TSY	2.20	1996	thymidylate synthase	2.1.1.45	UMP	306	161	Kd	11000	4.96
1TTM	1.95	2004	carbonic anhydrase II	4.2.1.1	667	309	104	Kd	45	7.35
1U33	1.95	2004	$\alpha$ -amylase, pancreatic	3.2.1.1	LM2	530	270	Ki	25000	4.60
1V2O	1.62	2004	trypsin	3.4.21.4	ANH	434	162	Ki	18450	4.73
1V48	2.20	2004	purine nucleoside phosphorylase	2.4.2.1	HA1	335	156	Ki	16	7.80
1ZC9	2.00	2006	2,2-dialkylglycine decarboxylase	4.1.1.64	PMP	247	143	Kd	600000	3.22
1ZS0	1.56	2006	neutrophil collagenase	3.4.24.34	EIN	397	137	Ki	700	6.16
2BAK	2.20	2005	mitogen-activated protein kinase 14	2.7.1.37	AQZ	584	114	Kd	37	7.43
2BRM	2.20	2005	serine/threonine-protein kinase Chk1	2.7.1.37	DFZ	330	70	Ki	1300	5.89
2CEQ	2.14	2006	$\beta$ -galactosidase	3.2.1.23	GIM	200	97	Ki	53	7.28
2CGR	2.20	1994	IgG2b-kappa NC6.8 Fab (light chain)		GAS	384	100	Kd	53	7.28
2D1O	2.02	2006	stromelysin 1	3.4.24.17	FA4	465	170	Ki	20	7.70
2ER9	2.20	1991	endothiapepsin	3.4.23.6	8MER	906	287	Ki	40	7.40
2FAI	2.10	2006	estrogen receptor		459	274	50	Ki	570	6.24
2H3E	2.30	2006	aspartate carbamoyltransferase	2.1.3.2	6PR	251	185	Kd	2000	5.70
2J78	1.65	2006	$\beta$ -glucosidase a	3.2.1.21	GOX	192	126	Kd	384	6.42
2QWD	2.00	1998	neuraminidase	3.2.1.18	4AM	290	167	Ki	14000	4.85
2STD	2.10	1999	scytalone dehydratase	4.2.1.94	CRP	335	29	Ki	0.14	9.85
3GSS	1.90	1997	glutathione S-transferase p1-1	2.5.1.18	GTT	608	258	Ki	1500	5.82
3PCH	2.05	1998	protocatechuate 3,4-dioxygenase	1.13.11.3	CHB	137	60	Ki	4000	5.40
5ABP	1.80	1992	L-arabinose-binding protein		GLA	180	110	Kd	230	6.64
8CPA	2.00	1994	carboxypeptidase a	3.4.17.1	AGF	462	167	Ki	0.71	9.15

## Redocking of the PDBBind2007 Database: Parameter Setups

**Table A.2:** Parameter setups chosen for the conformational search methods Lamarckian genetic algorithm (LGA), simulated annealing (SA) and stochastic tunneling (STUN). The following parameters were changed: population size (pop), mutational rate (mut), crossover rate (cross), initial annealing temperature  $T_0$ , temperature reduction factor  $rtrf$ , temperature  $T$  and the transformation parameter  $\gamma$ . Standard parameters are printed in bold type.

setup	LGA			SA		STUN	
	pop	mut	cross	$T_0$	$rtrf$	$T$	$\gamma$
1	50	0.02	0.7	400	0.8	400	0.010
2	50	0.02	0.8	400	0.9	400	0.025
3	50	0.02	0.9	400	0.95	400	0.050
4	50	0.05	0.7	500	0.8	400	0.070
5	50	0.05	0.8	500	0.9	400	0.100
6	50	0.05	0.9	<b>500</b>	<b>0.95</b>	500	0.010
7	50	0.10	0.7	600	0.8	500	0.025
8	50	0.10	0.8	600	0.9	<b>500</b>	<b>0.050</b>
9	50	0.10	0.9	600	0.95	500	0.070
10	125	0.02	0.7			500	0.100
11	<b>125</b>	<b>0.02</b>	<b>0.8</b>			600	0.010
12	125	0.02	0.9			600	0.025
13	125	0.05	0.7			600	0.050
14	125	0.05	0.8			600	0.070
15	125	0.05	0.9			600	0.100
16	125	0.10	0.7				
17	125	0.10	0.8				
18	125	0.10	0.9				
19	200	0.02	0.7				
20	200	0.02	0.8				
21	200	0.02	0.9				
22	200	0.05	0.7				
23	200	0.05	0.8				
24	200	0.05	0.9				
25	200	0.10	0.7				
26	200	0.10	0.8				
27	200	0.10	0.9				

---

## Redocking of the PDBBind2007 Database: Result Details

**Table A.3:** Spearman's ranking correlation coefficient between AutoDock score (estimated free energy of binding [kcal/mol]) and RMSD relative to the starting crystal structure. 50 protein-ligand complexes from the PDBBind2007 were redocked in 100 individual docking runs for each algorithm and setup: LGA 1-27, SA 1-9 and STUN 1-15. Details for parameter setups are listed in appendix Table A.2.

setup	algorithm		
	LGA	SA	STUN
1	0.486	0.550	0.539
2	0.468	0.622	0.596
3	0.503	0.639	0.613
4	0.493	0.542	0.531
5	0.492	0.623	0.599
6	0.487	0.636	0.619
7	0.509	0.544	0.541
8	0.512	0.626	0.621
9	0.502	0.645	0.610
10	0.467		0.534
11	0.467		0.595
12	0.474		0.617
13	0.497		0.543
14	0.495		0.590
15	0.493		0.633
16	0.518		
17	0.519		
18	0.517		
19	0.479		
20	0.494		
21	0.495		
22	0.506		
23	0.507		
24	0.497		
25	0.534		
26	0.529		
27	0.524		

## Redocking of BSc2118: Result Details

**Table A.4:** Result details of the redocking evaluation of BSc2118 in AutoDock using seven different setups and three different search algorithms: LGA, SA and STUN. The mean  $\mu$  and standard deviations  $\sigma$  of the estimated free energy of binding  $\Delta G$  and the RMSD are listed according to 100 individual runs. The RMSD of the best predicted pose of every run was calculated with respect to the crystal structure complex.

LGA	parameter setup						
	1	2	3	4	5	6	7
job duration [s]	$1.71 \times 10^4$	$1.98 \times 10^4$	$6.64 \times 10^4$	$6.67 \times 10^4$	$7.45 \times 10^4$	$7.96 \times 10^4$	$8.59 \times 10^4$
$\mu_{\Delta G}$	-4.818	45.453	46.856	47.828	50.378	48.569	1.453
$\sigma_{\Delta G}$	0.004	0.025	2.244	2.210	3.975	2.307	1.018
$\mu_{\text{RMSD}}$	0.993	0.883	8.155	8.406	8.204	8.318	8.056
$\sigma_{\text{RMSD}}$	0.016	0.005	0.800	0.918	0.830	0.883	1.016
SA	1	2	3	4	5	6	7
job duration [s]	7	46	94	98	crash	crash	crash
$\mu_{\Delta G}$	1.238	1344.888	2742.409	10948.129	-	-	-
$\sigma_{\Delta G}$	0.481	7925.556	12650.230	104955.200	-	-	-
$\mu_{\text{RMSD}}$	9.218	6.254	8.931	8.791	-	-	-
$\sigma_{\text{RMSD}}$	1.402	2.218	1.058	1.114	-	-	-
STUN	1	2	3	4	5	6	7
job duration [s]	7	43	26	88	crash	crash	crash
$\mu_{\Delta G}$	1.474	265.727	520.715	561.789	-	-	-
$\sigma_{\Delta G}$	0.692	88.344	185.737	176.506	-	-	-
$\mu_{\text{RMSD}}$	9.146	5.948	8.959	8.850	-	-	-
$\sigma_{\text{RMSD}}$	1.397	2.030	1.012	1.128	-	-	-

---

## Residue Integrity of Ku Structures

**Table A.5:** Details of the three different structures of the Ku protein. It is indicated which amino acid residues are present in the structures of Ku70 and Ku80. Note that the Ku80 C-terminal region (Ku80CTR) absent from both crystal structures, was determined isolated in solution. NMR: nuclear magnetic resonance.

PDB code	1JEQ	1JEY	1Q2Z
type	crystal structure	crystal structure	NMR structure
molecule	Ku70/Ku80, DNA-free	Ku70/Ku80, DNA-bound	Ku80CTR
residues of Ku70	35-223 231-538 559-609	34-222 231-532 -	
residues of Ku80	6-169 182-189 192-323 327-542 -	6-170 181-545 - - -	- - - - 590-709

## Organism Frequencies in Combined Alignments

**Table A.6:** Organism names present in the combined alignments after the permutation (p) and clustering (c) approach. In case of the clustering, sequence ids (ids) are specified indicating which exact sequence was kept.

organism name	p	c ids		p	c ids
<i>Acanthamoeba castellanii</i>	1	1 (1,1)	<i>Exophiala dermatitidis</i>	1	1 (1,1)
<i>Acromyrmex echinator</i>	1	1 (1,1)	<i>Fomitiporia mediterranea</i>	1	1 (1,1)
<i>Ajellomyces capsulatus</i>	2	1 (1,1)	<i>Galdieria sulphuraria</i>	4	1 (1,1)
<i>Ajellomyces dermatitidis</i>	4	1 (1,1)	<i>Glomerella graminicola</i>	1	1 (1,1)
<i>Alternaria alternata</i>	1	1 (1,1)	<i>Grosmannia clavigera</i>	1	1 (1,1)
<i>Anopheles gambiae</i>	1	1 (1,1)	<i>Harpegnathos saltator</i>	1	1 (1,1)
<i>Arabidopsis thaliana</i>	6	1 (2,1)	<i>Heterocephalus glaber</i>	1	1 (1,1)
<i>Arthroderma gypseum</i>	1	1 (1,1)	<i>Homo sapiens</i>	24	1 (1,1)
<i>Aspergillus kawachii</i>	1	1 (1,1)	<i>Hordeum vulgare</i>	1	1 (1,1)
<i>Aspergillus niger</i>	4	1 (1,1)	<i>Komagataella pastoris</i>	2	1 (1,1)
<i>Aspergillus oryzae</i>	2	1 (1,1)	<i>Leptosphaeria maculans</i>	1	1 (1,1)
<i>Aspergillus sojae</i>	1	1 (1,1)	<i>Macaca fascicularis</i>	1	1 (1,1)
<i>Auricularia delicata</i>	1	1 (1,1)	<i>Macrophomina phaseolina</i>	1	1 (1,1)
<i>Bos grunniens</i>	1	1 (1,1)	<i>Metarhizium acridum</i>	1	1 (1,1)
<i>Bos taurus</i>	1	1 (1,1)	<i>Metarhizium anisopliae</i>	1	1 (1,1)
<i>Brachionus ibericus</i>	1	1 (1,1)	<i>Mus musculus</i>	30	1 (2,1)
<i>Caenorhabditis briggsae</i>	1	1 (1,1)	<i>Mycosphaerella populorum</i>	1	1 (1,1)
<i>Caenorhabditis elegans</i>	1	1 (1,1)	<i>Neurospora crassa</i>	4	1 (1,1)
<i>Caenorhabditis remanei</i>	1	1 (1,1)	<i>Neurospora tetrasperma</i>	2	1 (1,1)
<i>Camponotus floridanus</i>	1	1 (1,1)	<i>Penicillium chrysogenum</i>	1	1 (1,1)
<i>Claviceps purpurea</i>	1	1 (1,1)	<i>Penicillium digitatum</i>	1	1 (1,1)
<i>Coccidioides immitis</i>	1	1 (1,1)	<i>Phaeosphaeria nodorum</i>	1	1 (1,1)
<i>Colletotrichum gloeosporioides</i>	1	1 (1,1)	<i>Piriformospora indica</i>	1	1 (1,1)
<i>Colletotrichum higginsianum</i>	2	1 (1,1)	<i>Polysphondylium pallidum</i>	1	1 (1,1)
<i>Columba livia</i>	1	1 (1,1)	<i>Pteropus alecto</i>	1	1 (1,1)
<i>Coniophora puteana</i>	1	1 (1,1)	<i>Punctularia strigosozonata</i>	1	1 (1,1)
<i>Coprinopsis cinerea</i>	1	1 (1,1)	<i>Pyrenophora tritici-repentis</i>	1	1 (1,1)
<i>Crassostrea gigas</i>	1	1 (1,1)	<i>Rattus norvegicus</i>	6	1 (1,1)
<i>Cricetulus griseus</i>	2	1 (1,1)	<i>Schizophyllum commune</i>	1	1 (1,1)
<i>Cryptococcus neoformans</i>	1	1 (1,1)	<i>Schizosaccharomyces japonicus</i>	1	1 (1,1)
<i>Dacryopinax sp.</i>	1	1 (1,1)	<i>Schizosaccharomyces pombe</i>	1	1 (1,1)
<i>Danio rerio</i>	4	1 (1,1)	<i>Sporisorium reilianum</i>	1	1 (1,1)
<i>Dichomitus squalens</i>	1	1 (1,1)	<i>Tetrahymena thermophila</i>	2	1 (1,1)
<i>Drosophila ananassae</i>	1	1 (1,1)	<i>Tetraodon nigroviridis</i>	1	1 (1,1)
<i>Drosophila erecta</i>	1	1 (1,1)	<i>Trametes versicolor</i>	1	1 (1,1)
<i>Drosophila grimshawi</i>	1	1 (1,1)	<i>Triticum aestivum</i>	1	1 (1,1)
<i>Drosophila melanogaster</i>	176	1 (7,16)	<i>Tupaia chinensis</i>	2	1 (1,1)
<i>Drosophila mojavensis</i>	1	1 (1,1)	<i>Ustilago hordei</i>	1	1 (1,1)
<i>Drosophila persimilis</i>	1	1 (1,1)	<i>Verticillium albo-atrum</i>	1	1 (1,1)
<i>Drosophila pseudoobscura</i>	1	1 (1,1)	<i>Verticillium dahliae</i>	1	1 (1,1)
<i>Drosophila sechellia</i>	1	1 (1,1)	<i>Vigna radiata</i>	1	1 (1,1)
<i>Drosophila simulans</i>	6	1 (1,5)	<i>Wallemia sebi</i>	1	1 (1,1)
<i>Drosophila virilis</i>	1	1 (1,1)	<i>Xenopus laevis</i>	4	1 (1,1)
<i>Drosophila willistoni</i>	1	1 (1,1)	<i>Yarrowia lipolytica</i>	1	1 (1,1)
<i>Drosophila yakuba</i>	1	1 (1,1)	<i>Zygosaccharomyces rouxii</i>	1	1 (1,1)

---

## Z-Score Thresholds

**Table A.7:** Z-score thresholds derived from the 75% quantile of the Z-score distribution of MI matrices before alignment combination (Intra) and after alignment combination (Intra' and Inter).

proteins(s)	MI matrix	Z-score threshold	
Ku70	Intra-Ku70	21.29	
Ku80	Intra-Ku80	17.85	
DNA-PKcs	Intra-DNA-PKcs	34.90	
		clustering	permutation
Ku70/Ku80	Intra'-Ku70	9.79	53.36
	Intra'-Ku80	9.08	55.27
	Inter	9.03	44.46
Ku70/DNA-PKcs	Intra'-Ku70	9.97	187.49
	Intra'-DNA-PKcs	7.29	129.85
	Inter	3.20	70.30
Ku80/DNA-PKcs	Intra'-Ku80	9.67	44.46
	Intra'-DNA-PKcs	7.06	53.35
	Inter	3.26	55.25

## Reproduction Capacity of Eigenvectors

**Table A.8:** Relevance of first eigenvectors yielded from singular value decomposition of MI matrices for different normalization variants. Pearson's correlation coefficient was computed for the degree of similarity between the reconstructed matrix and the original one.

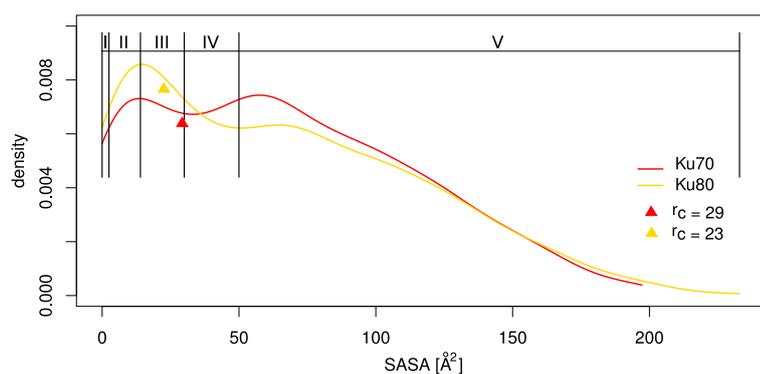
		Inter-MI							
		MI	MI.Z	MNE	clustering		APC.Z	RCW	RCW.Z
					MNE.Z	APC			
Ku70/Ku80	Intra'-Ku70	0.96	0.71	0.45	0.88	0.76	0.74	0.31	0.73
	Intra'-Ku80	0.96	0.73	0.4	0.86	0.76	0.73	0.25	0.72
	Inter	0.97	0.75	0.44	0.91	0.77	0.74	0.37	0.74
Ku70/DNA-PKcs	Intra'-Ku70	0.96	0.73	0.43	0.87	0.77	0.74	0.29	0.73
	Intra'-DNA-PKcs	0.96	0.72	0.59	0.79	0.76	0.55	0.38	0.68
	Inter	0.99	0.73	0.41	0.97	0.77	0.71	0.29	0.74
Ku80/DNA-PKcs	Intra'-Ku80	0.95	0.73	0.4	0.85	0.76	0.72	0.23	0.72
	Intra'-DNA-PKcs	0.97	0.7	0.57	0.84	0.78	0.62	0.4	0.71
	Inter	0.99	0.72	0.39	0.97	0.76	0.69	0.27	0.73
		MI	MI.Z	MNE	permutation		APC.Z	RCW	RCW.Z
					MNE.Z	APC			
Ku70/Ku80	Intra'-Ku70	0.98	0.8	0.56	0.93	0.83	0.83	0.45	0.82
	Intra'-Ku80	0.96	0.75	0.57	0.87	0.78	0.77	0.49	0.75
	Inter	0.98	0.86	0.56	0.94	0.83	0.82	0.4	0.81
Ku70/DNA-PKcs	Intra'-Ku70	0.97	0.79	0.47	0.92	0.81	0.80	0.28	0.79
	Intra'-DNA-PKcs	0.97	0.63	0.55	0.84	0.84	0.79	0.34	0.81
	Inter	0.99	0.91	0.49	0.98	0.91	0.9	0.49	0.9
Ku80/DNA-PKcs	Intra'-Ku80	0.95	0.76	0.42	0.86	0.81	0.79	0.26	0.78
	Intra'-DNA-PKcs	0.96	0.7	0.56	0.81	0.81	0.73	0.31	0.77
	Inter	0.99	0.86	0.45	0.97	0.88	0.86	0.33	0.85

---

## Surface Definition of Ku70 and Ku80 Proteins

**Table A.9:** Solvent-accessible surface area (SASA)-based classification of surface residues according to Jha et al. [101]. Ratios of Ku70 and Ku80 residues belonging to the five classes of residues that are completely exposed on the surface (I), regularly exposed (II), intermediately exposed (III), partially buried (IV) and totally buried (V) are listed and visualized in appendix Figure A.1.

class	SASA [ $\text{\AA}^2$ ]	type	surface residues [%]	
			Ku70	Ku80
I	$> 50$	completely exposed	59.3	56.0
II	$> 30 \leq 50$	exposed	13.5	12.5
III	$> 14 \leq 30$	intermediate	10.8	12.1
IV	$> 2.5 \leq 14$	partially buried	10.2	14.8
V	$\leq 2.5$	totally buried	6.2	4.6

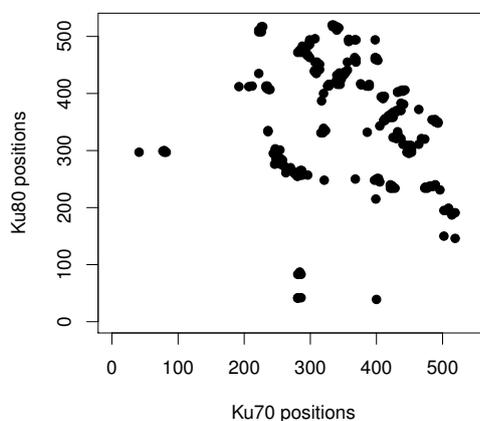


**Figure A.1:** Distribution of computed SASA values for each of the molecules Ku70 (red) and Ku80 (yellow) separately. The surface classification described in appendix Table A.9 is shown in Roman numerals. The SASA cutoff  $r_c$  for surface residue definition is marked by a triangle.

## Interaction Definition between Ku70 and Ku80

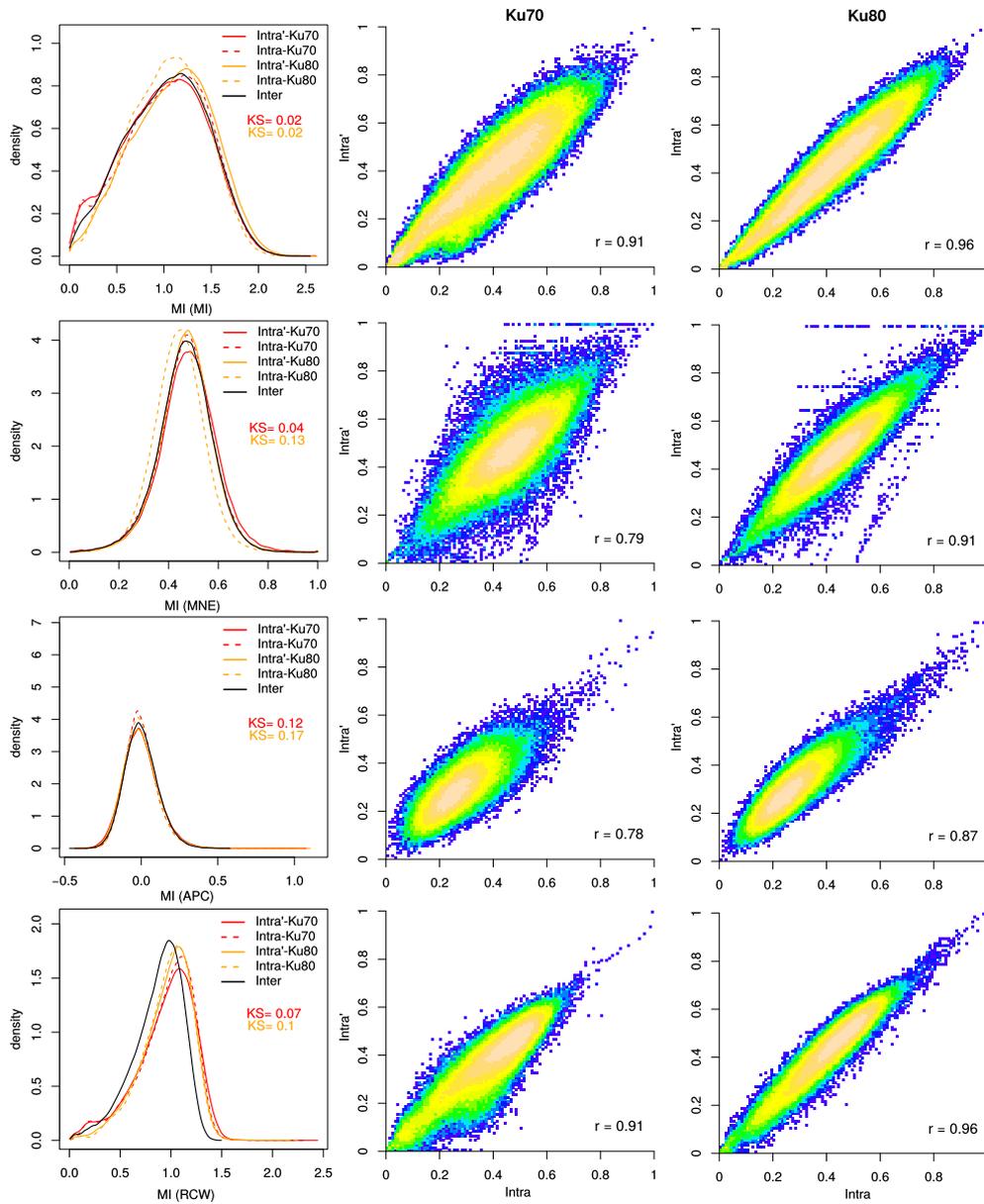
**Table A.10:** Interactions for Ku70 and Ku80 observed in the crystal structure (PDB code: 1JEQ), according to different interaction types and distances in Å. Optimization of the side chain conformations were further conducted using SCRWL 4.0 [112]. A total of 295 contacts was found.

Interaction type	atoms	possible residues involved	cutoff	count
hydrophobic		Ala, Val, Leu, Ile, Met, Phe, Trp, Pro, Tyr	5	171
H-bonds (b-b)	N-O	all	3.5	24
H-bonds (b-s)	N-O	all	3.5	32
H-bonds (s-s)	N-O	all	3.5	18
H-bonds (s-s)	S-S	(Cys, Met) + all	4	8
ionic		Arg, Lys, His, Asp, Glu	6	22
aromatic-aromatic		Tyr, Phe, Trp	7	10
aromatic-sulfur		Tyr, Phe, Trp + Met, Cys	5.3	4
cation- $\pi$		Lys, Arg + Tyr, Phe, Trp	6	6
disulfide bridges		Cys + Cys	2.2	0

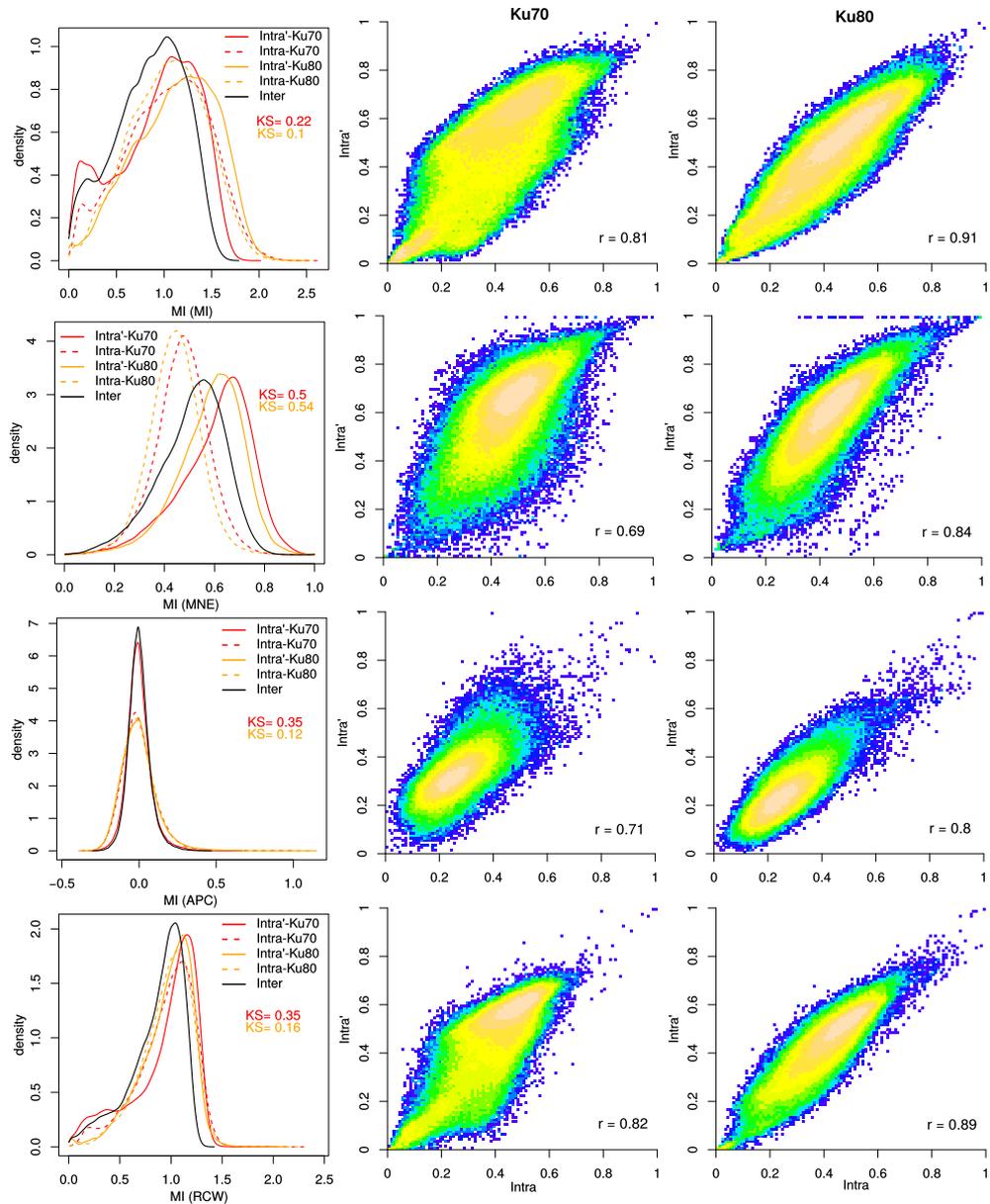


**Figure A.2:** Binary contact map constructed from the interactions between Ku70 and Ku80 observed in the crystal structure (PDB code: 1JEQ). Interactions are defined as described in appendix Table A.10. In the crystal structure, 548 residues are present of Ku70 and 520 of Ku80. Black and white positions mark residue pairs forming an interaction or not, respectively.

## Comparison of Alignment Combination Approaches

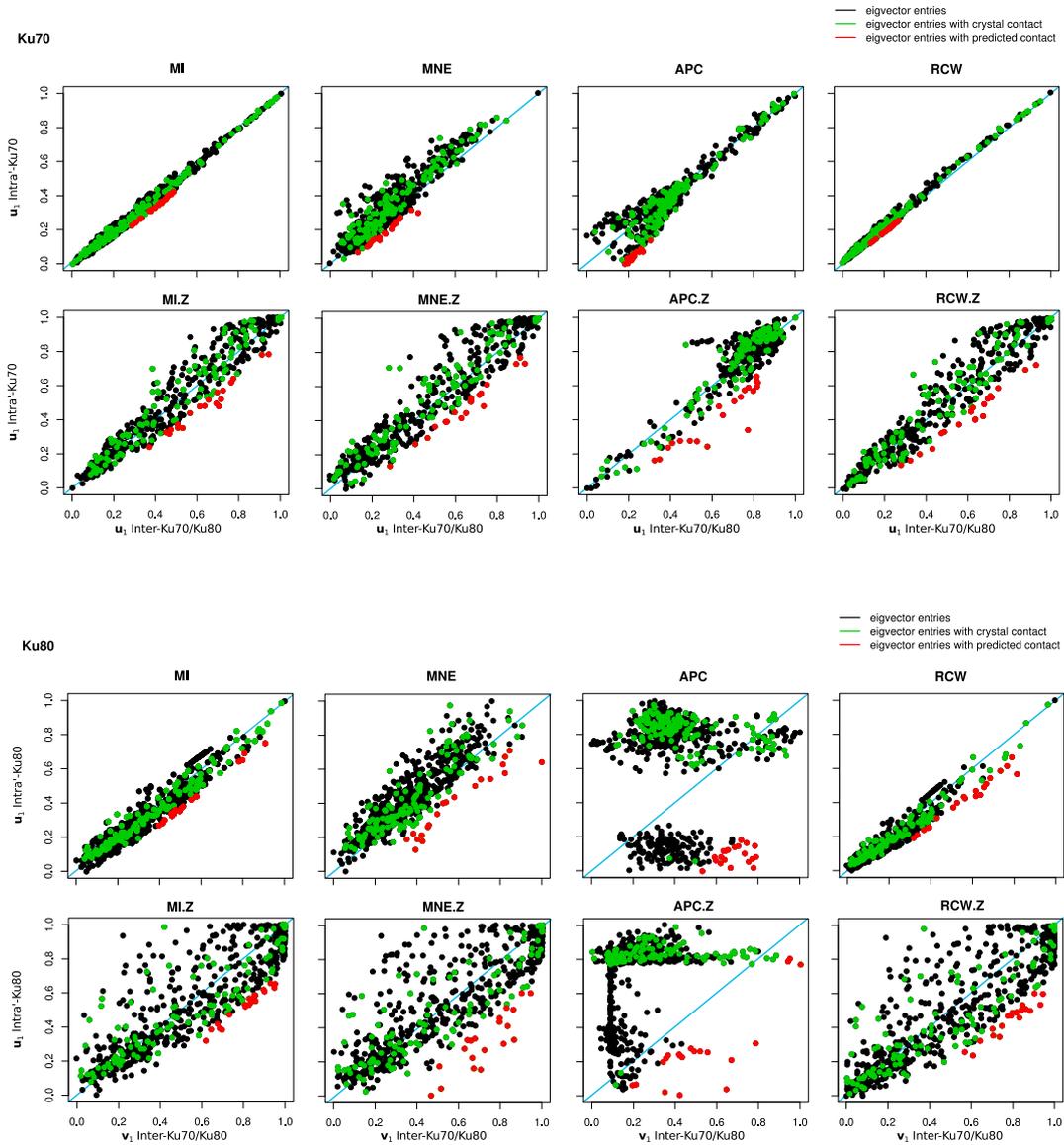


**Figure A.3:** Evaluation of the influence of the clustering alignment combination approach on MI. The Intra<sup>2</sup>-MI matrices after alignment combination are compared to the Intra-MI matrices before alignment combination. Left: MI value distributions from different matrices for Ku70 (red), Ku80 (yellow) and Ku70/Ku80 (black) are shown. Middle and right: Two-dimensional density plots with a color-scale (blue to yellow) exhibit the correlation between all matrix entries of Ku70 (middle) and Ku80 (right). The Pearson correlation coefficients  $r$  are indicated. For more information see Figure 3.9.



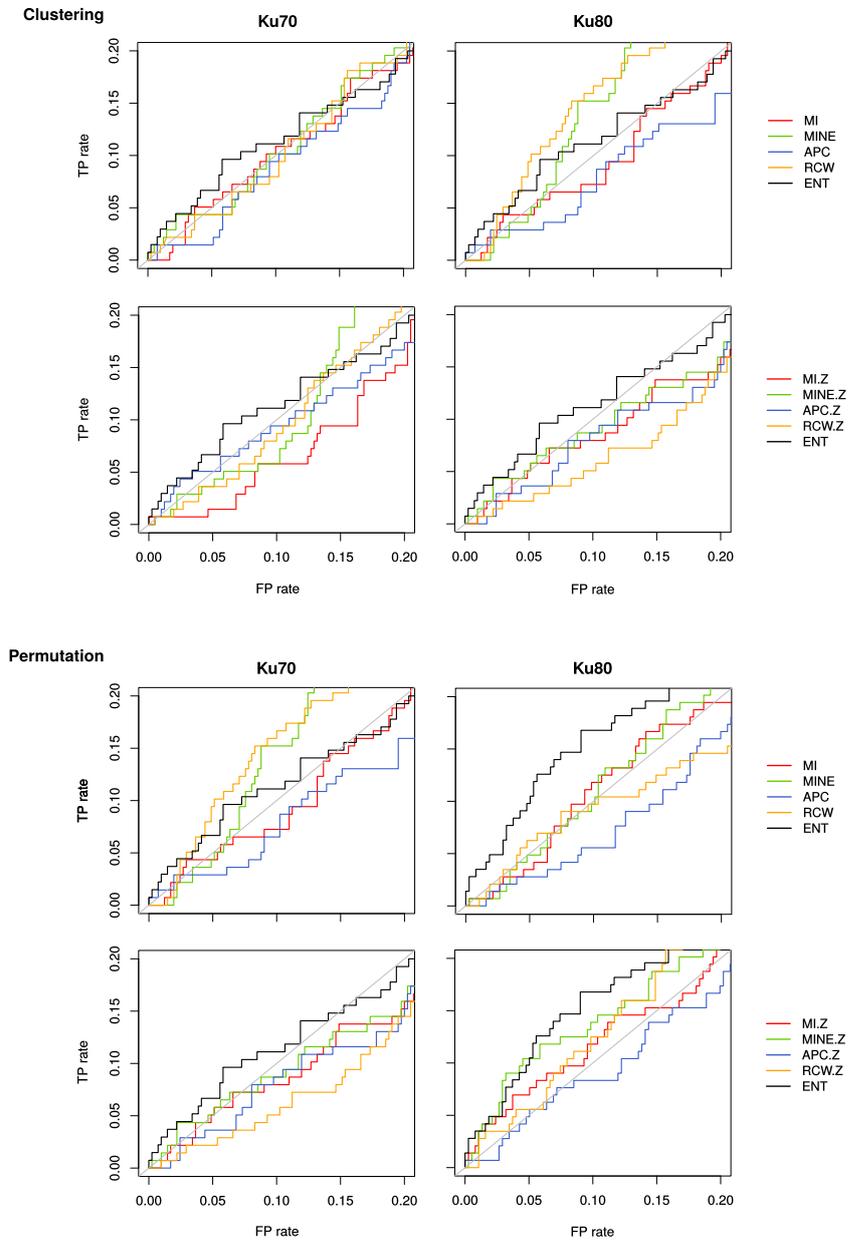
**Figure A.4:** Evaluation of the influence of the **permutation alignment combination approach** on MI. The Intra'-MI matrices after alignment combination are compared to the Intra-MI matrices before alignment combination. Left: MI value distributions from different matrices for Ku70 (red), Ku80 (yellow) and Ku70/Ku80 (black) are shown. Middle and right: Two-dimensional density plots with a color-scale (blue to yellow) exhibit the correlation between all matrix entries of Ku70 (middle) and Ku80 (right). The Pearson correlation coefficients  $r$  are indicated. For more information see Figure 3.9.

## Eigenvector Correlation: Inter- vs. Intra-MI Matrices



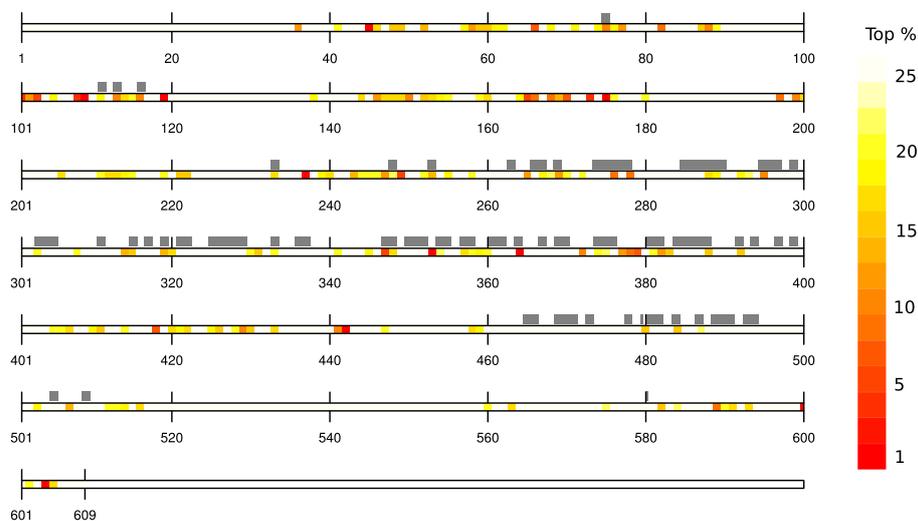
**Figure A.5:** Comparison of Inter- vs. Intra-MI signals for the **permutation approach**. The scatter plot shows the correlation of first eigenvectors. Upper panels: For Ku70, first left eigenvector  $u_1$  of the Ku70 Intra'-MI matrix and first left eigenvector  $u_1$  of the Ku70/Ku80 Inter-MI matrix. Lower panels: For Ku80, first left eigenvector  $u_1$  of the Ku80 Intra'-MI matrix and first right eigenvector  $v_1$  of the Ku70/Ku80 Inter-MI matrix. Eigenvectors were normalized to the range from 0 to 1. Red dots represent the top 20 data points being most distant to the diagonal (blue). Green dots represent the entirety of the remaining 295 contacts found in the Ku70/Ku80 crystal structure (PDB code: 1JEQ, see appendix Table A.10).

## Closeup of ROC Curves (Approach I)

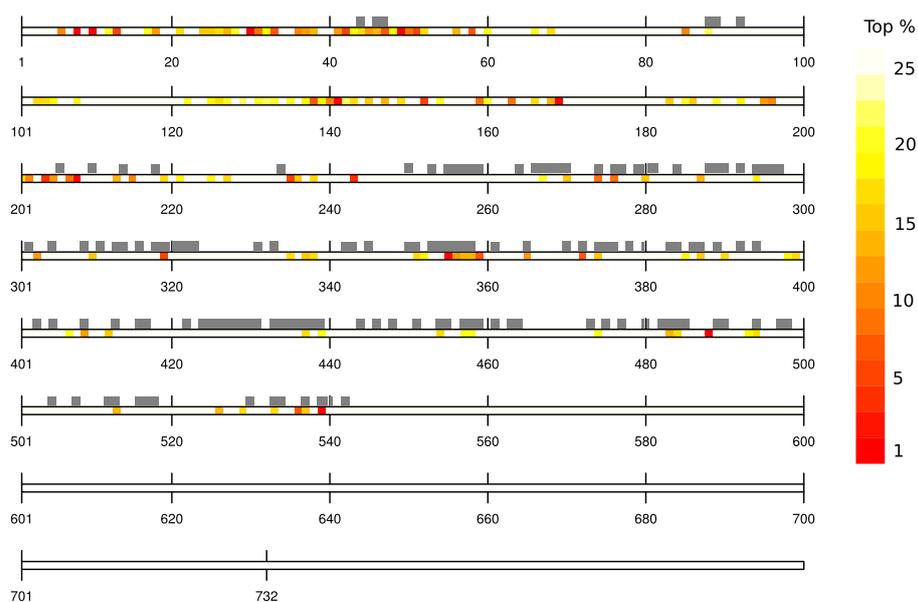


**Figure A.6:** Closeup ROC curves of Figure 3.13. Performance evaluation of **interacting residue prediction** for Ku70 (left) and Ku80 (right) separately using approach II. Different MI methods are evaluated for both alignment combination approaches (clustering and permutation): raw MI (MI), the three normalization variants (MNE, APC, RCW), their Z-score constrained counterparts (MI.Z, MNE.Z, APC.Z and RCW.Z) and, for comparison, the alignment column entropies (ENT). TP: true positives, FP: false positives.

## Linear Visualization of Predicted Top Residues



(a) Ku70 to Ku80

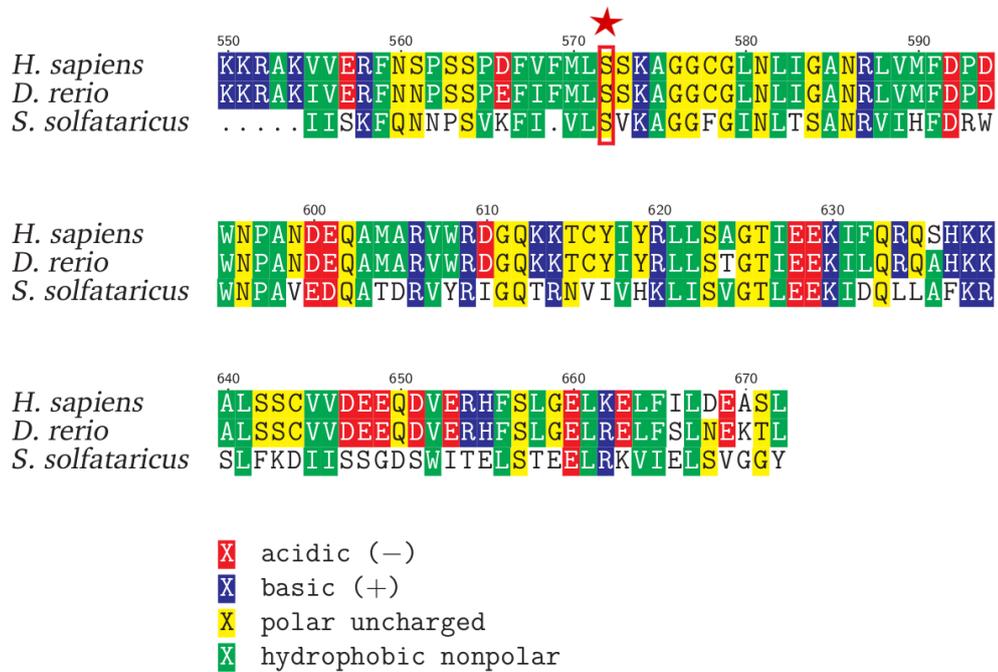


(b) Ku80 to Ku70

**Figure A.7:** Location of predicted interacting residues along the residue positions using approach II in a) Ku70 with respect to Ku80 and in b) Ku80 with respect to Ku70. Predicted residues are selected according to the top 1-25% quantiles (colors range from red to yellow) of Inter-MI values using RCW.ZS normalization and the permutation alignment approach. Interactions observed in crystal structure (PDB code: 1JEQ) according to the interaction criterion in appendix Table A.10 are colored gray.

## Rad54 Multiple Sequence Alignment

## Domain 2



**Figure A.8:** Multiple sequence alignment of human, zebrafish and archaea Rad54 sequences with alignment positions 550 to 672. Serine 572 that is phosphorylated in human Rad54 activation is equivalent to Ser566 in zebrafish *D. rerio* and Ser806 in archaea *S. solfataricus*. The phosphorylation site is marked by a red star.

# List of Abbreviations

aa	amino acid
ANM	anisotropic network model
APC	average product correction
APC.Z	Z-score constrained APC
ATP	adenosine triphosphate
AUC	area under curve
BLAST	basic local alignment search tool
BLOSUM	block substitution matrix
bp	base pair
BrAAP	branched amino acid preferring
CHARMM	Chemistry at Harvard Macromolecular Dynamics
ChTL	chymotrypsin-like
CL	caspase-like
CP	core particle
CTR	C-terminal region
DNA	deoxyribonucleic acid
DNA-PK	DNA-dependent protein kinase
DNA-PKcs	DNA-PK catalytic subunit
DSB	double-strand break
dsDNA	double-stranded DNA
EM	electron microscopy
FP	false positives
fs	femtosecond
GROMACS	Groningen Machine for Chemical Simulations
HD	helical domain
HR	homologous recombination
IC <sub>50</sub>	half maximal inhibitory concentration
<i>in silico</i>	<i>using computational methods</i>
<i>in vitro</i>	<i>using laboratory methods</i>
<i>in vivo</i>	<i>within a living organism</i>
IR	ionizing radiation
JS	Jensen-Shannon (divergence)
kDa	kilodalton
KL	Kullback-Leibler (divergence)
Ku	<i>initials K.U. of scleroderma patient</i>
LGA	Lamarckian genetic algorithm
LRT	linear response theory

MD	molecular dynamics
MI	mutual information
MI.Z	Z-score constrained MI
MJ	Miyazawa-Jernigan (interaction parameters)
MNE	<i>normalized to the minimum column entropy</i>
MNE.Z	Z-score constrained MNE
MOE	Molecular Operating Environment
MSA	multiple sequence alignment
NAMD	Nanoscale Molecular Dynamics
NCBI	National Center for Biotechnology Information
NEK	never in mitosis gene A-related protein kinase
NHEJ	non-homologous end joining
NMA	normal mode analysis
ns	nanosecond
Ntn	N-terminal nucleophile
NTR	N-terminal region
P1-P3	<i>primed substrate / inhibitor residue</i>
P1'-P3'	<i>non-primed substrate / inhibitor residue</i>
PDB	Protein Data Bank
PES	potential energy surface
PI3K	phosphatidylinositol-3-OH kinase
PIKK	PI3K-related kinase
RCW	row column weighting
RCW.Z	Z-score constrained RCW
RMSD	root mean square deviation
RMSF	root mean square fluctuation
ROC	receiver operating characteristic
RP	regulatory particle
S1-S3	<i>non-primed proteasome substrate binding pockets</i>
S1'-S3'	<i>primed proteasome substrate binding pockets</i>
SA	simulated annealing
SAR	structure-activity relationship
SASA	solvent-accessible surface area
SAXS	small angle X-ray scattering
SNAAP	small neutral amino acid preferring
STUN	stochastic tunneling
SVD	singular value decomposition
SVL	scientific vector language
SWI/SNF	switch/sucrose non-fermentable
TL	trypsin-like
TP	true positives
UPS	ubiquitin-proteasome system
VDW	van der Waals
VMD	Visual Molecular Dynamics

---

## List of Amino Acids

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine



# Bibliography

- [1] Allen C, Halbrook J and Nickoloff JA (2003). Interactive competition between homologous recombination and non-homologous end joining. *Mol Cancer Res* 1(12):913–20
- [2] Allinger NL, Yuh YH and Lii JH (1989). Molecular mechanics. The MM3 force field for hydrocarbons (part 1). *J Am Chem Soc* 111(23):8551–8566
- [3] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215(3):403–10
- [4] Amm I, Sommer T and Wolf DH (2014). Protein quality control and elimination of protein waste: the role of the ubiquitin-proteasome system. *Biochim Biophys Acta* 1843(1):182–96
- [5] Argyriou AA, Iconomou G and Kalofonos HP (2008). Bortezomib-induced peripheral neuropathy in multiple myeloma: a comprehensive review of the literature. *Blood* 112(5):1593–9
- [6] Asai A, Tsujita T, Sharma SV, Yamashita Y, Akinaga S, Funakoshi M, Kobayashi H and Mizukami T (2004). A new structural class of proteasome inhibitors identified by microbial screening using yeast-based assay. *Biochem Pharmacol* 67(2):227–234
- [7] Atilgan A, Durell S and Jernigan R (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 80(1):505–15
- [8] Bahar I (2010). On the functional significance of soft modes predicted by coarse-grained models for membrane proteins. *J Gen Physiol* 135(6):563–73
- [9] Bahar I, Atilgan A and Erman B (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des* 2(3):173–81
- [10] Berendsen H and Postma J (1981). Interaction models for water in relation to protein hydration. *Intermol Forces* 331–338
- [11] Berkers CR, Verdoes M, Lichtman E, Fiebiger E, Kessler BM, Anderson KC, Ploegh HL, Ovaa H and Galardy PJ (2005). Activity probe for *in vivo* profiling of the specificity of proteasome inhibitor bortezomib. *Nat Methods* 2(5):357–62

- [12] Bernier J, Hall EJ and Giaccia A (2004). Radiation oncology: a century of achievements. *Nat Rev Cancer* 4(9):737–47
- [13] de Bettignies G and Coux O (2010). Proteasome inhibitors: Dozens of molecules and still counting. *Biochimie* 92(11):1530–45
- [14] Blackburn C, Gigstad KM, Hales P, Garcia K, Jones M, Bruzzese FJ, Barrett C, Liu JX, Soucy Ta, Sappal DS, Bump N, Olhava EJ, Fleming P, Dick LR, Tsu C, Sintchak MD and Blank JL (2010). Characterization of a new series of non-covalent proteasome inhibitors with exquisite potency and selectivity for the 20S beta5-subunit. *Biochem J* 430(3):461–76
- [15] Blunt T, Finnie NJ, Taccioli GE, Smith GC, Demengeot J, Gottlieb TM, Mizuta R, Varghese AJ, Alt FW, Jeggo PA and Jackson SP (1995). Defective DNA-dependent protein kinase activity is linked to V(D)J recombination and DNA repair defects associated with the murine scid mutation. *Cell* 80(5):813–23
- [16] Boba P, Weil P, Hoffgaard F and Hamacher K (2010). Co-evolution in HIV enzymes. *BIOINFORMATICS2010* 39–47
- [17] Borissenko L and Groll M (2007). 20S proteasome and its inhibitors: crystallographic knowledge for drug development. *Chem Rev* 107(3):687–717
- [18] Born M and Oppenheimer R (1927). Zur Quantentheorie der Molekeln. *Ann Phys* 389(20):457–484
- [19] Bortolato A, Fanton M, Mason JS and Moro S (2013). Molecular docking methodologies. *Methods Mol Biol* 924:339–60
- [20] Bouchaert P, Guerif S, Debiais C, Irani J and Fromont G (2012). DNA-PKcs expression predicts response to radiotherapy in prostate cancer. *Int J Radiat Oncol Biol Phys* 84(5):1179–85
- [21] Braun HA, Umbreen S, Groll M, Kuckelkorn U, Mlynarczuk I, Wigand ME, Drung I, Kloetzel PM and Schmidt B (2005). Tripeptide mimetics inhibit the 20 S proteasome by covalent bonding to the active threonines. *J Biol Chem* 280(31):28394–401
- [22] Brooks B and Karplus M (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA* 80(21):6571–5
- [23] Bugreev DV, Rossi MJ and Mazin AV (2010). Cooperation of RAD51 and RAD54 in regression of a model replication fork. *Nucleic Acids Res* 39(6):2153–2164
- [24] Burger L and van Nimwegen E (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 6(1):e1000633

- 
- [25] Buslje CM, Santos J, Delfino JM and Nielsen M (2009). Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* 25(9):1125–31
- [26] Byfield JE (1989). 5-Fluorouracil radiation sensitization—a brief review. *Invest New Drugs* 7(1):111–6
- [27] Cardozo C and Kohanski R (1998). Altered properties of the branched chain amino acid-preferring activity contribute to increased cleavages after branched chain residues by the “immunoproteasome”. *J Biol Chem* 273(27):16764–16770
- [28] Carter T, Vancurová I, Sun I, Lou W and DeLeon S (1990). A DNA-activated protein kinase from HeLa cell nuclei. *Mol Cell Biol* 10(12):6460–71
- [29] Chatterjee S, Dunn D, Mallya S and Ator Ma (1999). P'-extended  $\alpha$ -ketoamide inhibitors of proteasome. *Bioorg Med Chem Lett* 9(17):2603–6
- [30] Codoñer FM and Fares Ma (2008). Why should we care about molecular coevolution? *Evol Bioinform Online* 4:29–38
- [31] conflexdock.svl (2013). *Scientific vector language (SVL) source code provided by the Chemical Computing Group Inc.* 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.
- [32] Coulomb CA (1785). Premier-[troisième] mémoire sur l'électricité et le magnétisme. Académie Royale des sciences
- [33] Coux O, Tanaka K and Goldberg AL (1996). Structure and functions of the 20S and 26S proteasomes. *Annu Rev Biochem* 65:801–47
- [34] Cron KR, Zhu K, Kushwaha DS, Hsieh G, Merzon D, Rameseder J, Chen CC, D'Andrea AD and Kozono D (2013). Proteasome inhibitors block DNA repair and radiosensitize non-small cell lung cancer. *PLoS One* 8(9):e73710
- [35] Curtin NJ (2013). Inhibiting the DNA damage response as a therapeutic manoeuvre in cancer. *Br J Pharmacol* 169(8):1745–65
- [36] Darwin C (1862). On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects: And on the Good Effects of Intercrossing. John Murray
- [37] Davidson D, Amrein L, Panasci L and Aloyz R (2013). Small Molecules, Inhibitors of DNA-PK, Targeting DNA Repair, and Beyond. *Front Pharmacol* 4(January):5

- [38] Deakyne JS, Huang F, Negri J, Tolliday N, Cocklin S and Mazin AV (2013). Analysis of the activities of RAD54, a SWI2/SNF2 protein, using a specific small-molecule inhibitor. *J Biol Chem* 288(44):31567–80
- [39] Ditzel L, Huber R, Mann K, Heinemeyer W, Wolf DH and Groll M (1998). Conformational constraints for protein self-cleavage in the proteasome. *J Mol Biol* 279(5):1187–91
- [40] Dobbins SE, Lesk VI and Sternberg MJE (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci USA* 105(30):10390–5
- [41] Dobbs TA, Tainer JA and Lees-Miller SP (2010). A structural model for regulation of NHEJ by DNA-PKcs autophosphorylation. *DNA Repair (Amst)* 9(12):1307–14
- [42] Downs JA and Jackson SP (2004). A means to a DNA end: the many roles of Ku. *Nat Rev Mol Cell Biol* 5(5):367–78
- [43] Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J and Kollman P (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24(16):1999–2012
- [44] Dunn SD, Wahl LM and Gloor GB (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333–40
- [45] Dürr H, Körner C, Müller M, Hickmann V and Hopfner KP (2005). X-ray structures of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core and its complex with DNA. *Cell* 121(3):363–73
- [46] Dynan WS and Yoo S (1998). Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. *Nucleic Acids Res* 26(7):1551–9
- [47] Endres D and Schindelin J (2003). A new metric for probability distributions. *IEEE Trans Inf Theory* 49(7):1858–1860
- [48] Essers J, van Steeg H, de Wit J, Swagemakers SM, Vermeij M, Hoeijmakers JH and Kanaar R (2000). Homologous and non-homologous recombination differentially affect DNA damage repair in mice. *EMBO J* 19(7):1703–10
- [49] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin D, Forman D and Bray F. GLOBOCAN 2012 version 1.0. Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer, Lyon, France. <http://globocan.iarc.fr> accessed on Jul 20, 2014

- 
- [50] Flechsig H, Popp D and Mikhailov AS (2011). In silico investigation of conformational motions in superfamily 2 helicase proteins. *PLoS One* 6(7):e21809
- [51] Fodor AA and Aldrich RW (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56(2):211–21
- [52] Frit P, Canitrot Y, Muller C, Foray N, Calsou P, Marangoni E, Bourhis J and Salles B (1999). Cross-resistance to ionizing radiation in a murine leukemic cell line resistant to cis-dichlorodiammineplatinum(II): role of Ku autoantigen. *Mol Pharmacol* 56(1):141–6
- [53] Game J and Mortimer R (1974). A genetic study of X-ray sensitive mutants in yeast. *Mutat Res Mol Mech Mutagen* 24(3):281–292
- [54] Gerstein M and Krebs W (1998). A database of macromolecular motions. *Nucleic Acids Res* 26(18):4280–90
- [55] Geurink PP, van der Linden WA, Mirabella AC, Gallastegui N, de Bruin G, Blom AEM, Voges MJ, Mock ED, Florea BI, van der Marel GA, Driessen C, van der Stelt M, Groll M, Overkleeft HS and Kisselev AF (2013). Incorporation of Non-natural Amino Acids Improves Cell Permeability and Potency of Specific Inhibitors of Proteasome Trypsin-like Sites. *J Med Chem* 56(3):1262–75
- [56] Gloor GB, Martin LC, Wahl LM and Dunn SD (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44(19):7156–65
- [57] Goktas S, Baran Y, Ural AU, Yazici S, Aydur E, Basal S, Avcu F, Pekel A, Dirican B and Beyzadeoglu M (2010). Proteasome inhibitor bortezomib increases radiation sensitivity in androgen independent human prostate cancer cells. *Urology* 75(4):793–8
- [58] Gomes M, Hamer R, Reinert G and Deane CM (2012). Mutual information and variants for protein domain-domain contact prediction. *BMC Res Notes* 5:472
- [59] Gottlieb TM and Jackson SP (1993). The DNA-dependent protein kinase: requirement for DNA ends and association with Ku antigen. *Cell* 72(1):131–42
- [60] Gouveia-Oliveira R and Pedersen AG (2007). Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol* 2:12

- [61] Gräwert MA, Gallastegui N, Stein M, Schmidt B, Kloetzel PM, Huber R and Groll M (2011). Elucidation of the  $\alpha$ -keto-aldehyde binding mechanism: a lead structure motif for proteasome inhibition. *Angew Chem Int Ed Engl* 50(2):542–544
- [62] Groll M (2003). Biochemistry, TU München. Co-crystal structure of the yeast 20S proteasome in complex with BSc2118 inhibitor. *unpublished data*
- [63] Groll M, Bajorek M, Köhler A, Moroder L, Rubin DM, Huber R, Glickman MH and Finley D (2000). A gated channel into the proteasome core particle. *Nat Struct Biol* 7(11):1062–1067
- [64] Groll M, Berkers CR, Ploegh HL and Ovaia H (2006). Crystal structure of the boronic acid-based proteasome inhibitor bortezomib in complex with the yeast 20S proteasome. *Structure* 14(3):451–456
- [65] Groll M, Ditzel L, Löwe J, Stock D, Bochtler M, Bartunik HD and Huber R (1997). Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* 386(6624):463–71
- [66] Groll M, Heinemeyer W, Jäger S, Ullrich T, Bochtler M, Wolf DH and Huber R (1999). The catalytic sites of 20S proteasomes and their role in subunit maturation: a mutational and crystallographic study. *Proc Natl Acad Sci USA* 96(20):10976–83
- [67] Groll M, Huber R and Potts BCM (2006). Crystal structures of Salinosporamide A (NPI-0052) and B (NPI-0047) in complex with the 20S proteasome reveal important consequences of  $\beta$ -lactone ring opening and a mechanism for irreversible binding. *J Am Chem Soc* 128(15):5136–41
- [68] Groll M, Kim KB, Kairies N, Huber R and Crews CM (2000). Crystal Structure of Epoxomicin: 20S Proteasome Reveals a Molecular Basis for Selectivity of  $\alpha'$ , $\beta'$ -Epoxyketone Proteasome Inhibitors. *J Am Chem Soc* 122(6):1237–1238
- [69] Groll M, Larionov OV, Huber R and de Meijere A (2006). Inhibitor-binding mode of homobelactosin C to proteasomes: new insights into class I MHC ligand generation. *Proc Natl Acad Sci USA* 103(12):4576–4579
- [70] Groll M, Nazif T, Huber R and Bogyo M (2002). Probing structural determinants distal to the site of hydrolysis that control substrate specificity of the 20S proteasome. *Chem Biol* 9(5):655–62
- [71] Grubbé EH (1933). Priority in the therapeutic use of X-rays. *Radiology* 21(2):156–162
- [72] Hall EJ (2012). Radiobiology for the Radiologist. 5th ed., vol. 224. Medical Dept., Harper & Row Hagerstown, Md. ISBN 0061410772

- 
- [73] Hamacher K (2008). Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. *Gene* 422(1-2):30–6
- [74] Hamacher K and McCammon JA (2006). Computing the Amino Acid Specificity of Fluctuations in Biomolecular Systems. *J Chem Theory Comput* 2(3):873–878
- [75] Hammel M, Yu Y, Mahaney BL, Cai B, Ye R, Phipps BM, Rambo RP, Hura GL, Pelikan M, So S, Abolfath RM, Chen DJ, Lees-Miller SP and Tainer Ja (2010). Ku and DNA-dependent protein kinase dynamic conformations and assembly regulate DNA binding and the initial non-homologous end joining complex. *J Biol Chem* 285(2):1414–23
- [76] Hammett L (1937). The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc* 343(1936)
- [77] Harris MN, Medrek TJ, Golomb FM, Gumport SL, Postel aH and Wright JC (1965). Chemotherapy With Streptonigrin in Advanced Cancer. *Cancer* 18:49–57
- [78] Harris R, Esposito D, Sankar A, Maman JD, Hinks JA, Pearl LH and Driscoll PC (2004). The 3D Solution Structure of the C-terminal Region of Ku86 (Ku86CTR). *J Mol Biol* 335(2):573–582
- [79] Havas K, Flaus A, Phelan M, Kingston R, Wade P, Lilley DM and Owen-Hughes T (2000). Generation of superhelical torsion by ATP-dependent chromatin remodeling activities. *Cell* 103(7):1133–42
- [80] Heinemeyer W, Fischer M, Krimmer T, Stachon U and Wolf DH (1997). The Active Sites of the Eukaryotic 20 S Proteasome and Their Involvement in Subunit Precursor Processing. *J Biol Chem* 272(40):25200–25209
- [81] Helleday T (2010). Homologous recombination in cancer development, treatment and development of drug resistance. *Carcinogenesis* 31(6):955–60
- [82] Henikoff S and Henikoff J (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89(November):10915–10919
- [83] Hershko A, Leshinsky E, Ganoth D and Heller H (1984). ATP-dependent degradation of ubiquitin-protein conjugates. *Proc Natl Acad Sci USA* 81(6):1619–23
- [84] Hoffgaard F (2011). Biomolecular Correlation in Physical and Sequence Space. Ph.D. thesis, TU Darmstadt
- [85] Hoffgaard F, Weil P and Hamacher K (2010). BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11:199

- [86] Holland JH (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA. ISBN 0262581116
- [87] Hosoya N and Miyagawa K (2014). Targeting DNA damage response in cancer therapy. *Cancer Sci* 105(4):370–88
- [88] Hou J, Li Z, Fang Q, Feng C, Zhang H, Guo W, Wang H, Gu G, Tian Y, Liu P, Liu R, Lin J, Shi YK, Yin Z, Shen J and Wang PG (2012). Discovery and extensive in vitro evaluations of NK-HDAC-1: a chiral histone deacetylase inhibitor as a promising lead. *J Med Chem* 55(7):3066–75
- [89] Huang F and Mazin AV (2014). Targeting the homologous recombination pathway by small molecule modulators. *Bioorg Med Chem Lett* 24(14):3006–13
- [90] Huang F, Motlekar Na, Burgwin CM, Napper AD, Diamond SL and Mazin AV (2011). Identification of specific inhibitors of human RAD51 recombinase using high-throughput screening. *ACS Chem Biol* 6(6):628–35
- [91] Huey R, Morris GM, Olson AJ and Goodsell DS (2007). A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* 28(6):1145–1152
- [92] Hug S (2013). Classical molecular dynamics in a nutshell. *Methods Mol Biol* 924:127–52
- [93] Humphrey W, Dalke A and Schulten K (1996). VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38
- [94] Ikeguchi M, Ueno J and Sato M (2005). Protein structural change upon ligand binding: linear response theory. *Phys Rev Lett* 94(7):1–4
- [95] Imai J, Yashiroda H, Maruya M, Yahara I and Tanaka K (2003). Proteasomes and molecular chaperones: cellular machinery responsible for folding and destruction of unfolded proteins. *Cell Cycle* 2(6):585–90
- [96] Imajoh-Ohmi S, Kawaguchi T, Sugiyama S, Tanaka K, Omura S and Kikuchi H (1995). Lactacystin, a specific inhibitor of the proteasome, induces apoptosis in human monoblast U937 cells. *Biochem Biophys Res Commun* 217(3):1070–7
- [97] Ismail IH, Martensson S, Moshinsky D, Rice A, Tang C, Howlett A, McMahon G and Hammarsten O (2004). SU11752 inhibits the DNA-dependent protein kinase and DNA double-strand break repair resulting in ionizing radiation sensitization. *Oncogene* 23(4):873–82

- 
- [98] Jaskelioff M, Van Komen S, Krebs JE, Sung P and Peterson CL (2003). Rad54p is a chromatin remodeling enzyme required for heteroduplex DNA joint formation with chromatin. *J Biol Chem* 278(11):9212–8
- [99] Jeggo PA, Geuting V and Löbrich M (2011). The role of homologous recombination in radiation-induced double-strand break repair. *Radiother Oncol* 101(1):7–12
- [100] Jekimovs C, Bolderson E, Suraweera A, Adams M, O’Byrne KJ and Richard DJ (2014). Chemotherapeutic compounds targeting the DNA double-strand break repair pathways: the good, the bad, and the promising. *Front Oncol* 4(April):86
- [101] Jha AN, Vishveshwara S and Banavar JR (2010). Amino acid interaction preferences in proteins. *Protein Sci* 19(3):603–16
- [102] Jones JE (1924). On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc R Soc A Math Phys Eng Sci* 106(738):463–477
- [103] Jones NJ, Stewart SA and Thompson LH (1990). Biochemical and genetic analysis of the Chinese hamster mutants *irs1* and *irs2* and their comparison to cultured ataxia telangiectasia cells. *Mutagenesis* 5(1):15–23
- [104] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW and Klein ML (1983). Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926
- [105] Kane RC, Bross PF, Farrell AT and Pazdur R (2003). Velcade: U.S. FDA approval for the treatment of multiple myeloma progressing on prior therapy. *Oncologist* 8(6):508–13
- [106] Kawamura S, Unno Y, Tanaka M, Sasaki T, Yamano A, Hirokawa T, Kameda T, Asai A, Arisawa M and Shuto S (2013). Investigation of the noncovalent binding mode of covalent proteasome inhibitors around the transition state by combined use of cyclopropylic strain-based conformational restriction and computational modeling. *J Med Chem* 56(14):5829–42
- [107] Kerr JF, Wyllie AH and Currie AR (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer* 26(4):239–57
- [108] Kirkpatrick S, Gelatt CD and Vecchi MP (1983). Optimization by simulated annealing. *Science* 220(4598):671–680
- [109] Kish-Trier E and Hill CP (2013). Structural biology of the proteasome. *Annu Rev Biophys* 42(February):29–49

- [110] Kisselev AF, van der Linden WA and Overkleeft HS (2012). Proteasome inhibitors: an expanding army attacking a unique target. *Chem Biol* 19(1):99–115
- [111] Korotkov VS, Ludwig A, Larionov OV, Lygin AV, Groll M and de Meijere A (2011). Synthesis and biological activity of optimized belactosin C congeners. *Org Biomol Chem* 9(22):7791–8
- [112] Krivov GG, Shapovalov MV and Dunbrack RL (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77(4):778–95
- [113] Krüger I, Rothkamm K and Löbrich M (2004). Enhanced fidelity for rejoining radiation-induced DNA double-strand breaks in the G2 phase of Chinese hamster ovary cells. *Nucleic Acids Res* 32(9):2677–84
- [114] Kruszewski M, Wojewódzka M, Iwaneńko T, Szumiel I and Okuyama a (1998). Differential inhibitory effect of OK-1035 on DNA repair in L5178Y murine lymphoma sublines with functional or defective repair of double strand breaks. *Mutat Res* 409(1):31–6
- [115] Kubinyi H (2001). Hydrogen bonding: The last mystery in drug design. In *Pharmacokinetic Optim. Drug Res. Biol. Physicochem. Comput. Strateg.* (eds Testa, B van Waterbeemd, H Folk. G. Guy, R). Verlag Helvetica Chimica Acta, Zürich. ISBN 9783906390437
- [116] Kullback S and Leibler RA (1951). On Information and Sufficiency. *Ann Math Stat* 22(1):79–86
- [117] Larkin Ma, Blackshields G, Brown NP, Chenna R, McGettigan Pa, McWilliam H, Valentin F, Wallace IM, Wilm a, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–8
- [118] Le Tallec B, Barrault MB, Courbeyrette R, Guérois R, Marsolier-Kergoat MC and Peyroche A (2007). 20S proteasome assembly is orchestrated by two distinct pairs of chaperones in yeast and in mammals. *Mol Cell* 27(4):660–74
- [119] Lecker SH, Goldberg AL and Mitch WE (2006). Protein degradation by the ubiquitin-proteasome pathway in normal and disease states. *J Am Soc Nephrol* 17(7):1807–19
- [120] Levenshtein V (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov Phys Dokl*
- [121] Lindert S, Stewart PL and Meiler J (2013). Computational determination of the orientation of a heat repeat-like domain of DNA-PKcs. *Comput Biol Chem* 42:1–4

- 
- [122] Lomax ME, Folkes LK and O'Neill P (2013). Biological consequences of radiation-induced DNA damage: relevance to radiotherapy. *Clin Oncol (R Coll Radiol)* 25(10):578–85
- [123] Lowe J, Stock D, Jap B, Zwickl P, Baumeister W and Huber R (1995). Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 3.4 Å resolution. *Science (80- )* 268(5210):533–539
- [124] MacKerell aD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FT, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D and Karplus M (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102(18):3586–616
- [125] Martin LC, Gloor GB, Dunn SD and Wahl LM (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21(22):4116–24
- [126] Mazin AV, Mazina OM, Bugreev DV and Rossi MJ (2010). Rad54, the motor of homologous recombination. *DNA Repair (Amst)* 9(3):286–302
- [127] Mazón G, Mimitou EP and Symington LS (2010). SnapShot: Homologous recombination in DNA double-strand break repair. *Cell* 142(4):646, 646.e1
- [128] Mermershtain I and Glover JNM (2013). Structural mechanisms underlying signaling in the cellular response to DNA double strand breaks. *Mutat Res* 750(1-2):15–22
- [129] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E (1953). Equation of State Calculations by Fast Computing Machines. *J Chem Phys* 21(6):1087
- [130] Miller DS, Laszlo J, McCarty KS, Guild WR and Hochstein P (1967). Mechanism of action of streptonigrin in leukemic cells. *Cancer Res* 27(4):632–8
- [131] Mimori T, Akizuki M, Yamagata H, Inada S, Yoshida S and Homma M (1981). Characterization of a high molecular weight acidic nuclear protein recognized by autoantibodies in sera from patients with polymyositis-scleroderma overlap. *J Clin Invest* 68(3):611–20
- [132] Miyazawa S and Jernigan RL (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256(3):623–44

- [133] Molecular Operating Environment (MOE) version 2013.08. Chemical Computing Group Inc. 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.
- [134] Monti MC, Margarucci L, Riccio R, Bonfili L, Mozzicafreddo M, Eleuteri AM and Casapullo A (2014). Mechanistic insights on petrosaspongolide M inhibitory effects on immunoproteasome and autophagy. *Biochim Biophys Acta*
- [135] Monticelli L and Tieleman DP (2013). Force fields for classical molecular dynamics. *Methods Mol Biol* 924:197–213
- [136] Moore EH (1920). On the reciprocal of the general algebraic matrix. *Bull Am Math Soc* 26:394–395
- [137] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS and Olson AJ (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30(16):2785–2791
- [138] Murata S, Sasaki K, Kishimoto T, Niwa SI, Hayashi H, Takahama Y and Tanaka K (2007). Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science* 316(5829):1349–53
- [139] Nandi D, Tahiliani P, Kumar A and Chandu D (2006). The ubiquitin-proteasome system. *J Biosci* 31(1):137–55
- [140] Neudert G and Klebe G (2011). fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics* 27(7):1021–2
- [141] Nobel Media AB 2014. The Nobel Prize in Chemistry 2004. [http://www.nobelprize.org/nobel\\_prizes/chemistry/laureates/2004/](http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2004/) accessed on Jul 5, 2014
- [142] O’Boyle NM, Banck M, James Ca, Morley C, Vandermeersch T and Hutchison GR (2011). Open Babel: An open chemical toolbox. *J Cheminform* 3(1):33
- [143] Ochi T, Sibanda BL, Wu Q, Chirgadze DY, Bolanos-Garcia VM and Blundell TL (2010). Structural biology of DNA repair: spatial organisation of the multicomponent complexes of nonhomologous end joining. *J Nucleic Acids* 2010
- [144] Ochi T, Wu Q and Blundell T (2014). The spatial organization of non-homologous end joining: From bridging to end joining. *DNA Repair (Amst)* 17:98–109
- [145] Oinonen C and Rouvinen J (2000). Structural comparison of Ntn-hydrolases. *Protein Sci* 9(12):2329–37

- 
- [146] Oostenbrink C, Villa A, Mark AE and van Gunsteren WF (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* 25(13):1656–76
- [147] Orłowski M, Cardozo C and Michaud C (1993). Evidence for the presence of five distinct proteolytic components in the pituitary multicatalytic proteinase complex. Properties of two components cleaving bonds on the carboxyl side of branched chain and small neutral amino acids. *Biochemistry* 32(6):1563–72
- [148] Orłowski RZ and Kuhn DJ (2008). Proteasome inhibitors in cancer therapy: lessons from the first decade. *Clin Cancer Res* 14(6):1649–57
- [149] Pál C, Papp B and Lercher MJ (2006). An integrated view of protein evolution. *Nat Rev Genet* 7(5):337–48
- [150] Palombella VJ, Rando OJ, Goldberg AL and Maniatis T (1994). The ubiquitin-proteasome pathway is required for processing the NF- $\kappa$ B1 precursor protein and the activation of NF- $\kappa$ B. *Cell* 78(5):773–785
- [151] Petukhova G, Stratton S and Sung P (1998). Catalysis of homologous DNA pairing by yeast Rad51 and Rad54 proteins. *Nature* 393(MAY):91–94
- [152] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L and Schulten K (2005). Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–802
- [153] Ponder JW and Richards FM (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J Comput Chem* 8(7):1016–1024
- [154] Postow L (2011). Destroying the ring: Freeing DNA from Ku with ubiquitin. *FEBS Lett* 585(18):2876–82
- [155] Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B and Lindahl E (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845–54
- [156] Pruitt KD, Tatusova T and Maglott DR (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–5
- [157] Rashid MH, Mahdavi S and Kuyucak S (2013). Computational studies of marine toxins targeting ion channels. *Mar Drugs* 11(3):848–69

- [158] R: A Language and Environment for Statistical Computing version 3.1.1 (2014). R Foundation for Statistical Computing. Vienna, Austria. <http://www.r-project.org>
- [159] Richardson PG, Mitsiades C, Hideshima T and Anderson KC (2006). Bortezomib: proteasome inhibition as an effective anticancer therapy. *Annu Rev Med* 57:33–47
- [160] Rosenberg B (1971). Some biological effects of platinum compounds. *Platin Met Rev* 15(2):42–51
- [161] Rothkamm K, Kühne M, Jeggo PA and Löbrich M (2001). Radiation-induced genomic rearrangements formed by nonhomologous end-joining of DNA double-strand breaks. *Cancer Res* 61(10):3886–93
- [162] Sanner MF (1999). Python: a programming language for software integration and development. *J Mol Graph Model* 17(1):57–61
- [163] Schaftenaar G and Noordik JH (2000). Molden: a pre- and post-processing program for molecular and electronic structures. *J Comput Aided Mol Des* 14(2):123–34
- [164] Schlick T (2010). Molecular Modeling and Simulation: An Interdisciplinary Guide, vol. 21 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY. ISBN 978-1-4419-6350-5
- [165] Scholz C, Voss C, Knorr S, Kuckelkorn U, Hamacher K, Kloetzel PM and Schmidt B (2014). Paradigm caught by a mouse trap: 20S proteasome's  $\beta 5$  subunit is not chymotrypsin-like. *manuscript in preparation*
- [166] Schüttelkopf AW and van Aalten DMF (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* 60(8):1355–63
- [167] Schwartz AL and Ciechanover A (1999). The ubiquitin-proteasome pathway and pathogenesis of human diseases. *Annu Rev Med* 50:57–74
- [168] Shannon CE (1948). A Mathematical Theory of Communication. *Bell Syst Tech J* 27(3):379–423
- [169] Shorter J (1985). Die Hammett-Gleichung – und was daraus in fünfzig Jahren wurde. *Chemie in unserer Zeit* 19(6):197–208
- [170] Shrivastav M, De Haro LP and Nickoloff Ja (2008). Regulation of DNA double-strand break repair pathway choice. *Cell Res* 18(1):134–47
- [171] Sibanda BL, Chirgadze DY and Blundell TL (2010). Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats. *Nature* 463(7277):118–21

- 
- [172] Spagnolo L, Rivera-Calzada A, Pearl LH and Llorca O (2006). Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. *Mol Cell* 22(4):511–9
- [173] Spies J and Löbrich M (2013). Radiation Biology and DNA Repair, TU Darmstadt. Phosphorylation at Ser572 in human Rad54. *personal communication*
- [174] van der Spoel D and Berendsen H (1997). Molecular dynamics simulations of Leu-enkephalin in water and DMSO. *Biophys J* 72(May):2032–2041
- [175] Stadtmueller BM and Hill CP (2011). Proteasome activators. *Mol Cell* 41(1):8–19
- [176] Steel GG (1979). Terminology in the description of drug-radiation interactions. *Int J Radiat Oncol Biol Phys* 5(8):1145–50
- [177] Stein ML, Cui H, Beck P, Dubiella C, Voss C, Krüger A, Schmidt B and Groll M (2014). Systematic Comparison of Peptidic Proteasome Inhibitors Highlights the  $\alpha$ -Ketoamide Electrophile as an Auspicious Reversible Lead Motif. *Angew Chem Int Ed Engl* 53(6):1679–83
- [178] Tamulevicius P, Wang M and Iliakis G (2007). Homology-directed repair is required for the development of radioresistance during S phase: interplay between double-strand break repair and checkpoint response. *Radiat Res* 167(1):1–11
- [179] Thomä NH, Czyzewski BK, Alexeev Aa, Mazin AV, Kowalczykowski SC and Pavletich NP (2005). Structure of the SWI2/SNF2 chromatin-remodeling domain of eukaryotic Rad54. *Nat Struct Mol Biol* 12(4):350–6
- [180] Tirion M (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett* 77(9):1905–1908
- [181] Toyama B and Hetzer M (2012). Protein homeostasis: live long, won't prosper. *Nat Rev Mol Cell Biol* 2(3):130–143
- [182] Turchi JJ, Henkels KM and Zhou Y (2000). Cisplatin-DNA adducts inhibit translocation of the Ku subunits of DNA-PK. *Nucleic Acids Res* 28(23):4634–41
- [183] Van Gunsteren WF and Berendsen HJC (1988). A Leap-frog Algorithm for Stochastic Dynamics. *Mol Simul* 1(3):173–185
- [184] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I and Mackerell AD (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 31(4):671–90

- [185] Vanommeslaeghe K and MacKerell A (2012). Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J Chem Inf Model* 52(12):3144–54
- [186] Vanommeslaeghe K, Raman EP and MacKerell A (2012). Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J Chem Inf Model* 52(12):3155–68
- [187] Velankar SS, Soultanas P, Dillingham MS, Subramanya HS and Wigley DB (1999). Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* 97(1):75–84
- [188] Venkitaraman AR (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 108(2):171–82
- [189] Verlet L (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys Rev* 159(1):98–103
- [190] Voss C, Scholz C, Knorr S, Beck P, Stein M, Zall A, Kuckelkorn U, Kloetzel P, Groll M, Hamacher K and Schmidt B (2014). Phenylamides as P1'-extended Proteasome Inhibitors. *accepted in ChemMedChem*
- [191] Wachters FM, van Putten JWG, Maring JG, Zdzienicka MZ, Groen HJM and Kampinga HH (2003). Selective targeting of homologous DNA recombination repair by gemcitabine. *Int J Radiat Oncol Biol Phys* 57(2):553–62
- [192] Walker JR, Corpina Ra and Goldberg J (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* 412(6847):607–14
- [193] Wang R, Fang X, Lu Y, Yang CY and Wang S (2005). The PDBbind database: methodologies and updates. *J Med Chem* 48(12):4111–4119
- [194] Waterhouse AM, Procter JB, Martin DMA, Clamp M and Barton GJ (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–91
- [195] Waters Ca, Strande NT, Wyatt DW, Pryor JM and Ramsden Da (2014). Non-homologous end joining: A good solution for bad ends. *DNA Repair (Amst)* 17:39–51
- [196] Weil P, Hoffgaard F and Hamacher K (2009). Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Comput Biol Chem* 33(6):440–4
- [197] Weissgraeber S, Hoffgaard F and Hamacher K (2011). Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins* 79(11):3144–54

- 
- [198] Wenzel W and Hamacher K (1999). Stochastic Tunneling Approach for Global Minimization of Complex Potential Energy Landscapes. *Phys Rev Lett* 82(15):3003–3007
- [199] Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6
- [200] Wilk S and Orłowski M (1983). Evidence that pituitary cation-sensitive neutral endopeptidase is a multicatalytic protease complex. *J Neurochem* 40(3):842–9
- [201] Wilson GD, Bentzen SrM and Harari PM (2006). Biologic basis for combining drugs with radiation. *Semin Radiat Oncol* 16(1):2–9
- [202] Yoshida T, Shimizu M, Harada M, Hitaoka S and Chuman H (2012). Reassessment of Hammett  $\sigma$  as an effective parameter representing intermolecular interaction energy-links between traditional and modern QSAR approaches. *Bioorg Med Chem Lett* 22(1):124–8
- [203] Zwickl P, Grziwa A and Puehler G (1992). Primary structure of the Thermoplasma proteasome and its implications for the structure, function, and evolution of the multicatalytic proteinase. *Biochemistry* 964–972



# Danksagung

An dieser Stelle möchte ich all diejenigen erwähnen, die mich während dieser Arbeit unterstützt haben. Ein herzlicher Dank gebührt den folgenden Personen:

- meinem Doktorvater **Prof. Kay Hamacher** für die Bereitstellung des Promotionsthemas, die angenehme Arbeitsatmosphäre und das mir entgegengebrachte Vertrauen. Ein Dankeschön geht an **Prof. Gerhard Thiel** für die Bereitschaft, meine Arbeit zu begutachten.
- **Prof. Boris Schmidt** für die gute Zusammenarbeit und den daraus entstandenen Publikationen zum Proteasom. Ein besonderer Dank geht an die Mitglieder der Arbeitsgruppe **Constantin, Christoph, Binia** und **Andrea**.
- **Prof. Löbrich** und **Julian Spies** für die gute Zusammenarbeit zum Rad54 Protein.
- den Mitgliedern der **Arbeitsgruppe Groll** (TU München) für die Einladung nach München. Besonders **Eva** und **Philipp**, die uns ins Kleinwalsertal begleiteten und dort tolle Vorträge hielten.
- den Mitgliedern der **Arbeitsgruppe Kast** (TU Dortmund), die uns in Dortmund herzlich aufnahmen. Ein besonderer Dank geht an **Franzi**, die den Fortschritt meiner Arbeit damals wie heute wohlwollend begleitet.
- meinen direkten Doktorgeschwistern **Steffi, Patrick, Frank** und **Martin(ez)** und denen der jüngeren Generation **Sven** und **Michael** für wertvolle wissenschaftliche Diskussionen und freundschaftlichen Beistand. Besonders erwähnt seien hier unsere Nachwuchstalente **Tine, Ben, Paul** und **Basti** sowie die Ehemaligen **Philipp** und **Miriam**.
- unserer rasenden Hilfssekretärin **Gisela**, die Probleme erledigt, bevor sie anfallen. Unserem Admin **Martin** aka Madmin, der die Technik gewährleistet und den Laden am Laufen hält.
- den Mitgliedern des **Graduiertenkollegs Strahlenbiologie** für gemeinsame Abende im Kleinwalsertal und Trifels.
- allen meinen **Freunden** und meiner **Familie**. Ein besonderer Dank geht an **Rudolf** und **Karin** für die ausführlichen Korrekturen zu dieser Arbeit. Meiner **Oma Elfriede** und ihrem Freund **Walter** für populärwissenschaftlichen Rat. **Klaus** und seiner Familie **Elisabeth, Beate** und **Ursula** für jeglichen Beistand.



# Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht.

Darmstadt, 8.8.2014

---

Sabine Knorr

# Curriculum Vitae

Dipl.-Biol. Sabine Knorr

## Education

---

since 04/2011

### **PhD Scholarship**

DFG-funded Graduiertenkolleg (GrK 1657):  
*"Molecular and Cellular Responses to Ionizing Radiation"*

Final Degree: Dr. rer. nat.

Department of Biology  
Group of Bioinformatics & Computational Biology  
Technische Universität Darmstadt (TU Darmstadt)

09/2006 - 06/2007

### **Academic Exchange Year**

Universidad de Alcalá de Henares (UAH),  
Comunidad de Madrid, Spain

09/2003 - 06/2010

### **Study of Biology**

Final Degree: Diploma

Department of Biology  
Group of Bioinformatics & Theoretical Biology  
Technische Universität Darmstadt (TU Darmstadt)

09/2002 - 08/2003

### **Study of Media System Design**

Fachhochschule Darmstadt (FH Darmstadt)

09/1993 - 06/2002

### **Gymnasium**

Final Degree: Abitur

Schillerschule, Frankfurt am Main