Imperial College London
Department of Chemistry

Atomistic graph analysis of protein dimers in disease

Léonie Strömich

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in the Department of Chemistry
at Imperial College London, February 2022

Declaration of Originality

I hereby declare that this Thesis and all work presented within are my own. I confirm that:

- In all cases where my work is based on the work of others, this is clearly stated and referenced.
- Where I consulted the work of others, I provide clear references.
- Wherever I include work that I have previously submitted as part of a degree at Imperial College London or at any other institution, it is clearly stated.
- Where my work was in collaboration with others, it is clearly stated which contributions were made by others and which is my own work.

Léonie Strömich, February 2022

Copyright Declaration

Abstract

Proteins are fundamental components of biological processes thus, they are often termed the molecular machinery of life. They commonly form dimers, in a process that is often essential for their functionality. Given the ubiquitous nature of protein regulation, many diseases are based on malfunctioning proteins and inhibiting them by binding to the active site is a widely chosen approach in drug development. However, due to acquired resistance mechanisms or high off-target effects, the active site might not always be a viable approach. This work presents an atomistic, structural investigation of dimeric proteins in the context of major disease processes, where we provide insights into potential alternative drug targeting approaches.

In this Thesis, novel diffusion-based methods are applied to characterise the intra-structural connectivity and signalling of protein dimers. The basis of our methods is the description of proteins as atomistic, energy-weighted graphs, where every atom represents a node, and every bond or interaction is encoded as a weighted edge. These graphs facilitate the study of connectivity and signal propagation within the protein through diffusion processes on the atom (node) and bond (edge) space. Two complementary methodologies are applied here, Markov Transients and bond-to-bond propensities, which have been successfully used in the context of allosteric site detection, the study of protein-protein interactions and the investigation of allosteric signalling on an atomistic level. This work explores the extension of these methodologies onto protein dimers and presents the investigation of allosteric mechanisms in three disease-relevant study systems:

1. Estrogen receptor alpha (ERα) is a homodimer and the main driver in breast cancer (BC) development and progression. Current chemotherapies based on inhibiting ERα become ineffective when recurrent tumours develop resistance against anti-estrogens. Our methodologies validate the molecular mechanism in ERα, and we further establish the prevalent role of the dimer interface in the inhibition process.

2. The main protease (M^{pro}) of the coronavirus SARS-CoV-2 is essential for virus replication in an early step of the viral life cycle. Since the beginning of 2020, we have seen this virus causing a global pandemic of COVID-19, with over 285 million cases of infection and over 5.5 million deaths by the end of 2021. To aid in combating COVID-19, we predict

highly connected allosteric hotspots and provide insights into how the disruption of the obligatory M^{pro} dimerisation presents a fruitful approach.

3. Cyclin-dependent kinases (CDKs) 4 and 6 are two essential cell cycle regulators that are often associated with cancer development, and in BC, their inhibition is part of an effective combinatorial treatment. This work contributes to understanding their activation process in complex with D-type cyclins and sheds light on the differential inhibitor patterns seen for CDKs.

By exploring these three systems with atomistic graph analysis, we describe intra-complex communication essential for activation in all three proteins. We further present implications for the respective dimer interface connectivities and how they could be a fruitful drug target. We conclude that ERα, the SARS-CoV-2 M^{pro} and CDK4/6 can be disrupted over allosteric mechanisms that include their dimer interfaces. These results provide scope for targeted drug development and provide a valuable contribution to the ongoing efforts to find efficient treatments for BC and COVID-19.

Acknowledgements

First and foremost, I want to express my gratitude to Sophia. Thank you for the opportunity to write this Thesis in your group and for guiding me throughout while still supporting me in following my own ideas. You often knew how to motivate and inspire me when I did not know myself. Further, I would like to thank Simak for his co-supervision on the estrogen receptor project and for believing in the power of computational methods. Our discussions were always fascinating, and your constructive suggestions were very appreciated.

A massive thank you to the former and present members of the Yaliraki group. Our interdisciplinary approach to science made for a beautiful working environment, and I appreciate all the input you provided. I am especially grateful to have worked with Sophia, Florian and Nan. Thanks for bouncing ideas back and forth to progress my work and, equally as important, all the walks, lunches, coffee breaks and Friday drinks. This extends to the fantastic people in the Chemistry department and particularly the computational office, especially Andrew, Megan, Sophie and Tamzin. You rock! Francesca and Maeve, I am grateful beyond words that you share your views and thoughts on literally any topic with me. You inspire me every day.

My thanks also include the Wellcome Trust for providing me with a scholarship that allowed me to pursue my research for the past four years. And even more importantly, for making me a part of the best PhD cohort anyone could have ever wished for. Heather, Jonathan, Maddy and Tara, I could not have done it without you, and I will forever be grateful we met and made it through this adventure together.

Last but not least, I am grateful for the ones who have been by my side since always and will be forever: Carina, Martin, my parents, Bernhard and Kerstin, and my sister Leslie. Knowing that you have my back carried me through, and I cannot thank you enough.

Dedication

This Thesis is dedicated to all the women who were part of my journey into and through academia. Whether you were a mentor or a fellow student, whether you guided me or you walked alongside me.

You deserve the space.

Sabine, Antonia, Nadine, Andrea, Nina, Sarah, Rebecca, Sarah, Kristin, Andrea, Elke, Susanne, Annette, Andrea, Katinka, Susi, Kathrin, Sophia, Isabel, Jasmin, Julia, Claudia, Jana, Sophie, Annika, Jana, Hannah, Sheena, Rita, Steffi, Chloé, Kristin, Sanja, Elli, Jana, Ann-Kathrin, Zhenni, Heike, Ulrike, Irmgard, Damjana, Rebecca, Kathrin, Dorothea, Lisa, Imme, Cornelia, Anne, Luisa, Monika, Eva, Julia, Becky, Kathrin, Kristina, Laura, Louisa, Anna, Fidi, Meike, Mareike, Wenke, Isabelle, Linda, Lea, Eva, Carolin, Julia, Claudia, Marie, Eva, Laura, Jenny, Claudia, Lisa, Fanny, Katharina, Bianca, Monica, Selina, Amy, Emily, Sarah, Bianca, Simone, Sabrina, Andrea, Astrid, Elke, Britta, Hannah, Maria, Alessandra, Valentina, Barbara, Qi, Lina, Aino, Sophia, Sabina, Eva, Natalie, Alessia, Megan, Lisi, Olivia, Sylvia, Sophia, Malica, Emily, Elodie, Karina, Sofia, Yixuan, Ching Ching, Eunjin, Claudia, Yara, Sophie, Patricia, Tamzin, Megan, Laura, Stefanie, Polly, Gitu, Vanessa, Aileen, Chloe, Hannah, Kim, Julia, Astrid, Sophie, Irene, Mikkaela, Shukai, Louise, Anna, Emma, Raya, Milia, Hannah, Michaela, Heather, Maddy, Tara, Maeve & Francesca.

Contents

List of Tables

List of Figures

List of Abbreviations

$^{+}$ssRNA	positive sense, single-stranded RNA
AF-1	transcription activation function 1
AF-2	transcription activation function 2
AI	artificial intelligence
AP-1	activator protein 1
ASD	Allosteric Database
ATP	adenosine triphosphate
BagPype	Biochemical atomistic graph construction software in Python for proteins etc
BBP	bond-to-bond propensity
BC	breast cancer
CADD	computer-aided drug design
CAK	CDK-activating kinase
CASP	Critical Assessment of Structure Prediction
CDK	cyclin-dependent kinase
CI	confidence interval
Cip	CDK interacting protein
COVID-19	coronavirus disease 2019
cryo-EM	cryogenic electron microscopy
DBD	DNA-binding domain
EBI	European Bioinformatics Institute
ENM	elastic network model

ER	endoplasmic reticulum
ERα	estrogen receptor α
ERβ	estrogen receptor β
ERE	estrogen response element
EST	17β-estradiol
FAQ	frequently asked questions
FDA	United States Food and Drug Administration
FIRST	Floppy Inclusions and Rigid Substructure Topology
G-loop	glycine-rich loop
GNM	gaussian network model
H12	helix 12
INK4	inhibitors of CDK4
Kip	kinase inhibitory protein
KNF	Koshland-Nemethy-Filmer
LBD	ligand-binding domain
M$^{\mathbf{pro}}$	main protease
MD	molecular dynamics
MERS	Middle East respiratory syndrome
MERS-CoV	Middle East respiratory syndrome coronavirus
ML	machine learning
MT	Markov transient
MWC	Monod-Wyman-Changeux
NHR	nuclear hormone receptor
NMA	normal mode analysis
NMR	nuclear magnetic resonance
NR	nuclear receptor
OHT	4-hydroxytamoxifen
p21	protein 21
p27	protein 27

PDB	Protein Data Bank
PIN	protein interaction network
PPI	protein-protein interaction
pRB	retinoblastoma protein
PROTAC	proteolysis-targeting chimera
QR	quantile regression
QS	quantile score
QSAR	quantitative structure-activity relationship
RdRp	RNA-dependent RNA-polymerase
RMSD	root mean square deviation
RMST	relaxed minimum spanning tree
RRIN	residue-residue interaction network
RSK4	ribosomal protein S6 kinase 4
SARS	severe acute respiratory syndrome
SARS-CoV	severe acute respiratory syndrome coronavirus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SASA	solvent-accessible surface area
SBSMMA	Structure-Based Statistical Mechanical Model of Allostery
SCA	statistical coupling analysis
SERCA	selective estrogen receptor covalent antagonist
SERD	selective estrogen receptor degrader/downregulator
SERM	selective estrogen receptor modulator
Sp-1	specificity protein 1
SRC	steroid receptor co-activator
SVM	support vector machine
TF	transcription factor
WHO	World Health Organization

Chapter 1

Introduction

1.1 Proteins - the molecular machinery of life

Proteins are involved in almost every task cells fulfil to sustain life. Based on their involvement across all areas of cellular function, they are often described as the molecular machinery of our cells. The complement of all proteins in an organism is called the *proteome* and the *interactome* describes how proteins interact to regulate and uphold biological function. The complexity of interactions between proteins is also mirrored in the scales of interactions within protein structures. The tertiary structure of proteins is formed by functional units called domains. Domains are made up of secondary structural elements like α-helices and β-sheets. Another zoom-in leads to the primary sequence of proteins: a chain of amino acids. These building blocks of proteins are formed by atoms, their chemical bonds and physical attractions and repulsions*. Ultimately, the synergy of these interactions on all scales dictates protein function and allows to maintain biological processes in an organism[2].

The plethora of roles proteins fulfil in the cellular environment are made possible over three main mechanistic principles:

- Proteins can bind to small molecules, which modulate their activity. These so-called

*For further insights into the basic principles of protein shape, structure and function, we refer the interested reader to Albert's "Molecular Biology of the Cell", Chapter 3[1].

ligand-modulated proteins are responsible for conferring the effects of small molecules. Once the respective ligand is bound, the protein shifts into an active conformation triggering downstream effects. An example of a ligand-activated protein is included in Chapter 4. Estrogen receptor α (ERα) binds estradiol hormones and triggers gene expression in target tissues[3].

- Proteins can catalyse chemical reactions. These proteins are called *enzymes*, and they facilitate the turnover of substrates into products[4]. Chapter 5 focuses on an enzyme of the protease class, which catalyses the reaction needed to break amino acid bonds in proteins. The M^{pro} of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the cleavage of viral poly peptides[5].

- Proteins can bind to other proteins. These *protein-protein interactions (PPIs)* are often required for communication steps in signalling or metabolic pathways[1]. One prevalent example is the interactions between different kinase and cyclin proteins which modulate the cell cycle phases[6]. Chapter 6 focuses on cyclin-dependent kinases (CDKs) 4 and 6, which interact with D-type cyclins to drive the G1 phase of the cell cycle.

It is worth noting here that these mechanisms are not mutually exclusive. For example, a multienzyme complex is formed by several proteins to allow the enzymatic turnover of substrates, as seen for the pyruvate dehydrogenase complex in the aerobic metabolism in eukaryotic mitochondria[7]. Similarly, proteins can require two input signals to reach an active state, e.g. binding of a ligand and a protein partner. This is often seen in receptor proteins[8].

1.1.1 Proteins in disease

We established above that protein-regulated processes are necessary to maintain cellular life. Consequently, it also follows that the dysregulation of a protein can lead to adverse effects and sometimes cumulate in a disease phenotype. The mechanisms through which proteins are implicated in disease are as manifold as their cellular functions. For diseases that can be traced back to a single protein, it is often point mutations changing a single amino acid residue that

lead to an altered behaviour. For example, the disease cystic fibrosis occurs when the cystic fibrosis transmembrane conductor is mutated[9]. The impaired protein can no longer fulfil its function in calcium transport, and the result is an impaired water transport on epithelial cells and the build-up of mucus[9].

However, many diseases have a more complicated network of protein signalling that underly their manifestation. Often it is alterations off PPIs that result in physiological effects of a disease[10]. Especially in cancer, protein-protein signalling pathways are dysregulated[11]. Further, Jubb et al.[12] described that mutations in PPI interfaces are linked to a wide range of genetic diseases and are involved in resistance mechanisms.

1.1.2 Proteins as therapeutic targets

Targeting proteins to overcome a certain disease is a fruitful approach, once a protein target is identified. The question then is, what is the best way to interact with the protein to recover its original function? Depending on the molecular mechanism that leads to the disease phenotype, a recovering interaction can either be inhibitory or activating.

A general approach is to use the wild-type function of the studied protein as a blueprint to develop targeting strategies. By investigating the natural ligand of a protein, scientists can develop competitive compounds that bind in a similar fashion but do not allow subsequent protein activity. These competitive inhibitors often bind at the orthosteric, also termed active site, of a protein where the natural ligand would also bind. Active sites usually are well-defined pockets, often located at the core of the protein[4]. Historically, the majority of approved drugs ($\sim 70\,\%$) target four types of protein families: protein kinases, ion channels, G-protein coupled receptors and nuclear hormone receptors[13]. For these protein families the active site binding modes are well defined.

However, targeting orthosteric sites does come with certain challenges. They are often highly conserved between proteins that fulfil related functions. Hence, targeting one active site might lead to off-target effects in other proteins[14]. This low selectivity is especially problematic

in closely related protein families like kinases[15] or G-protein coupled receptors[16]. Further, prolonged exposure to drugs binding at active sites often leads to acquired resistance in proteins to evade inhibition[17]. To overcome these limitations and open a wider chemical search space, two alternative targeting approaches have gained traction in drug development: disrupting protein dimerisation and allosteric modulation. They are described below.

1.2 Protein activity relying on dimerisation

As discussed above, proteins often function together with interaction partners. In the simplest form of a PPI, two proteins come together to form a dimeric assembly, as shown in Figure 1.1. Dimers can either be formed by two different proteins (= heterodimer, Fig. 1.1A) or by two copies of the same protein chain (= homodimer, Fig. 1.1B). For many dimers the interaction between the proteins is essential for functionality as seen in nuclear receptors[18] and in the cell cycle[19].

1.2.1 Interruption of dimerisation as drug targeting approach

Given a protein for which dimerisation is essential, the disruption of this process would mean inhibition can be achieved. This concept opens up an avenue for alternative drug targeting that does not involve the active site. Dimer interactions can be understood as a specific PPI class which exclusively involves two proteins. Homodimeric interfaces can be further distinguished and they tend to be larger and show less polarity than PPI interfaces formed between two different protein partners[20]. The study of PPIs, in general, has attracted more and more attention for drug design as they are fundamentally involved in many biological processes[21], and these concepts can be extended onto dimeric complexes.

To this end, two routes of inhibitory molecules are thinkable, as schematically shown in Figure 1.2. The dimerisation process can be disrupted when another agent binds with a higher affinity towards the monomer than the actual binding partner. This can either be a small

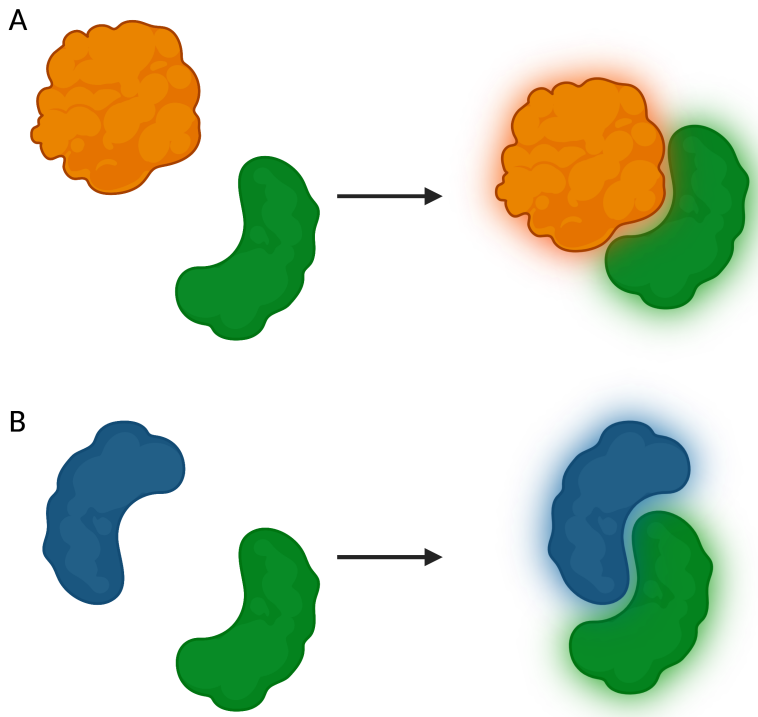


Figure 1.1: Schematic representation of dimeric protein interactions. Monomeric proteins come together to form active protein complexes, as indicated by a glow around the structure. These can be two different proteins for a heterodimeric assembly (**A**) or two copies of the same protein for a homodimeric complex (**B**).*

molecule inhibitor or a larger peptide molecule.

In the case of a small molecule inhibitor, a localised binding event would occupy an area essential for dimerisation[22,23], thereby preventing the assembly of the complex and thus activity (Fig. 1.2 left). A requirement for this approach would be to identify the specific residues in the interface that form a targetable and structurally-important area[24].

Another approach would be to develop larger molecules that mimic a PPI at the dimer interface[25] (Fig. 1.2 right). The design of these molecules is often given by the structure of the dimer interface itself: by replicating helices of the binding partner but engineering higher affinities, a blueprint is laid out. These peptide inhibitors offer the advantage of high specificity for their target as they can be modelled over a larger area, leading to fewer off-target effects[26].

*Created with biorender.com

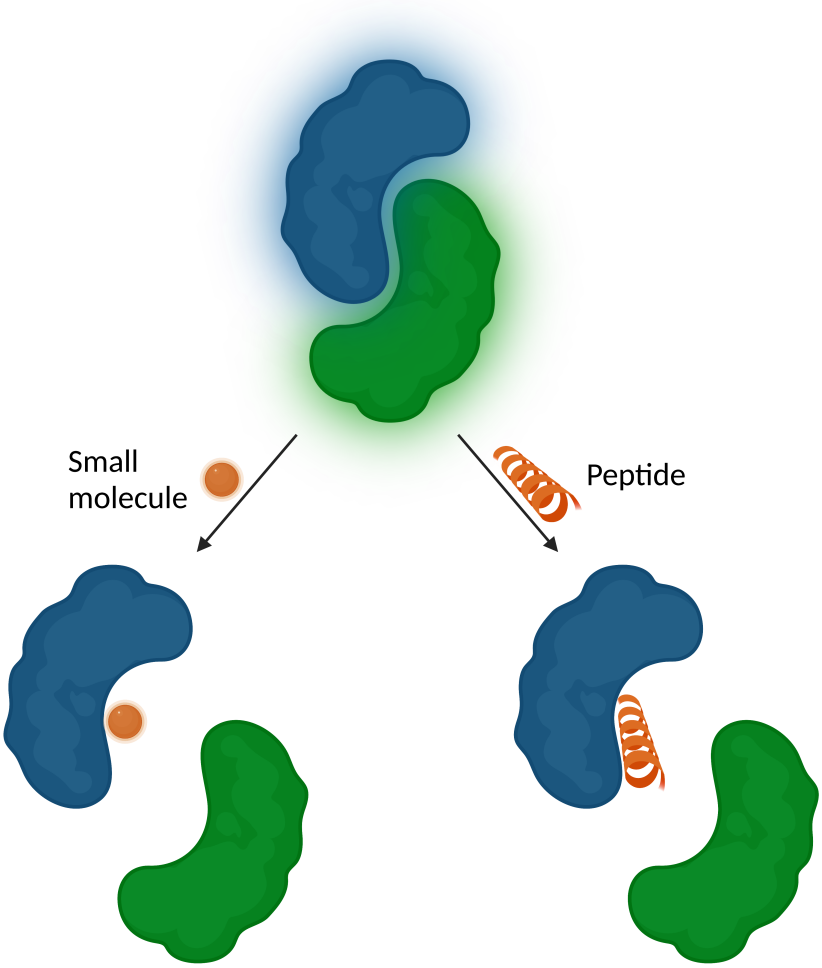


Figure 1.2: Schematic representation of dimerisation disruption. A dimeric protein with the two monomeric halves shown in blue and green and a glow indicating protein activity. The dimerisation can be disrupted by a small molecule or a peptide binding at the dimer interface.*

1.3 Allostery in proteins

Another concept that allows alternative drug targeting is allostery. The term has been coined as early as 1961[27] and describes the modulation of a protein at sites distal from the active site. Since then, different models of allostery have been developed to describe the functional modulation of proteins by distal effectors. The original descriptions of allostery were based on conformational changes based on cooperative binding of molecules. The Monod-Wyman-Changeux (MWC) model[28] considers a *concerted* conformational shift of the protein in two states, from inactive to active. In contrast, the Koshland-Nemethy-Filmer (KNF) model[29]

*Created with biorender.com

states that allosteric binding leads to a *sequential* subunit transition to the active state. Almost 20 years later, Cooper and Dryden[30] proposed a *dynamic-driven* model of allostery that did not rely on conformational changes but entropy contributions that confer allosteric effects.

The next step was the move towards conformational ensembles that represent the multiple states of a protein and where allosteric regulation based on perturbations leads to a *population shift* between two main states[31–33]. In the same year the population shift model was proposed, a *structure view* of allostery included the notion that the allostericity of a protein is encoded in its structure over intra-molecular paths[34,35]. In more recent years, there was a further extension towards an *ensemble view* of allostery[36,37] that explains proteins as existing in an ensemble of states in an energy landscape that is remodelled by perturbations like ligand-binding or protein interactions. Tsai and Nussinov[38] offered a unified view that includes the aspects of thermodynamics in two-state models, the free energy landscapes of ensembles and the structural aspects.

Coming back to the three different mechanisms that constitute protein function: ligand-mediated activity, enzymatic catalysis and protein-protein interactions (Sec. 1.1); allosteric mechanisms are mostly ligand-mediated. A molecule binds at a site distant from the active site and modulates protein activity. However, under the ensemble view, this concept can also be extended to PPIs, as the binding of a protein partner can lead to a conformational change that leads to activity.

1.3.1 Allostery as drug targeting approach

The concepts of allostery open up another alternative drug targeting approach that comes with a range of advantages. Allosteric drugs can provide a higher selectivity as allosteric sites are less conserved than active sites, and further, they allow both up and down regulating of protein function (reviewed in Wenthur *et al.*[17]). The potential of allosteric drugs is vast as protein function is fine-tuned by allosteric effects over a variety of mechanisms[38] and the number of proteins confirmed to be regulated by allostery is ever growing. It is further proposed

that the views of allostery as described above would also theoretically allow all proteins to be allosterically modulated[39].

Allosteric drug targeting is a concept that can apply to all assemblies of proteins, whether they are in monomeric or oligomeric form. Figure 1.3 is a schematic of the concepts of allosteric modulation in a protein dimer. Considering a homodimeric protein that contains one active and one allosteric site per monomer, in a symmetric modulation both allosteric sites would be

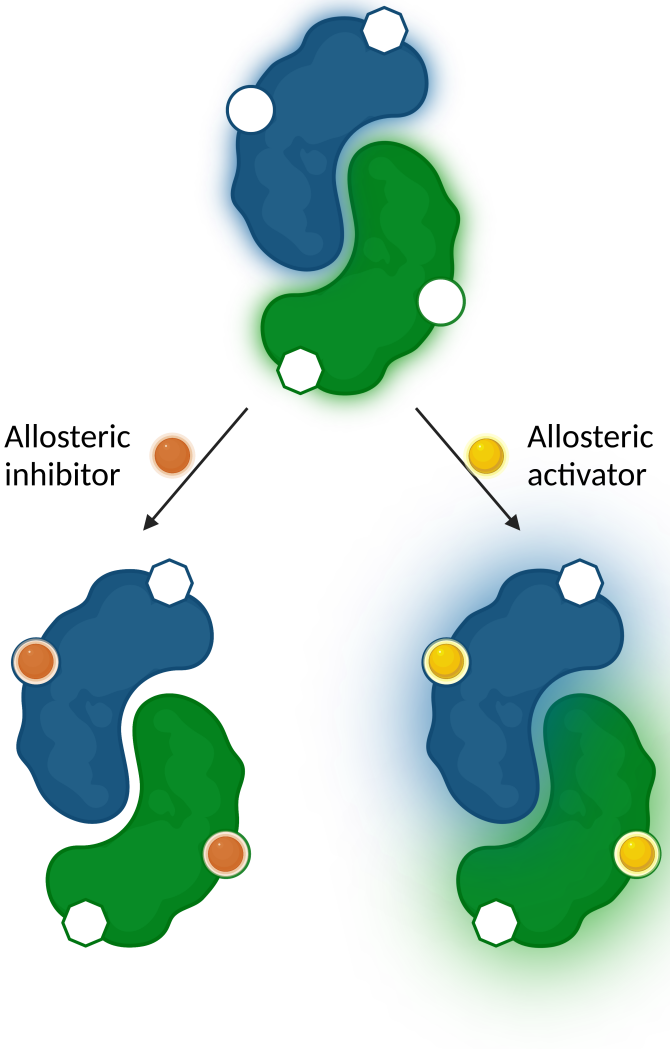


Figure 1.3: Schematic representation of allosteric modulation of a dimeric protein. A dimeric protein with two monomeric halves shown in blue and green. Two binding sites per monomer are shown: the active site as an octagon and an allosteric site as a circle. Binding at the active site can result in inhibition (orange) or activation (yellow). Activity of the dimer is indicated by a glow around the protein.*

*Created with biorender.com

occupied to lead to protein activity or inhibition*. However, for multimeric complexes, often a multi-state model of allosteric modulation is observed where the binding at one monomer can change the binding affinities for other subunits. These cooperative effects have, for example, been described for hemoglobin[40] and aspartate carbamoyltransferase[41]. Generally speaking, allosteric modulation means the binding of a molecule can either lead to inhibition or activation, and thus a variety of disease mechanisms could be targeted.

Furthermore, Gunasekaran et al.[39] suggested that allosteric effects can regulate all proteins once we discover the triggering points. However, allosteric discovery has often been serendipitous or requires time and resource-intensive high-throughput screens[42]. Hence, computational approaches can aid in efficiently elucidating allostericity in proteins, as discussed further in Section 2.3.

Lastly, we need to point out that these two alternative targeting approaches at dimer interfaces and allosteric sites are not mutually exclusive. On the contrary, it can be fruitful to apply these concepts together to tackle interrelated problems. For example, allosteric sites at a distance from the dimerisation interface could be used to overcome the ”undruggable” features of a large interface[43]. Vice versa, protein-protein interactions can be understood as a form of allosteric regulation, meaning the input over interface-binding peptides could lead to allosteric modulation of the protein activity[44].

1.4 Computational studies to facilitate drug design

We outlined above that targeting proteins to treat or overcome diseases is a fruitful endeavour. For many diseases, the in-depth study of the molecular mechanism will have identified a suitable protein target. In the next step, the aim is to identify an appropriate approach to interact with the target. Traditionally, the screening for active biomolecules against a given target and the following optimisation steps have been time and resource-intensive[45]. A considerable reduction in time and money can be expected if the number of biomolecules that have to be explored

*For the purpose of illustrating allosteric modulation, we chose to show a symmetric modulation in Figure 1.3 rather than a more complicated multi-state mechanism that can vary between proteins.

experimentally is narrowed down to promising lead compounds. This is where computational methods come in. They can present a time and resource-saving shortcut to identify, model and develop bioactive compounds against certain targets or even identify new targetable proteins[46]. The computational methods that can provide biologically meaningful insights are plentiful and are summarised in more detail in Chapter 2. They all have in common that they aim to provide biologically relevant insights and point the drug design process in the right direction in an efficient manner.

The era we live in sees major methodological advances on a regular basis. On the level of data availability, it is the advancement of methods for structure determination[47] and the subsequent deposition of more, and often also structurally more complex, proteins in the Protein Data Bank (PDB)*,[48]. In the area of computational advancements, we are seeing a continuous growth of artificial intelligence (AI) for drug discovery. One of these scientific breakthroughs was recently achieved by AlphaFold, a deep learning-based structure prediction tool that achieved the highest predication accuracy in the 2020 Critical Assessment of Structure Prediction (CASP) competition[49]. In a collaboration with the European Bioinformatics Institute (EBI), AlphaFold predictions were made publicly available for the proteomes of 21 organisms†. The integration of AI into modelling approaches, advances in computational representations of physicochemical properties and the sheer increase in computational power drive the field towards computer-driven drug design[50]. But it is also the increase in international data sharing, open access initiatives and scientific collaborations facilitated by computational means that speed up scientific discoveries.

*With over 600 depositions in January 2022 alone, the PDB now contains over 186000 structures: www.rcsb.org/stats/growth/growth-released-structures

†Available at: alphafold.ebi.ac.uk

1.5 Thesis outline - the study of protein dimers with atomistic graph analysis

The work contained in this thesis constitutes an application of computational methods based on graph theory to three protein systems in the context of diseases. We* aimed to provide insights into activation mechanisms and alternative targeting approaches on a molecular level. Indeed, we showed that dimer interactions are essential in the mechanisms of the studied proteins, and we were able to deduce which interactions can be target points for drug design. This work highlights strategies that allow to overcome future drug resistance and broadens the range of targeting approaches by including allosteric signalling and dimer interfaces.

Chapter 2 constitutes an introduction to the field of computer-aided drug design (CADD) which describes the power of computational approaches for the purpose of combating diseases. We put a particular focus on targeting approaches that are distal from the active site. Two main approaches are discussed in detail here: the prediction of allosteric sites for allosteric modulation of proteins and the study of dimer interfaces. We further discuss the graph theoretical methodologies that were developed in our group in the context of computational methods to study protein structures and highlight some advantages of our approach.

The methodologies that find application in this work are introduced in Chapter 3. We provide procedural details on how protein structures are translated into atomistic graphs and summarise the different bond and interaction types that are encoded in these protein graphs. We further introduce Markov transient and bond-to-bond propensity analysis as two methodologies that can be used to explore fast and strong connectivity within graphs and provide mathematical details for both. The Chapter also details how we post-process the obtained data to allow quantitative insights and scoring of residues and sites of interest. Finally, we describe ProteinLens, a web server that makes our atomistic graph analysis available to the community in the form of a user-friendly web application[51].

*This thesis is largely written in the first-person plural to stylistically reflect that research is always influenced and supported by the work of previous and current members of a group. However, as laid out in the Declaration of Originality, all work was done by the author except for where it is clearly stated otherwise.

In Chapter 4, we demonstrate the first application of our atomistic graph analysis on ERα. The homodimeric protein modulates the cellular response to estrogens by initiating gene expression. ERα is also one of the most studied proteins in the context of breast cancer, the leading type of cancer in women worldwide[52]. We use bond-to-bond propensity analysis to validate the molecular mechanism of the ERα ligand-binding domain (LBD). We further study the dimer interface of the ERα LBD in agonist and antagonist-bound conformation and highlight important structural features. Finally, this Chapter explores how our methodology can contribute to reveal mechanisms of chemotherapeutic resistance observed in cancer mutations of the protein.

Chapter 5 goes on to explore another homodimeric protein in the context of a highly relevant disease: coronavirus disease 2019 (COVID-19). The underlying agent is SARS-CoV-2, and M^{pro} is one of the most important drug targets in the virus. M^{pro} is a proteolytically active protein that cuts the viral polyproteins and is essential for viral replication. We describe the role of the dimer interface in conferring activity of the protein and how mutated residues regulate it on a molecular level. Using Markov transient and bond-to-bond propensity analyses, we identify four putative allosteric hotspots and explore their targetability with small fragments. This Chapter demonstrates how atomistic graph analysis can aid in identifying target points for drug development in a time-sensitive setting.

Chapter 6 sees the application of our analysis on a less studied system which allows us to explore the predictive power of our approach. The heterodimeric complexes of CDK4/6 with D-type cyclins are essential cell cycle regulators and their dysregulation is implicated in cancer growth[53]. We provide insights into the multi-factorial activation mechanism of CDK4 in complexes with D-type cyclins. Markov transient and bond-to-bond propensity analyses reveal the interplay of different activation signals in a complementary manner. Our atomistic graph analysis also sheds light on different communication paths in the dimer interface of this heterodimeric system. We further explore the applicability of our methods to study differential effects of chemotherapeutics and present the first indications of diverging inhibition mechanisms in CDK6 and CDK2.

1.6 Publications

ProteinLens, the web server described in Chapter 3 has been published in the 2021 web server issue of *Nucleic Acids Research*. The author of this Thesis holds co-first authorship of the paper with the title "ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules"[51] (Mersmann, S.F.; **Strömich, L.**; Song, F.J.; Wu, N.; Vianello, F.; Barahona, M. & Yaliraki, S.N.).

The benchmarking study of bond-to-bond propensities mentioned in Chapters 2 and 3 has been published in *Patterns* as "Prediction of allosteric sites and signaling: Insights from benchmarking datasets"[54] (Wu, N., **Strömich, L.** & Yaliraki, S.N.).

The work in Chapter 4 has been drafted as a manuscript for submission to the *Journal of Molecular Biology* with the title "Molecular mechanisms in estrogen receptor alpha using atomistic graph analysis" (**Strömich, L.**; Ali, S. & Yaliraki, S.N.).

The work in Chapter 5 has been submitted to the *Journal of Molecular Biology* as a paper titled "Allosteric hotspots in the main protease of SARS-CoV-2" (**Strömich, L.**; Wu, N.; Barahona, M. & Yaliraki, S.N.). The paper can be accessed as a preprint[55] with DOI: 10.1101/2020.11.06.369439.

Chapter 2

Computational approaches for the study of proteins

2.1 Computer-aided drug design

The pharmaceutical drug discovery process is lengthy and resource intensive. The average time for a drug from discovery to market is 12.5 years and costs around 1.15 billion GBP*. To provide a time and resource short cut in the process, computational analyses have long been integrated into the pipeline[57]. Computer-aided drug design (CADD)† describes the scientific field that combines computational chemistry with structure-based methods to facilitate a rational drug discovery process.

Traditionally, CADD has been understood as a way to identify compounds that are active against a certain target (= "hit") and further optimise these molecules until they become a "lead" compound against the target. To identify "hit" molecules, the field distinguishes between ligand-based and structure-based virtual screening approaches[58]. In ligand-based virtual screenings, molecules of known activity are classified based on chemical features and molecular

*The Pharmaceutical Journal by the Royal Pharmaceutical Society collated this data for the British pharmaceutical industry[56]: pharmaceutical-journal.com/article/feature/drug-development-the-journey-of-a-medicine-from-lab-to-shelf

†The term drug design is used for the individual steps and applications of the process, while drug discovery describes the development of a drug from beginning to end.

descriptors to develop quantitative structure-activity relationship (QSAR) models[59]. These machine learning (ML) approaches can then be used to predict which compounds in a large compound library would be active against a given target.

In a structure-based approach, a virtual screening would mean the *docking* of compound libraries into the three-dimensional structure of the protein target. The work by Kitchen et al.[60] reviewed docking approaches that simulate the binding event of a ligand to a protein structure, try to determine the best binding pose and predict binding affinities. In the context of CADD, this would encompass determining the binding pose for large numbers of compounds and ranking them by a docking score that can predict binding affinities. Docking is a widely applied technique in virtual screenings because it rapidly narrows down the search space[60]. However, it is not without pitfalls as it is highly reliant on well-curated input data and chosen parameters and results need to be treated with caution[61,62].

In further steps of the drug discovery pipeline, only the top-scoring compounds would be optimised to identify "lead" compounds. These optimisation approaches can see the application of further docking experiments as well as QSAR models or detailed molecular dynamics (MD) simulations. The latter models the dynamics of biomolecular structures by simulating the forces between interacting atoms over time. MD techniques can predict energies of target-ligand interactions and are widely applied to optimise binding affinities[63].

Ultimately, it is a combined approach of ligand-based and structure-based methods that lead to success in drug discovery and the above-described methods are often applied at various stages of the pipeline[57].

2.1.1 Advances in structure-based drug design

It is important to acknowledge that all structure-based drug design methods are reliant on the quality of the underlying structural data. However, the past decade has seen major advances in structure determination techniques that improved data quality and quantity[64]. Previously, structures deposited in the PDB were mainly determined with X-ray crystallography and nu-

clear magnetic resonance (NMR), but we now see an increase of structure determination with cryogenic electron microscopy (cryo-EM)[64]. The technique has contributed to the field by facilitating the determination of larger and more complex biomolecules at continuously increasing resolution[65].

Furthermore, the advances in ML algorithms like deep learning led to an increase in structure predictions[66]. The most notable development in that field is AlphaFold[49] which predicted unsolved structures for 21 organisms and allows the study of disease systems that have been inaccessible before as no structural data was available. Utilising computational approaches provides flexibility that allows reacting quickly to new threats as seen for COVID-19 in the past two years[67]. Given the rapid advancements in the computational fields of structural biology as well as ML methods for drug design, we can presume that CADD will continue to be a driving force in drug discovery[50].

In a traditional sense, the CADD technologies are used to find and optimise a drug that binds to the active site of a target protein. However, active or orthosteric site targeting can be challenging for closely related proteins with structurally similar binding sites as it might be difficult to find a compound that binds selectively to only one target protein. Low selectivity of a drug leads to off-target effects, which are a leading cause of, for example, cancer drugs failing at later stages in pharmaceutical development[68]. In the context of this work, we explore CADD to reach beyond the active site and discuss two subfields that focus on alternative targeting approaches distal from the active site.

2.2 Dimer interaction studies with computational methods

In Section 1.2.1, we introduced the concept of interrupting the dimerisation process for protein inhibition. We here aim to provide an overview of how computational approaches can aid in discovering how to predict and target dimers.

Although some work has been done exclusively on the level of dimer interfaces, the much larger body of work is in the realm of general protein-protein interaction (PPI) studies. The largest application of computational PPI drug design is in the prediction of PPIs and determining the networks of protein communication. An excellent review of the computational approaches for PPI prediction can be found in Keskin et al.[69]. These methodologies can broadly be categorised by the data that is processed as an input: genomic information is used in methods like gene/domain fusion[70] or gene co-expression[71], whereas many ML algorithms use amino acid sequence information[72–74]. Lu et al.[75] also acknowledged the importance of incorporating structural data when predicting PPI networks which continues to be a fruitful approach in light of the continuously growing structural resources in the PDB. Information on PPIs that is obtained from predictions or experiments is collated in a range of publicly accessible repositories[69]. STRING is one PPI database that integrates experimental and predictive data and finds frequent application[76].

The integration of structural data further allows predicting the PPI interface on an atomistic level for which Laddach et al.[77] proposed a detailed workflow. Given two proteins of interest that interact with each other, the interface can either be found in structurally annotated databases or has to be predicted using different approaches[77]. One widely used methodology is PDBePisa, which evaluates biological assemblies if a solved structure is available[78]. Another large class of methodologies that are based on structural input data revolve around protein-protein docking, as reviewed by Vakser[79]. Further methods for the characterisation of PPI interfaces make use of ML algorithms that use structural descriptors[80] or MD which allows to model how two protein partners come together in a stable conformation[81]. All of these methodologies aim to provide knowledge of the PPI interface which is needed to then develop strategies to target the interface interaction for drug discovery.

Once an interface has been defined, the next step in the drug discovery pipeline would be to elucidate a mechanism to disrupt the interface. Traditionally, PPIs have been considered "undruggable" due to their large, shallow surface that does not show deep sub-pockets for molecule binding[22,23]. To overcome these limitations, two approaches are thinkable: targeting the interface with small molecules at residue clusters relevant for the binding affinity of the

protein complex or developing larger peptide inhibitors that mimic the binding of the protein partner (Fig. 1.2.1). For both approaches, computational methods can be applied to provide insights that facilitate a rational drug design process.

The first approach has its basis in a body of work that has revealed that PPI interfaces have different regions, and some, often more conserved, residues contribute to the binding energy more significantly than others (reviewed in Moreira et al.[82]). Identifying these "hot spots" of energetically critical residues would allow targeting the interface at a much smaller region which might be more suitable for a small molecule. One experimental approach is to use systematic alanine mutations of each position in the interface and study the effect on binding energies[83]. However, these alanisation scans are resource-intensive and unsuitable for large-scale data screens; thus, computational methodologies are used (reviewed in Morrow and Zhang[84]). The methods range from *in silico* mutational scans[85,86], over knowledge-based approaches that incorporate physical properties at the residue level[87–89], to MD simulations[90,91]. Ultimately, these methodologies aim to predict which residues are the most likely to contribute to a binding event, and thus might constitute an anchor point for an effective inhibitory molecule.

Another approach for disrupting the dimerisation process would be to design peptides that bind at the interface and thus block the binding of the protein partner. To this end, computational approaches can be used in two ways: identification of a binding pose for the peptide on the interface, and the computational peptide design for a given target. The work by N. London and colleagues extended the above mentioned concepts of residue "hot spots" in the interface to "hot segments", which would be best targeted by peptides[92,93]. Another possibility is the above-mentioned docking technique, which can be extended from identifying PPI interfaces[79] onto protein-peptide interfaces. If the protein-peptide docking is done *ab initio* (without a narrowed search space), it is computationally intensive as the whole surface area of a protein would need to be considered. However, if the docking is guided towards a pre-defined PPI interface, peptide docking can be applied to identify inhibitory peptides[94]. Finally, computational methodologies find application in designing peptides against known targets where a blueprint is given by the surface structure of the binding partner[92,95,96]. Further, computational approaches can then be used to predict binding affinity and optimise the peptide inhibitor. Chakraborty

et al.[97] demonstrated this workflow for a peptide binding to the homodimeric ERα and later demonstrated its inhibitory potential[98].

2.3 Allosteric site discovery with computational methods

Going hand-in-hand with the ongoing search for a unifying model for allostericity in biomolecules is the development of computational methodologies to study allosteric processes. Given that the allosteric effect in proteins can be conferred over different mechanisms, it is difficult to develop a universally applicable prediction tool. This is mirrored by the range of computational methodologies that have been developed to study and predict allosteric behaviour, as summarised in excellent reviews over the past decade[99–101].

The newfound relevance of allosteric effects for drug design and the increase in computational power has led to the rise of computational methods to predict allosteric sites, which can be broadly classified into the following areas:

- Allosteric effects are often regulated over conserved residues, and coupling between allosteric and orthosteric sites might be encoded in co-evolution patterns[102]. Statistical coupling analysis (SCA)[103] is based on these ideas and predicts allosteric sites from multiple protein sequence alignments[104,105].

- Based on the idea that allosteric and orthosteric sites have distinct characteristics, ML approaches extract these and other structural and physicochemical features to determine allosteric sites. Successful allosteric site prediction ML algorithms have been proposed based on support vector machine (SVM)[106], random forest models[107], or Naive Bayes classifiers and artificial neural networks[108]. These approaches can predict allosteric sites with decent success rates. However, they do not allow studying of the underlying allosteric mechanism that describes the effect of a distant binding event on the active site.

- Another set of methodologies are based on structural data of a protein and aim to describe the dynamic properties of allosteric effects. A large class of these methodologies are based on MD to model the dynamics of protein allostery. MD tools are often used to describe allosteric communication paths in proteins but have also been used to deduce allosteric sites that can be used for protein inhibition[109]. However, all-atom MD approaches are still computationally heavy and thus can not cover the allosteric effects that might occur on larger scales in the protein. To elevate the computational burden, coarse-grained models which investigate proteins on the residue level are more suitable. One example would be elastic network models (ENMs) of proteins for which the dynamics are described with normal mode analysis (NMA). These methods have successfully been applied to predict allosteric sites with decent predictive accuracy up to 65 %[110,111]. Other allosteric site prediction methods have been proposed, which reflect the dynamics of allosteric proteins by simulating protein ensembles on which they studied the effect of a perturbation that could mimic a binding event[112,113].

- Many methodologies developed over the years can not be counted exclusively to one of the above-described classes but are rather integrated approaches of several methodologies (reviewed in Amamuddy et al.[114]). For example, Song et al.[115] proposed the combination of NMA with a previously developed structure-based ML algorithm[106] for allosteric site prediction. Another example of integrated approaches is a group of methods that study the correlation between orthosteric and allosteric sites based on results from MD or ENM simulations[116–118]. Xie et al.[119] recently introduced CorrSite2.0 that calculates correlations between pockets on the protein surface from ENM results and achieved 90 % prediction accuracy in allosteric proteins.

To facilitate the study of allosteric proteins, the Allosteric Database (ASD) has been developed to record experimentally confirmed allosteric proteins and their modulators[120–123]. The structural information contained in ASD also allows to assess the predictive power of algorithms for allosteric site detection. To compare the prediction accuracy between different tools, two benchmarking datasets have been curated. ASBench[124] and CASBench[125] contain a wealth

of allosteric proteins and record information on allosteric and orthosteric ligands*. To further the progress in the field of allosteric modulator discovery, some of the above-described methods have been developed as publicly accessible web servers. Section 3.2.2 provides more details on available web applications and introduces ProteinLens, a web server developed in our group[51] based on graph theoretical methods to predict protein allostery.

2.4 Graph theoretical methods to study protein structure and function

Graphs or networks† have been introduced on all levels of biological complexity, and we will briefly summarise the different scales in protein studies towards atomistic resolution as shown in Figure 2.1.

For a graph at the whole protein level, each node represents a protein. These protein interaction networks (PINs) are often used to encode the interactome to describe the network of PPIs (Fig. 2.1A). These PINs can hold information on the physical interactions and the functional linkage between proteins and they can be established for biologically diverse contexts, i.e. the interactions in different organisms[126] or the proteins involved in a metabolic pathway[127]. PINs have found applications in revealing signalling cascades in diseases[128] as well as in drug discovery[129].

The next finer-grained level is occupied by graphs built from interacting protein chains or domains[130] (Fig. 2.1B) which are, for example, used to allow the classification of assemblies in the PDB[131]. Another class of protein graphs that resides between protein and residue-level granularity is constituted from secondary structure elements like α-helices and β-sheets[132]. These graphs aim to elucidate the topology of proteins and are collated in the Protein Topology Graph Library[133,134]. Applications of these graphs are primarily in the realm of structural

*We recently published a benchmarking study of our methodologies which incorporates ASBench and CasBench[54].

†The terms graph and networks are used interchangeably in literature, but for the level of atomistic resolution (which is at the heart of this work) the term graph is more common.

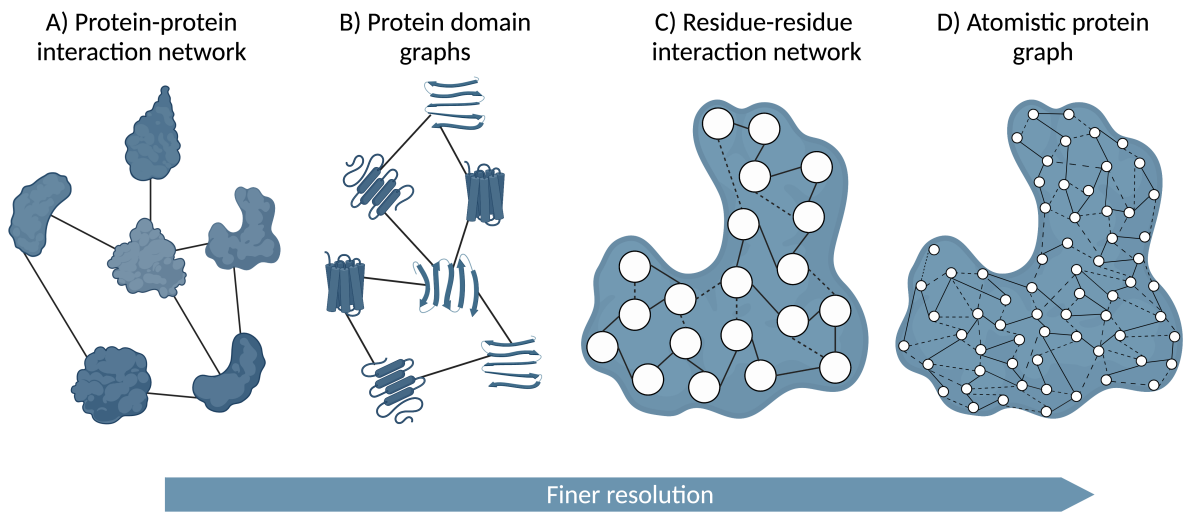


Figure 2.1: Overview of graph descriptions of proteins structures. A) Protein-protein interaction networks record the communication between different proteins. **B)** Domain networks are established from protein chains and domain data. **C)** On the next level, a protein graph can be described by a residue-residue interaction network. **D)** The most detailed description are atomistic graphs which are built by considering each all atoms in the structure and connecting them by edges that represent bonds and interactions.*

protein classification as reviewed by Koch and Schäfer[135].

The next zoom-in step brings us to the residue level where protein graphs are constructed by using every residue as a node (Fig. 2.1C). Edges of the residue-residue interaction networks (RRINs)† are inferred by different approaches. In the simplest form, RRINs are built by using distance cutoffs between Cα atoms of residues[136]. Others deduce contacts from atom interactions between residue pairs and weight the edges accordingly[137,138]. Ribeiro and Ortiz[139] found that incorporating energetic weights into RRINs is essential to correctly represent signal propagation in the protein. It has also been a commonly used approach to deduce residue-residue interactions from MD or ENM simulations and thus encode a dynamical aspect in the RRINs[140]. Compared to graphs on the atom level, RRINs still record fewer interactions and are hence considered coarse-grained. Nonetheless, they are widely applied to study protein dynamics to answer biological questions. Pacini et al.[141] constructed protein graphs on the residue level and found that the neighbourhood of each node explained protein dynamics. Brinda et al.[142] and Del Sol and O'Meara[143] studied dimeric proteins as RRINs and identified interface "hot spot" residues, which ties in with the scope of CADD in targeting dimeric

†In literature, RRINs are also termed protein contact networks or amino acid networks.

interfaces as discussed in Section 2.2.

RRINs are also an important ingredient for the study of protein allostericity. One understanding of allostery is rooted in the notion that every protein has pre-existing intrinsic pathways encoded in its residue network. Allosteric signals are transmitted through these paths depending on where the modulator binds on the protein surface[35]. By describing proteins as connected residues, it is possible to study allosteric communication and identify signalling paths within the proteins[144]. Popular approaches to study allosteric communication paths in proteins are rooted in Monte Carlo algorithms[145], NMA of ENM[146] or the Structure-Based Statistical Mechanical Model of Allostery (SBSMMA)[147]. Finally, we encounter atomistic protein graphs on the most fine-grained level of detail obtained from structural data (Fig. 2.1D).

2.4.1 Atomistic graph analysis

Atomistic protein graphs are built by considering each atom in the protein as a node and defining edges between them. Edge definition on an atomistic level lends itself to the logical step of assigning weights based on naturally occurring physical interactions and chemical bonds and their respective energy contributions. Similarly to what Ribeiro and Ortiz[139] found on the RRIN level, this can be assumed to be advantageous to model the real-life physicochemical energies within a protein. Work by Sen et al.[148] constructed ENMs on several scales of resolution and found support for the importance of atomistic detail in dynamic modelling. Amor et al.[149] showed that atomistic detail is required to detect allosteric sites and identify trigger interactions and bonds.

Jacobs et al.[150] and Thorpe et al.[151] introduced the earliest approach to model proteins as atomistic graphs, built into the Floppy Inclusions and Rigid Substructure Topology (FIRST) program. One motivation was to study dynamics in proteins over ”floppy” and ”rigid” modes while being computationally less expensive than classic MD simulations at the time. FIRST constructs a graph from covalent bonds, hydrogen bonds, salt bridges and hydrophobic interactions and can be used to study protein flexibility and folding[151,152]. A large body of work

[†]Created with biorender.com

that incorporates the FIRST algorithm is presented by the group of H. Gohlke. Over the years the group studied protein thermostability[153], flexibility[154] and allosteric signalling[155,156] at atomistic resolution. Another work that built an atomistic protein graph was presented by Veloso et al.[157] in which they identified biologically significant residue clusters in myoglobins.

The work of our group is situated in the realm of atomistic protein graphs that capture physicochemical properties of proteins with high accuracy. Originally based on the highly efficient FIRST algorithm described above, we construct protein graphs at atomistic resolution to study protein function. The graph construction and edge weighting process was first described by Delmotte et al.[158] and was refined over the years[149,159,160] to include more interaction types as described in Section 3.1.2.

We apply graph theoretical methods to model diffusion processes on these atomistic graphs to determine biological function. Two main approaches have been developed over the years and applied in the context of intra-molecular protein communication. *Markov transient* analysis is applied on the node-space of the graph which in proteins means it provides a measure for how fast each atom in the protein is reached by a signal originating at a chosen source, e.g. the active site. The method is based on the idea of a random walker in a network where the bond energies in the weighted graph represent the transition possibilities as described in detail in Section 3.1.3. This method has successfully been used to reveal allosteric sites and pathways in caspase-1[159] and aided in drug repurposing against allosteric sites in ribosomal protein S6 kinase 4 (RSK4)[161].

Another measure called *bond-to-bond propensity* models the effect of a perturbation at defined source edges on every other edge of the network. Translated onto proteins, this describes the connectivity of a source site, i.e. a ligand, with any other bond in the protein and has been found to discover allosteric sites[149]. The successful prediction of allosteric sites was initially confirmed for 19 out of 20 proteins[149] but has recently been extended onto two large benchmarking sets: ASBench[124] and CASBench[125]. We were able to predict allosteric sites for 127 of 146 proteins (407 of 432 structures) with six statistical measures that highlight different aspects of allosteric binding[54]. Bond-to-bond propensities have also been an effective

approach to studying the cooperativity of the allosteric effect in multimeric proteins[162]. In a recent study, the methodology was extended to predicting PPI sites as another form of protein modulation at distal sites[163].

Our methods are based on sparse matrix descriptions and together with algorithmic advancements, this leads to computationally inexpensive methods that retain atomistic details of protein function. Prompted by the computational efficiency, we recently made our methodologies available to the community in the form of a user-friendly, interactive web server called ProteinLens*, [51]. The web application allows to construct atomistic graphs of biomolecules and run Markov transient and bond-to-bond propensity analyses. ProteinLens provides the results in intuitive visualisations that can be accessed interactively. The user is thus able to investigate a protein of choice and discover allosteric properties like signalling paths, residues, and hotspots. Section 3.2.2 provides a detailed description of the web server and its functionalities.

2.5 Conclusions

Two main factors contribute to the increasing importance of integrating computational approaches into drug discovery. Evolving experimental approaches lead to an influx of data amounts that can only be analysed to full potential by computational means. Further, increasing computational power, as well as ML advances, allow exploring the high-dimensional chemical search space *in silico*[50]. CADD ranges from identifying viable target proteins[164] to designing high-affinity compounds against them[57]. One commonly utilised drug discovery approach is to design drugs against the orthosteric site of a target protein which can have downfalls regarding selectivity and potency in structurally related protein families[17]. To overcome the limitations of active site inhibition, PPI interfaces and allosteric sites are proposed as distant effectors of protein activity[21,165].

Especially for these alternative targeting approaches, computational guidance is of the essence as they tend to be less studied and often serendipitously detected[165]. In the context of this

*Accessible at: proteinlens.io

work, we highlight two alternative targeting mechanisms by investigating the dimer interface in dimeric proteins and studying protein allostery. Section 2.2 provided an overview of how computational methods can be exploited to study PPI interfaces and the design of inhibitors that target these interactions in dimeric proteins[23]. In Section 2.3, we summarised computational approaches detecting allosteric sites that can be used to modulate protein activity with small molecules[165].

Section 2.4 provides a detailed explanation of how graphs are used to study protein function and dynamics on different scales, introducing PPI networks, domain and structural feature networks and RRINs. The latter find wide application in the modelling of communication within proteins and have been used to find important residues in dimeric protein interfaces[142] and allosteric signalling[144]. On the most fine-grained level of protein graph modelling, we introduce atomistic protein graphs for which each node represents an atom and bonds or interactions can be represented as weighted edges[151].

The work in our group revolves around modelling biomolecules as atomistic protein graphs, which are constructed from three-dimensional structural data[158,160]. Markov transient and bond-to-bond propensity analyses are two approaches to study diffusion processes on these graphs that can be related to biological function. In several studies by our group, the atomistic graph analyses were primarily used to detect allosteric pathways and sites in small[149,159] and large protein datasets[54]. The approach has also been extended to study allostery in large multimeric protein assemblies[162] and predict PPI interfaces[163].

Chapters 4, 5 and 6 demonstrate how the information provided by atomistic graph analyses is used to study protein systems relevant to disease contexts. We apply Markov transient and bond-to-bond propensity analyses to shed light on molecular mechanisms and elucidate how protein activation is achieved. Furthermore, we show how these approaches can be utilised to discover residues and sites used for alternative targeting approaches for protein inhibition. Chapter 3 provides a more detailed explanation of the underlying methodologies and mathematical concepts. Further, it contains a description of ProteinLens, an interactive web server to study allosteric effects in proteins[51].

Chapter 3

Methodology

Our approach to studying biologically relevant concepts like allostery and protein-protein interaction signalling is based on structural information. The underlying idea is that the biological function of proteins is encoded in their three dimensional structure which allows long and short-range physical interactions to occur. To facilitate an investigation of structural data with computational means we make use of atomistic graphs which are a computationally efficient representation. All atoms in the protein are represented as nodes and all bonds or interactions as edges in the graph. The weighting of these edges is rooted in chemical and physical knowledge of bond and interaction energies and hence our protein graphs retain physicochemical information. We further apply methodologies that are based on graph diffusion processes and reveal fast and strong connectivities within the protein that are biologically meaningful. The following Sections elaborate on these concepts and highlight key features.

3.1 Atomistic graph analysis

3.1.1 Data collection and processing

All structural data was downloaded from the Protein Data Bank (PDB) in `.pdb` file format[48]. Files had to be pre-processed to ensure the right biomolecule was contained, to consider struc-

turally important water molecules, and to clean ambiguous side-chain confirmations. Detailed descriptions of the structures used for each study system and of our data processing workflow can be found in Appendix A.1.

3.1.2 Atomistic graph construction

As described above, our approach to investigate biological concepts is underpinned by computationally efficient graph representations of biomolecular structures, constructured from the cartesian coordinates that are stored in `.pdb` files obtained from the PDB[48]. The graph construction process for biomolecules has been developed in our group, and is described by S. Meliga[166] and A. Delmotte[167]. It has further been detailed in Delmotte et al.[158], Amor et al.[159] and by B. Amor[168].

On the atomistic level, constructing a *protein graph* G means to represent every atom as a *node or vertex* V and every bond or interaction between the atoms as an *edge* E. Thus we obtain a graph object $G(V, E)$ which represents the atomistic protein structure (Fig. 2.1D). By weighting the edges, we can then capture physicochemical properties of the protein like hydrogen bond strength and the hydrophobic effect[158,169]. Figure 3.1 illustrates the graph construction process. Firstly, structural data is downloaded from the PDB[48] and pre-processed (details are given in App. A.1). This processing can include the addition of hydrogen atoms with a command line tool called *Reduce*[*,170].

The next step is the edge detection, which was originally performed using the Floppy Inclusions and Rigid Substructure Topology (FIRST) algorithm[150,171]. The command line tool FIRST detects edges over distance cut-offs and chemical knowledge constraints. Finally, these edges are weighted according to bond or interaction energy. The following Sections provide an overview of all bonds and interactions that are detected and how the weights are assigned.

[*]Available at: github.com/rlabduke/reduce

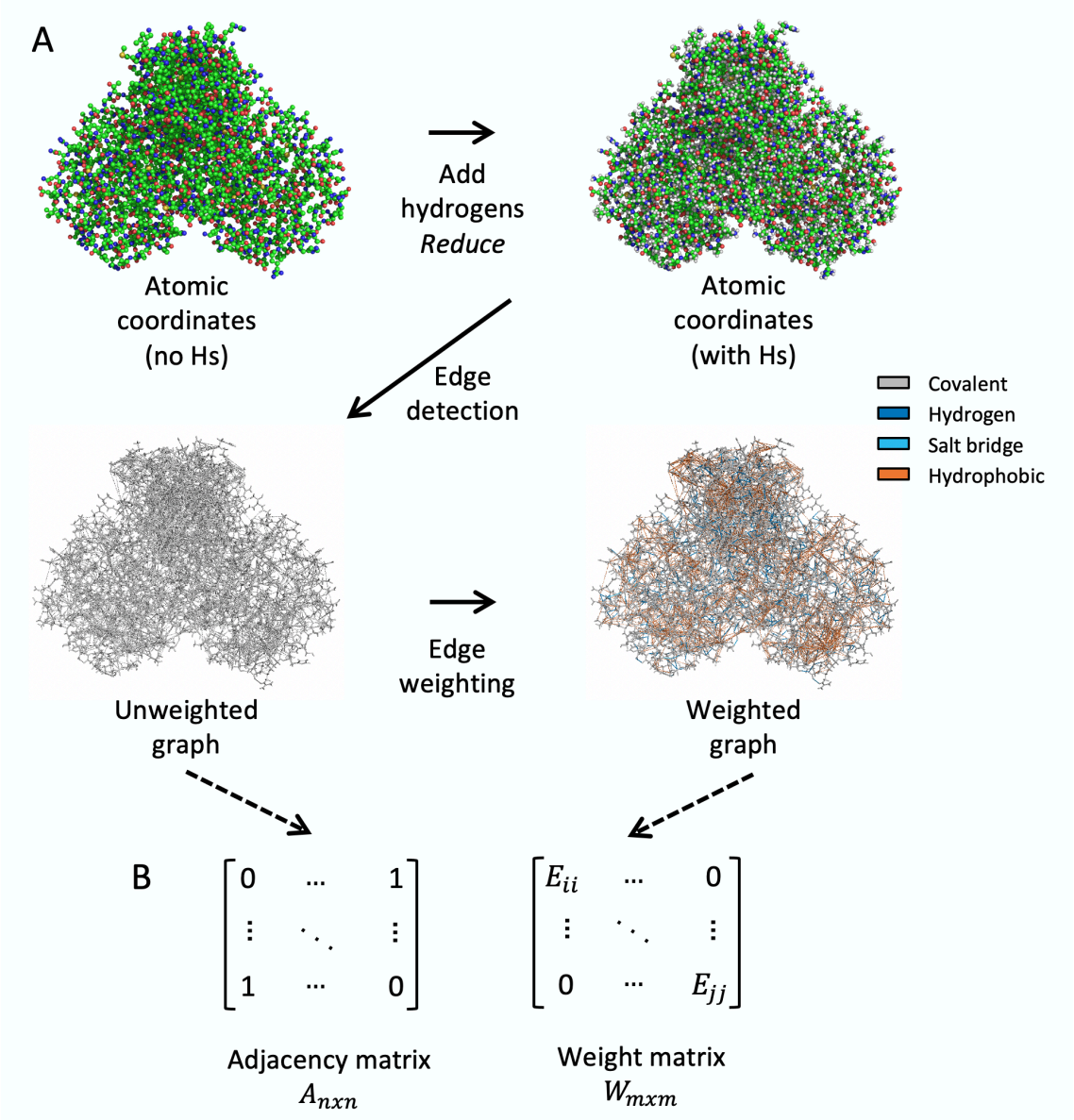


Figure 3.1: Graph construction process. A) The graph construction process, using the structure of the SARS-CoV-2 M^{pro} (PDB id: 6Y2E[5]). After the atomic coordinates are obtained from the PDB[48], hydrogen atoms are added using the command line tool *Reduce*[170]. In the next step, atoms are defined as nodes in the graph and edges are detected. These edges are then weighted according to the bond or interaction type they represent. For the given structure covalent and hydrogen bonds as well as salt bridges and hydrophobic interactions were detected and are coloured as indicated. **B)** Graphs can be described as a set of matrices where the adjacency matrix $A_{n\times n}$ indicates which nodes n are connected. The weight matrix $W_{m\times m}$ stores the interaction energy of all edges m on the diagonal. Adapted from Strömich et al.[55].

An updated graph construction

During the course of this Thesis, a new graph construction workflow was developed and released, to include new features and allow its incorporation into a single Python package. This new tool,

called Biochemical atomistic graph construction software in Python for proteins etc (BagPype) was developed in Song et al.[160] and finds further elaboration by F. Song[169]. BagPype was developed in Python[172], a hugely popular programming language in the community. BagPype refines the way edges were detected previously by FIRST and incorporates a range of distance constraints and chemically relevant characteristics[160].

Another change from the previous graph construction is the elaborate detection and weighting of hydrophobic interactions as discussed in detail by F. Song[169]. This extension aims to model the hydrophobic effect on proteins which is a many-body effect[173] and as such cannot be described by a single edge in the graph. BagPype uses a set of constraints to find potential hydrophobic interactions between atoms, weights them according to the hydrophobic potential of mean force[173] and then uses the relaxed minimum spanning tree (RMST) method[174] to sparsify the sub-graph. The sparsification step is included to reflect the hydrophobic many-body effect, whilst lowering graph complexity and preventing overly connected regions of the graph[169].

Furthermore, BagPype incorporates $\pi - \pi$ stacking interactions as well as DNA backbone interactions to describe the physicochemical synergies that occur in DNA strands. The detection of DNA interactions is partly based on previous work by Delmotte[167], and BagPype extends on that by allowing the construction of atomistic graphs for structures that contain both, proteins and DNA. This extension of the graph construction means that an even larger part of the PDB is accessible with our methodology.

In this work, we used the original graph construction based on FIRST for Chapter 4 and the updated tool BagPype for Chapters 5 and 6. The different projects presented in these chapters are self-contained entities and individual results are not directly compared across chapters. Hence, we do not believe the switch in graph construction methodology has an impact on the statements that are concluded from our results.

Weighted edges

The bond and interaction types that are included in our graph construction are listed below. Generally speaking, the algorithm determines which edges an atom can form by investigating the neighbourhood of each atom. The spatial coordinates that are found in the `.pdb` files allow the definition of distances and geometric criteria between atom pairs, which we use to inform bond/interaction types.*

- **Covalent bonds** are detected according to distance constraints and are weighted with empirical bond dissociation energies[175].

- **Hydrogen bonds** can be formed between pairs of donor and acceptor atoms that share electrostatic attractions over a proton. Their detection is based on distance and angle constraints between the atoms. The so found hydrogen bonds are then weighted according to the Mayo potential[176,177].

- **Salt bridges** can be considered as charged hydrogen bonds as they occur between positively and negatively charged atoms. Hence, their weight is determined by a modified Mayo potential[176] as applied in FIRST[150].

- **Hydrophobic interactions** model a many-body hydrophobic effect in the protein and their detection differs between graph construction processes as described above. The previous graph construction based on FIRST[150] assigns hydrophobic tethers between C-C and C-S atom pairs based on proximity if their van der Waals' radii are within 2 Å. BagPype extends on these constraints with a general distance cutoff of 9 Å to later sparsify the weighted hydrophobic edges using the RMST method[174]. Both approaches assign weights by applying the hydrophobic potential of mean force[173]. But where the previous process included only two values corresponding to the valleys in the potential, BagPype uses the continuous potential to assign bond energies.

*For completeness, the full list of edge types that are encoded by BagPype is given here. However, for the protein structures in this work, only the following bond types are detected: covalent bonds, hydrogen bonds, salt bridges and hydrophobic interactions.

- **Electrostatic interactions** are often especially important between small molecules like ligands and the main protein chains. We get information on these interactions from the LINK entries in PDB files and weight them according to a Coulomb potential as defined by Gilson and Honig[178], where atom charges are obtained from the OPLS-AA force field[179].

- **π-π stacking interactions** are found between two aromatic rings and play an important role in DNA structural stability. The edges are assigned using an energetic threshold and weighted by combining van der Waals and electrostatic contributions as modelled by Hunter and Sanders[180] and Warshel et al.[181].

- **DNA backbone interactions** are modelled via edges placed between consecutive nucleotides and are weighted as electrostatic interactions between the phosphate groups as described in detail in Delmotte[167].

Protein graphs

Following the workflow described above, we obtain a protein graph $G(V, E)$ with a set of nodes V and edges E which represent the atoms and bonds/interactions, respectively. These protein graphs can be represented as a set of matrices with n nodes and m edges that contain relevant information.

- The *adjacency matrix* A is an $n \times n$ matrix with entries indicating whether two nodes are connected in the graph ($a_{ij} = 1$) or not ($a_{ij} = 0$). For a protein graph, the adjacency matrix indicates whether two atoms are connected as determined in the graph construction process (Fig. 3.1B). The adjacency matrix can also be weighted as A^w where the entries are w_{ij} if two nodes are connected and 0 otherwise.

- The *weight matrix* W is an $m \times m$ diagonal matrix where the entries contain the edge weights which represent bond and interaction energies for a protein graph (Fig. 3.1B). This makes the weight matrix the protein equivalent to the conductance matrix G for electrical grids defined by Schaub et al.[182].

- The *incidence matrix* $B_{n\times m}$ provides information on which edge correlates with which node. If a node i is incident on an edge b, B_{bi} is recorded as 1 and otherwise as 0.

- The *degree matrix* $D_{n\times n}$ is a diagonal matrix that describes the degree of each node in the graph. The degree is the number of all edges attached to a node.

- The combinatorial *Laplacian matrix* $L_{n\times n}$ is a matrix to describe protein graph dynamics[137,183] and is defined as $L = D - A$. For the weighted *adjacency matrix* A^w it follows:

$$L = \begin{cases} -w_{ij}, & i \neq j. \\ \sum_j w_{ij}, & i = j. \end{cases} \tag{3.1}$$

Representing biomolecules as graphs that can be described with matrices is advantageous as matrices are mathematically efficient objects that allow for a lowered computational cost. Additionally, due to the primarily local nature of bonds and interactions within proteins, our protein matrices are sparse, which also decreases algorithmic running time. Our group further draws from graph theoretical concepts to model biological processes on graphs of biomolecules[149,158,159,162,163,166,168,184–188]. Two of these methodologies that find application in this work for the purpose of studying protein systems in disease are described in more detail below.

3.1.3 Markov Transients

The analysis of graphs with Markov processes allows us to assess the dynamics of a system across all scales, represented by different Markov times. This approach can be applied in an unsupervised manner to detect communities in graphs and understand their intrinsic structure and organisation[184,185]. In the case of protein graphs, these concepts revealed that community organisation within proteins is found on multiple scales and protein dynamics are governed by an interplay of partitions across these scales[158,186,188]. Leading on from there, Amor et al.[159] explored the application of Markovian random processes starting from a pre-defined source.

This random walk on a protein graph has been found to be able to reveal allosteric sites and pathways in caspase-1 [159] and ribosomal protein S6 kinase 4 (RSK4) [161].

Figure 3.2 provides a schematic of the so-called Markov transient (MT) analysis and shows how a random walker would be sourced on a protein from e.g. the atoms in an active site. For every *Markov time step* t the random walk can be described as:

$$p_{t+1} = p_t T \tag{3.2}$$

Here T is the *Markov transient matrix* and every entry T_{ij} is the probability for the random walk to transition from node i to node j in one time step. The vector p_t gives the probabilities of the current state at each node at a given time step. This process allows the modelling of signal propagation through a protein where the weights of bonds or interactions are encoded in the Markov transient matrix as $T = D^{-1}A^w$.

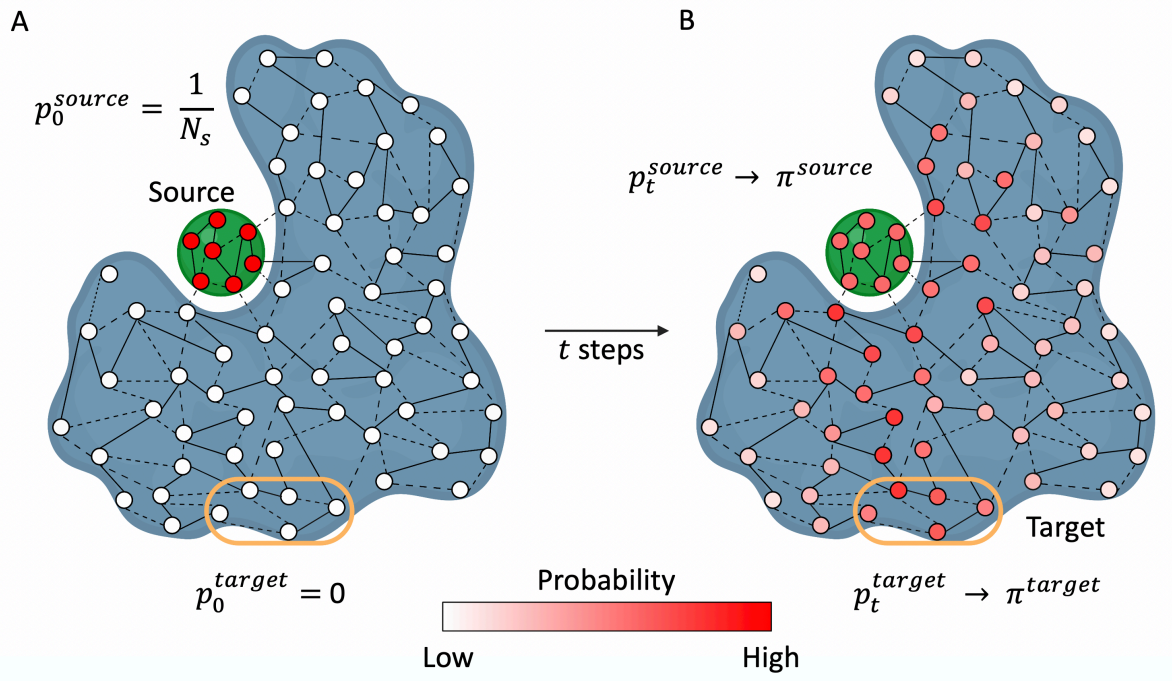


Figure 3.2: Schematic representation of Markov transient analysis. Representation of a protein graph with several nodes connected by different edges. At t_0 **(A)** the probability is equally distributed over the source nodes. The Markovian random process on the graph leads to the time evolution of the probability of each node which eventually reaches a stationary value **(B)**. The $t_{1/2}$ time until the stationary distribution π is reached presents a measure for the connectivity between a source (green) and a target (yellow).*

*Created with biorender.com

When we take into consideration that a dynamical process on the protein graph can be described by a combinatorial Laplacian, we can substitute the Markov transient matrix from 3.2 with $-L$ as such:

$$\dot{p}(t) = -p_t L \tag{3.3}$$

This equation can then be solved as:

$$p(t) = p_0\, exp(-tL) \tag{3.4}$$

The probability distribution $p(t)$ for any given time point t can then be calculated based on the initial probability distribution p_0. As schematically shown in Figure 3.2A, p_0 is defined as a uniform probability distribution over all source nodes:

$$p(0) = \left(0 \;\cdots\; \overbrace{\frac{1}{N_S} \;\cdots\; \frac{1}{N_S}}^{p_0^{\text{source}}} \;\cdots\; \overbrace{0 \;\cdots\; 0}^{p_0^{\text{target}}} \;\cdots\; 0 \right) \tag{3.5}$$

where the number of source nodes is given by N_S. The signal propagation between the source and any given target can be monitored by the change in probability at the target nodes:

$$p(t) = \left(\cdots \; \left[p_t^{\text{source}}\right] \; \cdots \; \left[p_t^{\text{target}}\right] \; \cdots \right) \tag{3.6}$$

When the Markov time t tends to infinity, the probability vector p_t converges to the stationary distribution π of the random walk (Fig. 3.2B). Amor et al.[159] introduced a measure of speed for the random walk started at a source towards a target by considering half the time steps it takes to reach the stationary probability value in any given target node i.

$$t_{1/2}^{(i)} = arg\, \min_t \left[p_t^{(i)} \geq \frac{\pi^{(i)}}{2} \right] \tag{3.7}$$

This characteristic *transient time* $t_{1/2}$ provides a measure for how connected every atom in the

protein is to the source site. The $t_{1/2}$ of a residue is then calculated as the average over all atoms in a residue. MT analysis provides us with a measure for intra-protein communication and highlights atoms and residues that are particularly fast reached by a signal propagation from a site of interest.

3.1.4 Bond-to-bond propensities

Another method is called bond-to-bond propensity (BBP) analysis and is based on the idea of a perturbation at a source edge and how this affects any other edge in a network. First developed for the general application in networks like power grids or traffic flow networks by Schaub et al.[182], Amor et al.[149] extended the concept onto protein graphs. In proteins, bond-to-bond propensities provide a measure for how strongly coupled the source bonds (e.g. within an active site) are to any other bond. Previous studies have leveraged BBP analysis to identify allosteric sites in known allosteric proteins[149] and to investigate allostery in multimeric complexes[162] and protein-protein interactions[163]. Wu et al.[54] recently published a large BBP benchmarking study in two allosteric benchmarking sets (ASBench[124] and CASBench[125]) where we found a combined prediction accuracy of 87 % in 146 proteins. Figure 3.3 provides a schematic visualisation of the BBP concept and we here summarise the details of the methodology as developed in Schaub et al.[182] and Amor et al.[149] and further described by B. Amor[168].

The approach is defined on the edge space of a graph where the *transfer matrix* $M_{m\times m}$ describes a discrete Green's function that quantifies how a perturbation at edge i would instantaneously affect any other edge j[182]. Amor et al.[149] then established that in proteins the transfer matrix M can be defined by

$$M = \frac{1}{2} W B^T L^{\dagger} B \tag{3.8}$$

with W, B^T, L^* and B being the graph matrices defined in Section 3.1.2, which describe

*Importantly, the full pseudo-inverse of the Laplacian $L^{\dagger}$ does not need to be solved in Equation 3.8. Instead, a sparse linear system containing the combinatorial graph Laplacian can be solved which allows the approach to run in almost linear time. With a running time of $O(E\ log^2\ (N))$ dependent on the number of edges E and

relevant protein properties.

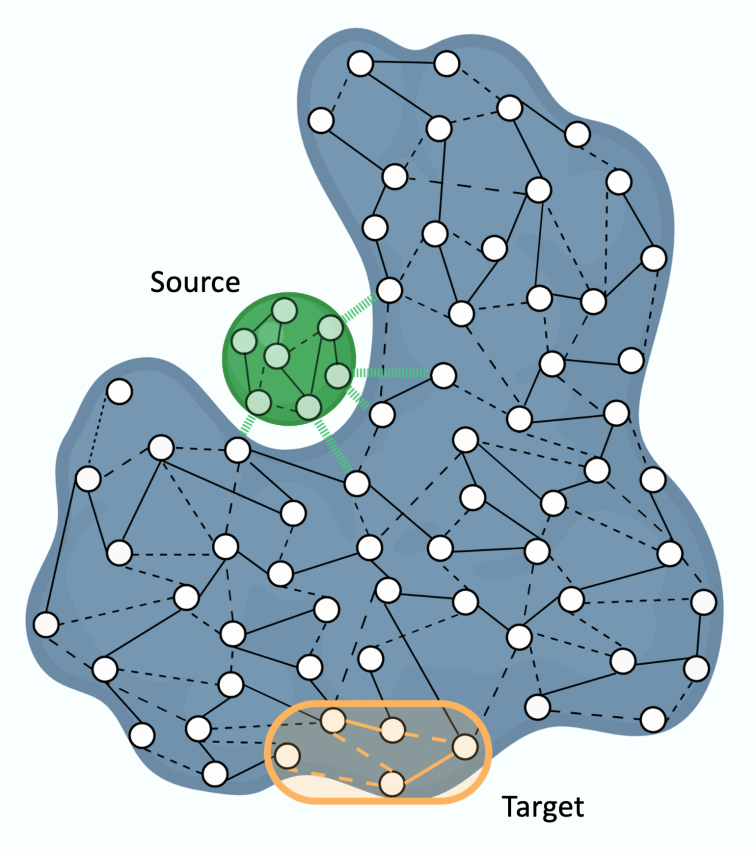


Figure 3.3: Schematic representation of BBP model. The protein (blue) is described as a graph where atoms are nodes connected by edges that represent bonds or interactions. The instantaneous effect of bond fluctuations at a source (green) on a target (yellow) is calculated on the edge space.*

For the purpose of representing protein graphs, we can now deduce that the off-diagonal entries $M_{b_1 b_2}$ reflect how a perturbation at bond b_2 is affecting a bond b_1. Given the incorporation of the weight matrix W, the perturbation effect is weighted by the strength of bond b_1 meaning perturbing a stronger bond is considered to be more important.

To answer biologically relevant questions, we are only interested in slices of M which represent the impact of meaningful source bonds on the rest of the protein. Hence, we obtain for the propensity of any bond b the combined effect of all source bonds:

$$\Pi_b = \sum_{b' \in \text{source}} |M_{bb'}| \tag{3.9}$$

nodes N, BBP analysis is applicable to large protein structures and complexes[149].

*Created with biorender.com

Importantly, this sum only considers weak bonds in the source and the protein object as these weak bonding patterns are known to drive intra-protein communication. A. Delmotte[167] and B. Amor[168] extensively studied the impact of covalent and non-covalent bonds by randomising edges in the graph, the results of which supported the focus on non-covalent bonds.

Throughout this work, our results are based on the detailed propensity calculations at the bond level. We further introduce a measure of connectivity to the source for every residue in the protein by calculating the *residue propensity* as the sum over the bond propensities of all bonds in a residue R:

$$\Pi_R = \sum_{b \in R} \Pi_b \tag{3.10}$$

BBP analysis allows us to the investigate the instantaneous effect of a perturbation at a given source and highlights bonds and residues that are particularly strongly connected to a site of interest.

3.1.5 Quantile scoring and site scores

Due to the nature of interactions within proteins, which is largely defined by local chemical bonds and interactions, we obtain sparse protein graphs. This means that we observe a data pattern that is dependent on the distance from the source in our analysis. For Markov Transients, we observe a general increase of $t_{1/2}$ values with distance from the source, while bond-to-bond propensities decline the further away an edge is from the source. To account for this distance bias, we use a technique called *quantile regression (QR)*[189].

Quantile regression

Whereas standard least squares regression estimates a model for the mean of samples, QR is based on estimating models for conditional quantile functions. This allows highlighting atoms and bonds that are in the tails of the distribution rather than in the mean and thus are of

more interest in our analysis[168]. Further, the quantile functions can reflect non-normal data distributions like the ones we see for Markov transient (MT) and BBP values. Given the exponential decay of BBP values over distance (Fig. 3.4B), we use a linear function of the logarithm when fitting the model. For Markov Transients, we use cubic splines as they allow more flexibility to fit the model to the distribution of $t_{1/2}$ values (Fig. 3.4A). Further details on the QR process are described by B. Amor[168] and in Amor et al.[149] for MT and BBP analyses, respectively.

Figure 3.4 provides example distributions of MT and BBP values across a given protein graph. As can be seen in the figure, atoms X and Y share the same $t_{1/2}$ value, and bonds E and F share the same Π_b value. Thus, if distance was not taken into account, they would rank equally in terms of connectivity to the source site. However, atom Y and bond F are further away from the source in comparison to atom X and bond E. Considering their similar values while at a greater distance, we can deduce atom Y and bond F must be more impacted by the source site signal. Hence, QR is applied to rank each atom or bond in relation to all other bonds at a similar distance. The resulting *quantile score (QS)* of each atom or bond provides us with a quantitative measure of their signalling significance from 0 to 1. We further extend the QR workflow onto the residue level in both MT and BBP analyses.

By incorporating this ranking step which accounts for the distance bias in the data, we can identify areas and residues that are significantly correlated with a source site. These high scoring residues can be understood as biologically meaningful in the context of proteins as they represent distant sites that have the potential to exhibit an impact on the source.

Site scoring and structural bootstrap

To further characterise areas of high connectivity we score them by calculating the average residue QS across multiple residues as:

$$\overline{p_{R,\text{site}}} = \frac{1}{N_{R,\text{site}}} \sum_{R \in \text{site}} p_R \tag{3.11}$$

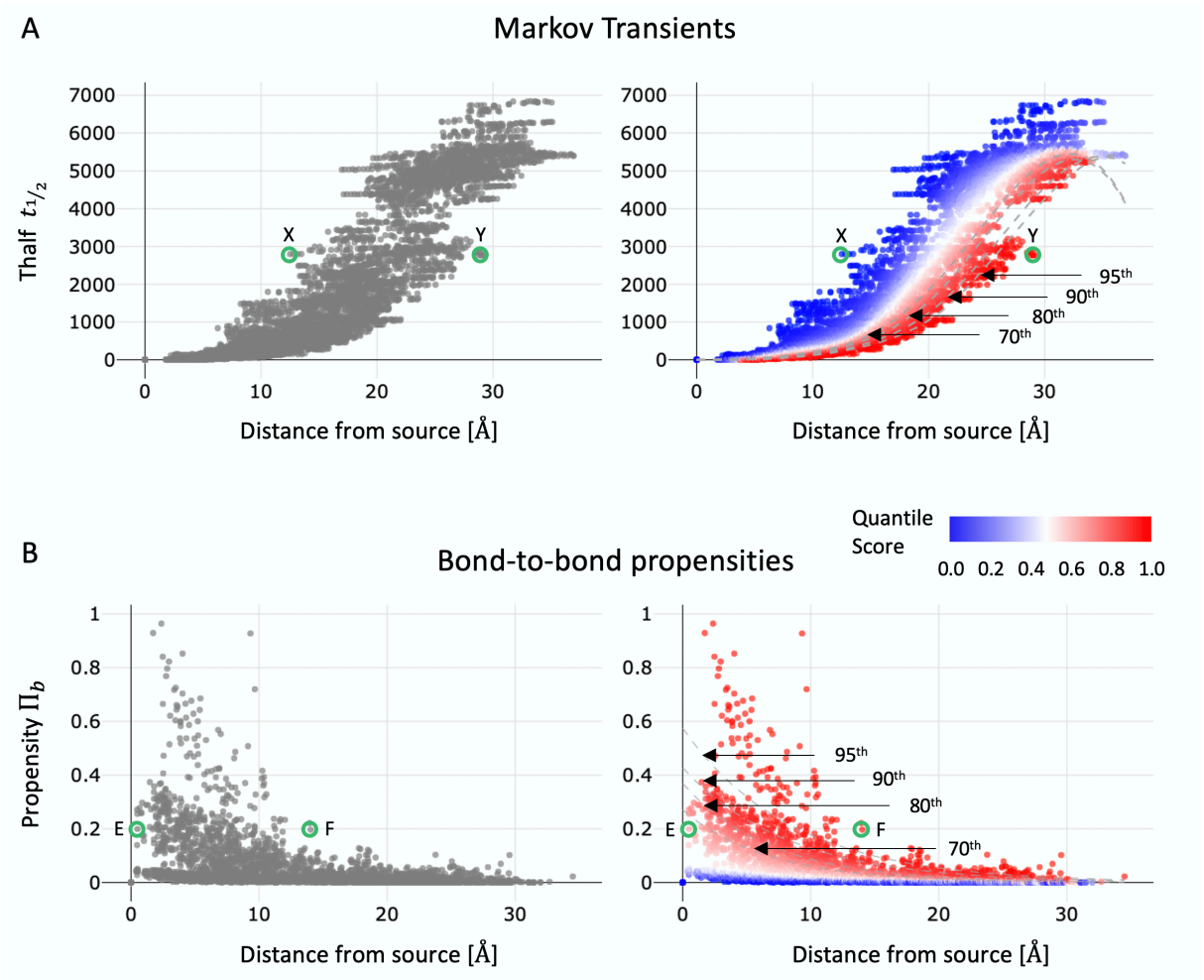


Figure 3.4: The effect of quantile regression for MT and BBP analyses. A) left: Shown are $t_{1/2}$ values over distance from the source for each atom (in grey). Highlighted in green are two atoms X and Y with similar $t_{1/2}$ values but different distances from the source. **Right:** Quantile regression assigns a QS between zero (blue) to 1 (red) to each atom to assess which ones are significantly more connected to the active site. **B)** A similar approach is performed on the propensity values of each bond/interaction in the protein. Two bonds E and F are highlighted that have a similar propensity values Π_b but different distances from the source. The 95th, 90th, 80th and 70th quantiles are indicated with dotted lines in both panels.

This can be applied to score previously known areas of interest, like allosteric sites or protein-protein interaction sites, or other relevant structural features.

By applying a *structural bootstrap* we can assess the significance of a scored site against the rest of the protein as described in detail in Amor et al.[149]. We sample 1000 random surrogate sites of the same residue number and size as the site of interest. For each surrogate site we calculate the average QS $\overline{p_{R,\text{surr}}}$ and then average across the ensemble E of 1000 surrogate sites

to obtain:

$$\langle \overline{p_{R,\mathrm{surr}}} \rangle_E = \frac{1}{1000} \sum_{\mathrm{surr} \in E} \overline{p_{R,\mathrm{surr}}} \tag{3.12}$$

This average of averages can be compared to the average QS of a site of interest. To provide statistical significance, we apply a bootstrap with 10000 resamples with replacement[190] to the ensemble values to calculate a 95 % confidence interval (CI).

3.1.6 Conclusion

Atomistic graph analysis serves as a workflow to analyse biomolecular structures with the aim to uncover allosteric effects and intra-molecular communication. The workflow was developed and refined over the last decade[149,158,159,167,168] and has since been proven an effective tool in protein systems[161–163]. The two methodologies used here, MT and BBP analyses, are of complementary nature and highlight different aspects of communication within proteins by providing measures for fast and strong connectivity, respectively. This Thesis sees the application of atomistic graph analysis to three protein dimers in the context of disease in Chapters 4, 5 and 6.

The following Sections describe methodological additions to the workflow that were used in this Thesis. We further introduce the web server ProteinLens which was recently deployed and published[51] to provide the public with a tool to study protein allostery with atomistic graph analysis.

3.2 Development of additional tools for atomistic graph analysis

3.2.1 Structural features and visualisations

PyMol[191] was used throughout this work to investigate protein structures in 3D and visualise results. This graphic interface implements a multitude of visualisation techniques and allows for custom scripts. These custom scripts were used to map QS results onto protein structures and highlight high scoring residues.

Definition of protein-protein interfaces

One of the aims of this work is to explore the interface that is formed between monomeric protein chains in a dimeric assembly. To investigate the connectivity within and towards these dimer interfaces we need to have a definition of their location and which residues are involved in the formation. For this purpose, we used *PDBePisa*[*], an online tool that can be used to explore macromolecular interfaces[78]. We provided PDBePisa with the pre-processed protein structures in `.pdb` format. From these crystalline states the tool infers a list of residues that are at the interface by considering the area that becomes inaccessible if two protein chains are brought into contact[192]. PDBePisa further detects whether interface residues form bonds, in which case it distinguishes between hydrogen bonds, salt bridges, disulphide bonds and covalent links.

Solvent accessible surface area

Our methodologies enable us to detect residues and hotspots in a protein that show allosteric potential. For the definition of allosteric hotspots, we provide an indication of how buried or accessible the residues in the hotspot are. We find the residue wise solvent-accessible surface

[*]Accessible at ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver

area (SASA) using the `get_area` function in PyMol[191] with default settings: a dot density of 2 and a solvent radius of 1.4.

3.2.2 ProteinLens - a user-friendly interactive webserver

> The implementation of the web server in Django was coded in large parts by S. Mersmann, the concept, general structure and website texting were led by the author of this Thesis. ProteinLens can be accessed at proteinlens.io.

The methodologies in previous Sections have been shown to effectively predict allosteric sites and signalling pathways. Further evidence of their predictive power has been provided in small[149] and in large datasets[54]. Additionally, a wide portfolio of biological systems has been described through atomistic graph analysis: from small proteins[149,159] to large complexes[162], in the setting of allergenicity[186] to drug repurposing in cancer[161]. Hence, we consider it our responsibility to provide easy access to these versatile methodologies to allow community usage for further scientific explorations of allosteric signalling and site detection. We implemented our atomistic graph analysis pipeline in an interactive web server, incorporating the two methods Markov Transients and bond-to-bond propensities. The only requirement for the user is to provide source residues and a PDB id of a structure of interest. The web server then provides complementary insights into the speed and strength of communication within the structure which allows to study the allosteric effect in the system. Another major advantage of our methodology is the computational efficiency as described in more detail above. This means the user can obtain the results for their structure within minutes*. We put a particular focus on user-friendliness and an intuitive presentation of the results. ProteinLens provides fully interactive 3D visualisations which can be screenshotted, and all data can be downloaded. We further allow the user to score sites of interest to assess the significance of a found hotspot.

*An overview of running times for proteins of different sizes can be found in the frequently asked questions (FAQ) page of ProteinLens: proteinlens.io/webserver/faq

Web applications for the prediction of allosteric sites and pathways

Allostery is a ubiquitous concept in protein regulation[35] and using allosteric principles for targeting proteins is a fruitful approach. Allosteric sites allow more selective and robust targeting of disease-causing proteins especially for members of large protein families[17] as described in Section 1.3.1. However, the experimental discovery of allosterically regulated proteins is often a product of chance or requires high-throughput screenings[165]. Computational approaches allow for a faster exploration and prediction of protein allostery, ushering in a new era for allosteric discovery. To benefit the wider community, some of these approaches have been released in the form of publicly accessible* web servers as listed in Table 3.1. Mostly, these web servers either focus on allosteric site, pathway, or functional residue prediction.

Table 3.1: Web servers to predict allosteric sites and signalling paths.

Type	Name	Link	Methodology
Allosteric sites	AllosMod[116]	modbase.compbio.ucsf.edu/allosmod	MD
	AlloSite[106]/ AlloSitePro[115]	mdl.shsmu.edu.cn/AST/Allosite	NMA & ML
	CorrSite 2.0[119]	within CavityPlus[197]: pkumdl.cn:8000/cavityplus	GNM
	PASSer[198]	passer.smu.edu	ML
Allosteric signalling	MCPath[145]	safir.prc.boun.edu.tr/clbet_server	Monte Carlo path simulations
	Dynomics[146]	dyn.life.nthu.edu.tw/oENM	ENM
	AlloSigMA 2[199]	allosigma.bii.a-star.edu.sg/home	SBSMMA

Workflow of ProteinLens

ProteinLens presents an all-encompassing tool to study atomistic communication pathways and connectivity within single protein chains as well as at the scale of protein multimers. The underlying theoretical methods are computationally inexpensive because they rely on sparse matrices as described in Sections 3.1.2 and 3.1.4. This makes them well suited to be deployed within an interactive web application. Figure 3.5 provides a summary of the whole workflow on the website which the user can follow. At each step the user can provide input in an easily

*Further web servers in the field have been described but are currently partly or fully inaccessible: PARS[193], SPACER[194], STRESS[195], OHM[196].

accessible and interactive way. The following Sections provide a short overview of each web server step and describe the insights that can be gained.

Input

The input to ProteinLens is structural files of biomolecules which can either be sourced directly from the PDB with their respective identifier or uploaded by the user. BagPype was built into ProteinLens for constructing atomistic graphs[160]. As discussed in Section 3.1.2, BagPype can process DNA molecules in PDB structures and further provides a range of useful options. These options provide flexibility to the user in terms of stripping certain entries from the PDB file, choosing how to handle NMR models and processing multimeric proteins. The settings provided by the user are taken into consideration when converting the input structural data into atomistic graph representations[160]. After successful processing of the structure, the user is automatically forwarded to the next page which presents a quick summary of the constructed graph. The page summarises the main features of the graph, like the number of nodes (atoms) and edges (bond/interactions). At this stage, the user also receives feedback on the graph's connectivity as a connected graph* is required for our methodologies. If the graph is disconnected, the user can choose which subset of the graph to use for the rest of the analysis.

Computational settings

The next step in the ProteinLens workflow is the calculation of bond-to-bond propensities[149] and Markov transients[159] on the biomolecular graph. To do so the user must provide the source residues, either as a list of residue numbers or by choosing a ligand that was identified in a protein chain (Fig. 3.5B). In the most common use case for ProteinLens, a source could be chosen to be the active site ligand or binding site residues. The web server's results would then indicate which parts of the protein exhibit a functional connectivity to the active site. Finally, the user can choose which methodology to run. By default, both BBP and MT analyses will be performed.

*In a connected graph, every node is in contact with at least one other node over at least one edge.

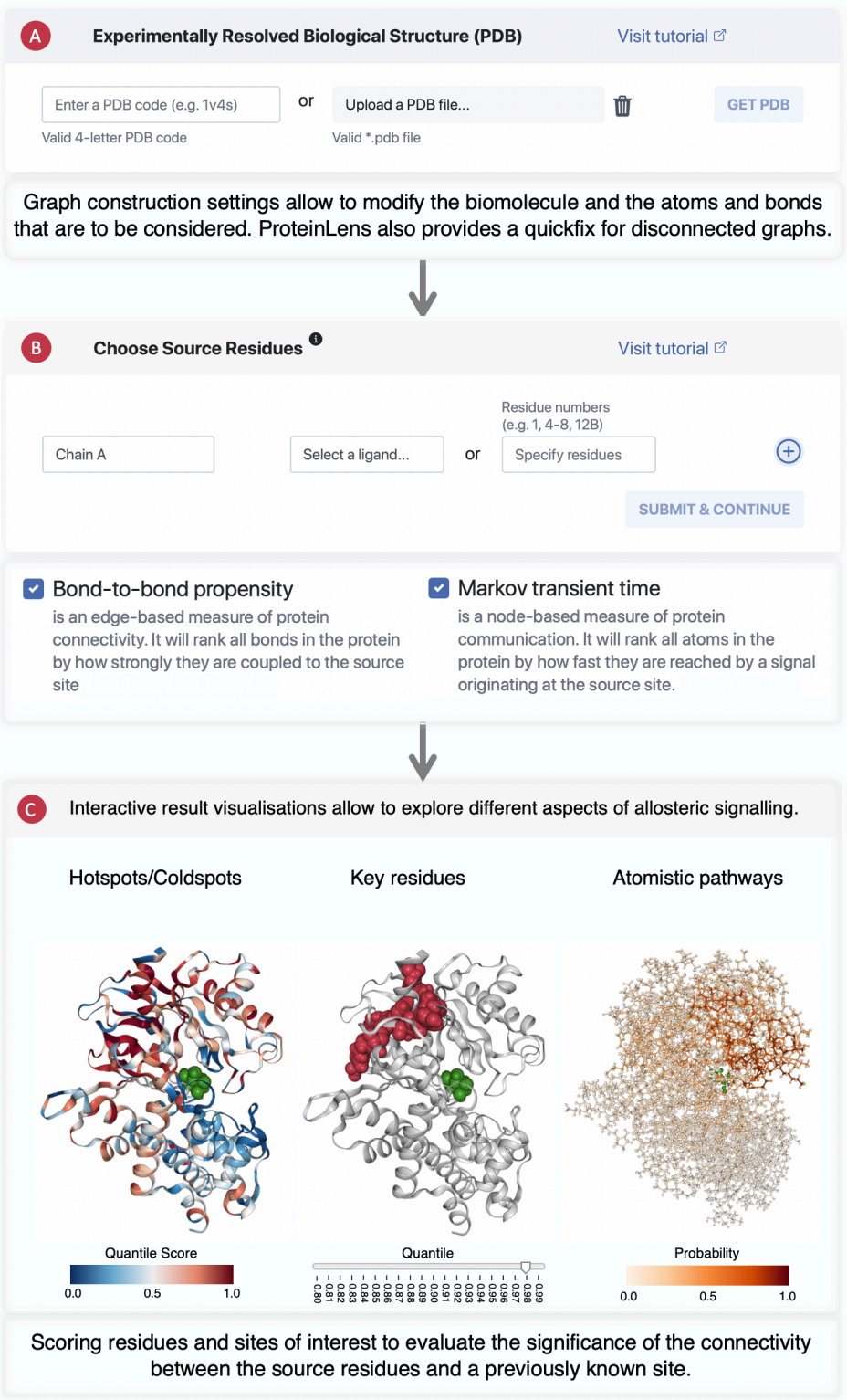


Figure 3.5: Workflow of ProteinLens. **A)** The user can choose to retrieve a PDB entry or upload a `.pdb` file. The user can then choose which biomolecule to consider for graph construction and advanced options allow to strip certain atoms or ligands. **B)** After the graph is constructed, the user has to specify which residues to choose as a source. ProteinLens features a dropdown menu for all ligands that were detected in each chain or the user can specify residue numbers. The user can then choose which methodology to run. **C)** The results are presented in interactive, complementary visualisations that highlight different elements of allosteric communication. ProteinLens also provides a scoring panel where the significance of residues or sites of interest can be assessed. Adapted from Mersmann et al.[51].

Results

The final page to which the user is redirected presents the results of the analyses. The user is provided with a summary of the structure and the settings they chose and can then toggle out a variety of different panes. ProteinLens provides complementary result visualisations that highlight different aspects of the results, allowing to detect different allosteric features.

The hotspots view can be accessed for both BBP and MT results. In an interactive 3D-viewer, the chosen structure is coloured according to quantile score results (Fig. 3.5C left). For this panel, ProteinLens also provides an interactive plot of the data distribution which is fully linked to the structure viewer. This visualisation allows identifying areas of the protein that are hotspots in our analysis and hence hold potential for allosteric regulation.

The relevant residues view is also available for results from both methodologies. This visualisation highlights residues that are above a certain QS threshold which can be interactively set by the user. These high scoring residues are highlighted on the 3D structure (Fig. 3.5C middle) and on the interactive data distribution plot that is provided for this panel. This visualisation allows the user to investigate relevant residues that contribute to the allosteric behaviour of the protein.

The scoring panel provides the option to score a site of interest in the context of bond-to-bond propensities. The user can provide residues that belong to a functionally relevant site like a known allosteric site. ProteinLens will then score this site and randomly sample surrogate sites to allow a comparison to the score of a random site. Details on this structural bootstrap approach are provided in Section 3.1.5.

The random walker visualises the MT time steps that underly the $t_{1/2}$ calculations. Each atom in the interactive structure is coloured by the probability of the random walker being at this node at a time step t (Fig. 3.5C right). With a user-controlled slider, the probability propagation at different time steps can be shown, allowing the user to investigate signalling paths within the structure.

All panels provide the option to take screenshots of the results and the structures in a chosen

orientation. Further, the user is provided with the option to download all raw data and results with one click.

Implementation

ProteinLens is coded in Python[172] and incorporates the methodologies of our group as described above. The web server was built using Django (v.3.1)[200] with an SQLite database in the back. The front-facing design is coded with Bootstrap (v.4.3.1)* and visualisations are based on the D3.js library† and the NGL viewer[201]. The initial structure and scaffold of the backend were set up by the author of this work. The bulk of the coding and the integration of the visualisation tools was done by S. Mersmann.

Documentation

ProteinLens features a detailed tutorial‡ of the whole analysis workflow. Furthermore, each panel on the website is linked out to the relevant section in the tutorial for easy access to further information. Many concepts are also explained by providing information boxes which can be accessed by hovering over particular terms. If the user wishes to learn more about the underlying methodology, they can do so by accessing the extensive background§ page. We also provide a page with FAQ¶ for the purpose of troubleshooting.

Conclusion

Taken together, the features of ProteinLens allow the community to explore allosteric signalling in their own case studies within minutes and in an intuitive manner. The underlying methodologies have been benchmarked and used across a variety of study systems. We are confident that this will be a valuable contribution to the field of protein allostery.

*Accessible at: getbootstrap.com/docs/4.3/getting-started/introduction/
†Accessible at: d3js.org
‡Accessible at: proteinlens.io/webserver/tutorial
§Accessible at: proteinlens.io/webserver/background
¶Accessible at: proteinlens.io/webserver/faq

The publication in *Nucleic Acids Research* also contained a small case study on human glucokinase where we demonstrate how the web server can be used to analyse a protein with a known allosteric site[51]. We recovered the allosteric site with bond-to-bond propensities and exemplified how the scoring functionality in ProteinLens can be used. In the following Chapters we analyse more complex and less studied systems and demonstrate how atomistic graph analysis can provide valuable insights and contribute to our understanding of these systems in disease.

Chapter 4

Estrogen receptor alpha

This Chapter builds on work done by L. Strömich[202], and some of these results are shortly summarised here. Where this is the case, we clearly indicate how this current work is an extension and goes beyond what was previously found. Parts of this Chapter were motivated by experimental results obtained by our collaborators Fui Lai and Simak Ali, and this is clearly indicated throughout.

This Chapter presents our atomistic graph analysis of estrogen receptor α (ERα). This first study system demonstrated the validity of our approaches for the study of molecular mechanisms in dimeric proteins. We discuss the results of bond-to-bond propensity (BBP) analysis in the system from different biologically relevant source sites and how this methodology can aid in determining specific residues that are involved in dimer interface signalling.

4.1 A nuclear hormone receptor regulating gene expression

One of the major roles of proteins in our cells is the regulation of gene expression which ranges from DNA remodelling to transcription initiation[203]. Proteins that induce transcription are

often known as transcription factors (TFs), and they are unified by their ability to bind to DNA and initiate gene expression[204]. One of the largest TF families are nuclear receptors (NRs) which are involved in most biological processes and regulate gene expression in all tissues in our body[205]. The NR mode of action roughly follows this pattern: binding of a ligand molecule promotes localisation into the nucleus where the receptor binds to DNA. NRs are active as monomers, homodimers and heterodimers and they often recruit co-activators to regulate gene expression. Natural ligands of NRs are small lipophilic molecules that range from thyroid/steroid/retinoid hormones* to molecules that are involved in lipid metabolism[18].

Two members of the classic hormone NRs are estrogen receptor α (ERα) and estrogen receptor β (ERβ), which bind to estrogen steroid hormones[206]. ERα (also known as ER1 or Esr1) was identified as receptor for the modulation of estrogen signalling in the 1960s and was one of the first known ligand-activated TFs[207,208]. The second member of nuclear estrogen receptors was identified three decades later in 1996 and termed ERβ or ER2/Esr2[209]. The two ERs bind estrogens and anti-estrogens but diverge in their binding modes, physiological effects and tissue specificity (reviewed in Jia et al.[210]). The focus of this work is on ERα, as it is widely described and studied as the primary driver of breast cancer (BC) development and progression, as discussed below.

4.1.1 Molecular mode of action of ERα

ERα has a variety of functions in estrogen-targeted tissues, and its three main roles are categorised by the cell compartments that ERα is active in (reviewed in Yasar et al.[3]). At the plasma membrane, ERα interacts with G proteins to trigger kinase signalling cascades that mediate a rapid response to estradiol signals[211]. Secondly, ERα is found in mitochondria where it regulates gene expression of mitochondrial TFs[212]. Lastly, ERα is found in the nucleus, where it fulfils its most important role: initiating gene transcription[3]. This transcription initiation can happen in a direct manner by binding to an estrogen response element (ERE) on DNA[213] or indirectly by recruiting and assembling other TFs[3]. Once ERα binds directly to

*NRs binding to hormones are called nuclear hormone receptor (NHR).

an ERE, further co-factors like p160 regulators and histone modulating proteins are recruited and allow the transcription activation of an extensive array of human genes[214,215]. For the indirect genomic mechanism, also known as ERE-independent genomic mechanism, ERα is responsible for activating TFs like activator protein 1 (AP-1)[216] and specificity protein 1 (Sp-1)[217]. These TFs bind to their own response elements, further increasing the number of genes that are under regulation of estrogen signalling through ERα. Crucial for either transcription activation mechanism is the formation of ERα homodimers[218–220], as mutations that impact dimerisation render the protein defective[221]. This essential dimerisation is also observed for the ERα signalling pathways at the cell membrane and within mitochondria[3].

Structural features of ERα

From a structural perspective, ERα follows the general pattern seen for NHRs, as shown in Figure 4.1. The receptors are modular proteins where different domains are fulfilling distinct functions. The N-terminal region is intrinsically disordered but important in conferring transcription activation function 1 (AF-1)[222]. AF-1 of ERα is dependent on interactions with the protein C-terminus and involved in binding co-activator proteins[223,224]. The DNA-binding domain (DBD) (highlighted in red in Fig. 4.1A) is the part of the protein that interacts directly with DNA by binding to the ERE[213]. Connected to the DBD over a hinge region is the hormone or ligand-binding domain (LBD) (shown in orange in Fig. 4.1A), which mainly consists of 12 α helices that are oriented in a three-layered antiparallel fold. The hydrophobic ligand-binding cavity that forms within this fold can be occupied by small lipophilic molecules[225], such as estradiol, which is shown in Figure 4.1B and C. Most structures available for ERα are of the LBD only and show the protein in the unbound apo state or bound to estrogens as well as anti-estrogens. All of these structures are deposited as dimers, as the dimerisation is essential for function and is regulated by ligand binding events[226].

Unfortunately, to this date there is no full-length structure of estrogen receptors available which can be explained by the assumption that the DBD and LBD are connected by disordered regions, making structure determination a challenging task*. However, there are some

*For an AlphaFold prediction[49] that confirms large disordered loops in the full-length monomeric ERα,

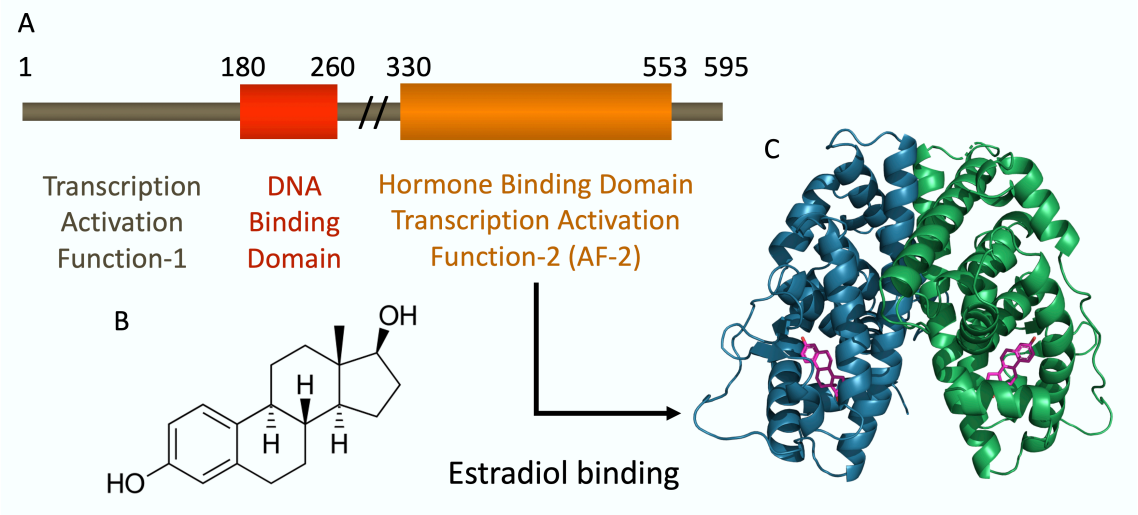


Figure 4.1: Human estrogen receptor alpha and its functional domains. A) A schematic overview of the full-length ERα. The functional domains are indicated with the DNA binding domain in red and the hormone or ligand-binding domain in orange. **C)** The structure of the LBD bound to the natural ligand estradiol (**B**, pink) with the two monomer halves shown in blue and green (PDB id: 1G50[227]).

structures available of full-length NHRs that provide insights into the allosteric couplings of the distinct domains[228]. The structures of the heterodimeric PPAR-RXR complex (PDB id: 3E00/3DZU/3DZY[229]), the homodimeric HNF-4α (PDB id: 4IQR[230]), the heterodimeric RXR-LXR complex (PDB id: 4NQA[231]) and the heterodimeric RARβ-RXRα complex (PBD id: 5UAN[232]) are all solved bound to DNA and reveal the quaternary structures of nuclear receptors. Although the structures have large gaps in the inter-domain loop regions, they allow deducing how the domains are oriented towards each other and suggest allosteric signalling through domain interactions (reviewed in Rastinejad et al.[228]).

Ligand binding triggers activation function-2

The natural ligands of ERα are estrogens, the most potent one of which is 17β-estradiol (EST)[233]. Upon binding of estrogen hormones like EST to the LBD, the highly dynamic helix 12 (H12) moves over the ligand-binding site and forms a co-activator binding groove together with parts of helices three, four and five[219]. This so-called agonist-bound conformation of the LBD is essential for co-activator assembly, also known as transcription activation function 2 (AF-2) of ERα[225]. For example, co-activators of the p160 steroid receptor co-activator

see Figure B.1.

(SRC) family present an LXXLL* motif that allows them to occupy the binding groove in the agonist-bound structure of ERα[234]. The structural rearrangement of H12 that underlies the activity of the agonist-bound conformation, also plays a major role in the inhibition of ERα when bound to anti-estrogens. When anti-estrogens like tamoxifen or raloxifen bind to the LBD, H12 itself localises into the above-mentioned groove with its L^{540}-L^{541}-E^{542}-M^{543}-L^{544} motif[219,235]. In this antagonist-bound conformation, no co-factors can assemble, and AF-2 of ERα is inhibited. Figure 4.2 provides a visual comparison of the agonist (to EST) and antagonist-bound (to 4-hydroxytamoxifen (OHT)) conformations of the LBD of ERα. The most striking difference is the localisation of H12, as highlighted in orange.

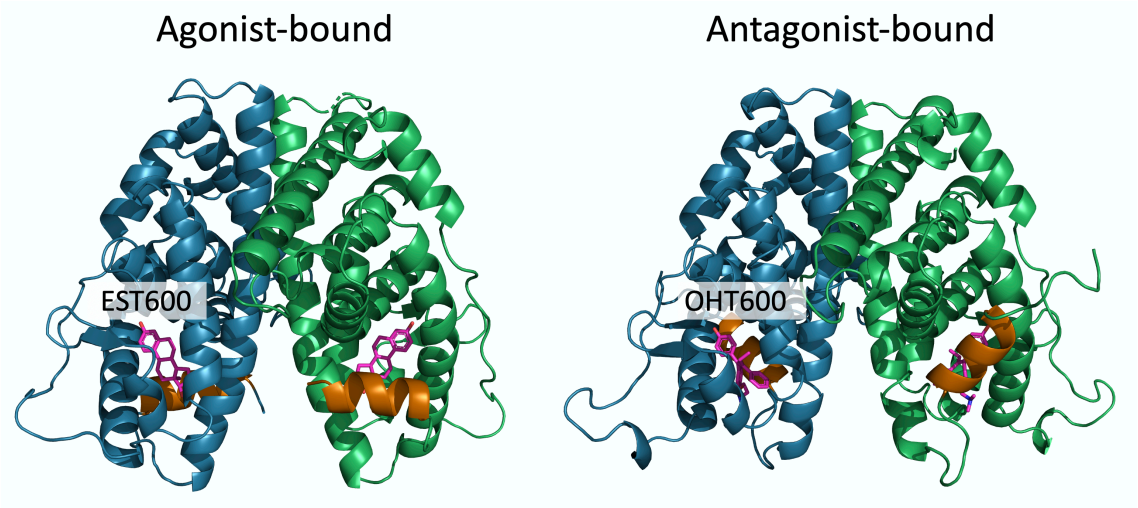


Figure 4.2: Agonist and antagonist-bound conformations of the ERα ligand-binding domain. Shown are the agonist (PDB id: 1G50[227]) and antagonist-bound (PDB id: 3ERT[235]) conformations of the ERα LBD. The two dimer halves are shown in blue and green. The bound ligands in pink are 17β-estradiol (EST) and 4-hydroxytamoxifen (OHT). H12 is highlighted in orange in the two different conformations.

4.1.2 ERα in breast cancer

Apart from having a wide range of essential physiological functions[233], estrogen signalling also plays a vital role in the development of malignancies in the respective target tissues. One tissue under control by estrogen receptors is the human mammary gland, where a misbalance in exposure to ER mediated signalling can lead to the formation of BC tumours[236]. Although BC

*L - leucine, X - any amino acid

is a highly diverse disease in phenotype[237] and genotype[238], most tumours (70 - 80 %) express ERα. The roles of ERα in BC development and progression are complex and have been reviewed extensively[239,240]. Figure 4.3A provides a simplified summary of the role of ERα in BC, focusing on the structural mechanism. The binding of natural ligands like estradiol stimulates AF-2 by complexing H12 into an agonist-bound conformation that allows co-activator assembly and promotes transcription initiation. If there is too much exposure to this agonistic signal, gene expression is overstimulated, leading to enhanced cell growth and tumour formation.

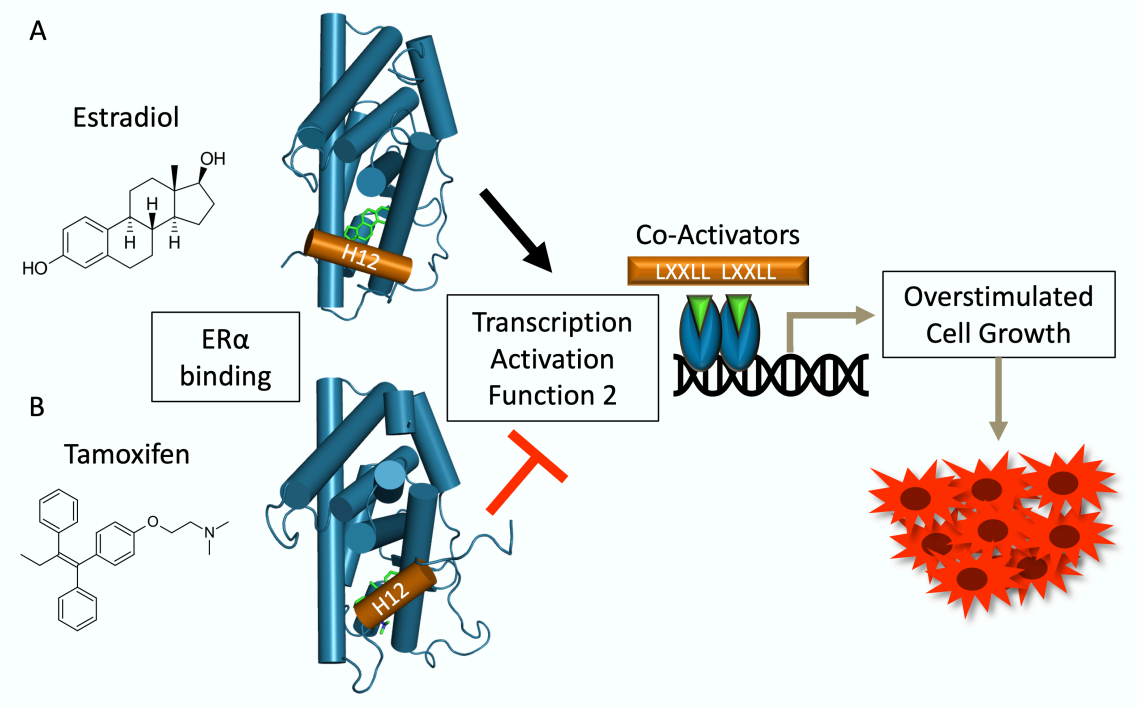


Figure 4.3: Schematic representation of the structural features of ERα LBD and molecular mechanism in cancer. The LBD of ERα is shown in agonist and antagonist-bound conformations with helices shown as barrels. **A)** Upon agonist (estradiol) binding H12 (shown in orange) forms a co-activator binding groove. This so-called transcription activation function 2 means co-activators with an LXXLL binding motif can assemble, and gene expression is activated. If this input signal is out of control, overstimulated cell growth occurs, and tumours can form. **B)** When an anti-estrogen like tamoxifen binds, H12 localises into the co-activator binding groove. AF-2 is inhibited, and none of the downstream effects occurs. Adapted from L. Strömich[202].

Inhibiting ERα thus becomes a treatment strategy against BC tumours, which makes the receptor an important target for drug development[241], leading to ERα-focused therapeutic strategies for BC being developed as early as the 1960s[239]. The concept of anti-estrogen chemotherapeutics like tamoxifen is based on the antagonist-bound conformation, which inhibits AF-2 and all

downstream effects (Fig. 4.3B).

Different classes of chemotherapeutics

Tamoxifen was developed in 1966 and is the earliest example of an anti-estrogenic BC therapy[242]. However, much research has been done since, and there are now several classes of chemotherapeutics active against ERα. The selective estrogen receptor covalent antagonists (SERCAs) bind covalently to the binding site in the LBD and show potency against cancer mutants[243]. Another class comprises proteolysis-targeting chimeras (PROTACs), which tag ERα for degradation through ubiquitination[244]. The two most common classes by far, which have undergone several generations of inhibitor optimisation are listed below:

Selective estrogen receptor modulators (SERMs) are one commonly used class against ERα[245]. They are called modulators because they can have agonistic as well as antagonistic function, dependent on the tissue they work in[246]. Examples of this class are tamoxifen, raloxifene, bazedoxifene and lasofoxifene. The mode of action relies on forcing the localisation of H12 into the co-factor binding groove, yielding the antagonist-bound conformation and the consequent inhibition of ERα function. This conformation of H12 has been structurally confirmed for SERMs like raloxifene[219], 4-hydroxytamoxifen[235], lasofoxifene[247] and bazedoxifene[248].

Selective estrogen receptor degrader/downregulators (SERDs) are drugs that do not display any agonist activity and lead to the degradation of ERα. The earliest member of that class is fulvestrant (Faslodex), developed as second-line* endocrine therapy approved for usage in BC[249]. Upon binding of fulvestrant, ERα dimerisation and localisation into the nucleus is impaired, and the protein is in an unstable conformation, leading to its degradation[250]. Other SERDs showing a similar mechanism of action are in various stages of development and approval: AZD9496[251], GDC-0810[252], RAD1901[253] and AZD9833[254]. Structures are solved for AZD9496 (PDB id: 5ACC[255]) and AZD9833 (PDB id: 6ZOR[254]) and show the dimeric ERα LBD with H12 in antagonist-bound conformation.

*Second-line therapies are coming into play once a patient shows recurrent tumour growth that is resistant against first-line BC chemotherapeutics.

Cancer mutations and overcoming resistance

Every new class and generation of ERα inhibitors aims to overcome the disadvantages of previous anti-estrogens. This search for new therapeutic agents is often fuelled by the acquired resistance against prior hormonal therapies[256], and a range of somatic mutations implicated in endocrine resistance has been described (reviewed in Dustin et al.[257]). Most of them are located in the LBD, and some of the most prevalent ones lead to estrogen hypersensitivity or estrogen-independent activity. Two well-explored and structurally solved mutations are Y537S and D358G which are known to stimulate AF-2[258]. These mutations are proposed to stabilise H12 in an agonist-bound conformation without requiring a ligand-binding event. LEU^{536} is a nearby, much less studied position whose mutations have also been found in many cancer genomes[257].

To continue BC treatment after the emergence of mutation-driven resistance against primary therapies, a major focus of ERα research is the search for new inhibitors and novel therapeutic approaches. Computational explorations of the protein aid the continuous efforts to find new target sites and methods, as they are often a fast and inexpensive way of obtaining predictions that allow to guide rational drug design (see Sec. 2.1). Bafna et al.[259] extensively reviewed the scope of computer-aided drug design (CADD) for ERα, highlighting alternative targeting options like the DBD and the C-terminal F-domain. For the LBD, the estrogen binding pocket remains the main focus for development of new therapeutics belonging to the SERM, SERD or covalent inhibitor classes described above. However, some studies have explored the potential of other structurally important sites like the co-activator groove[260]. Another possibility is targeting the LBD dimer interface to disrupt ERα dimerisation which is essential for functionality[226].

A molecular dynamics study by Fratev[261] has shown the interplay between dimerisation and the localisation of H12 into agonist and antagonist-bound conformations. This study further suggests that estrogens and anti-estrogens control H12 conformations over inter-dimer interactions[261]. The work by Chakraborty and colleagues explored the possibilities of targeting ERα by disrupting the LBD dimer interface. Based on *in silico* design, they developed helical peptide

grafts to bind against the dimer interface motifs located in H10, H11 and prevent dimerisation[97,98]. Although targeting the dimer interface with helical peptides has been suggested as early as 2001[262], to the best of our knowledge no further work has been done on this inhibition mechanism.

Objective

ERα has a key role in BC development and progression, and current day therapies focus on the inhibition of its activity. To tackle resistance of recurrent tumours, finding novel drugs and targeting mechanisms against ERα is of high importance. This Chapter examines the molecular mechanism of ERα with our graph analysis on atomistic resolution. In the scope of this Thesis, this study system provides a good starting point as it is well described and allows us to show the validity of our approach in homodimeric proteins. We further describe new strategies to inhibit AF-2 of ERα, including a detailed investigation of the dimerisation process and how it can be disrupted. We also include an analysis of cancer mutations resistant to certain classes of chemotherapeutics. In doing so, we show how bond-to-bond propensities can be applied to understand resistance mechanisms and which insights can be gained to guide the development of future BC therapies.

4.2 Bond-to-bond propensities validate the molecular mechanism of ERα

> The results in this Section have previously been described by L. Strömich[202]. They are shortly summarised here, as they are the basis for further studies that were done in the scope of this Thesis and are presented from Section 4.3 onwards.

The prevalent role of ERα in BC and the continuous search for alternative targeting strategies motivated us to employ our atomistic graph analysis across the protein. Firstly, we aimed to elucidate the connectivity and signalling differences between the agonist and antagonist-bound

conformations of the ligand-binding domain (LBD). The structures were obtained from the PDB and prepared as described in Appendix A.1.1. It is important to note here that the structural file of the agonist-bound confirmation (PDB id: 1G50[227]) contained a fully solved homodimer, while the antagonist-bound confirmation was deposited as one protein chain (PDB id: 3ERT[235]) and the second monomer was modelled based on symmetry information in the `.pdb` file. As a result, we detect minor differences between residues in the two monomers in the agonist-bound conformation and hence, provide all results as average between the two chains. After atomistic graphs were constructed as described in Section 3.1.2, we ran Markov transient (MT) and bond-to-bond propensity (BBP) analyses sourced from the respective bound ligands. We chose to source in a symmetrical fashion from both dimer halves to mirror the binding events that would happen *in vivo**. Details of the used source residues can be found in Appendix A.1.1.

As shown in previous studies[159,161], MT analysis is especially powerful in proteins with catalytic activity. The MT random walker efficiently highlights atoms and residues, which contribute to a fast signal propagation and might be involved in allosteric signalling[159]. The propagation of these random walks can be captured for every Markov time step as shown for selected steps in the agonist-bound conformation of the ERα LBD in Figure 4.4 and Figure B.2 for the antagonist-bound structure. When we sourced the MT analysis from the ligands in the ERα LBD, we did not detect any divergent patterns in the $t_{1/2}$ times. Instead, we can see a signal propagation that diffuses uniformly through the graph without highlighting particular residue areas or paths. As ERα is a receptor protein and hence does not confer enzymatic catalysis, these observations tie in with what we know about the general use case of MT analysis for catalytically active proteins of interest[159,161].

Based on these first MT results, we decided to focus on the study of the ERα LBD with the means of BBP analysis and present these results in the following.

*To the best of our knowledge, the binding of estrogens and anti-estrogens to ERα has always been described to happen simultaneously and there are no experimental or computational studies that describe the effect of one sided binding events.

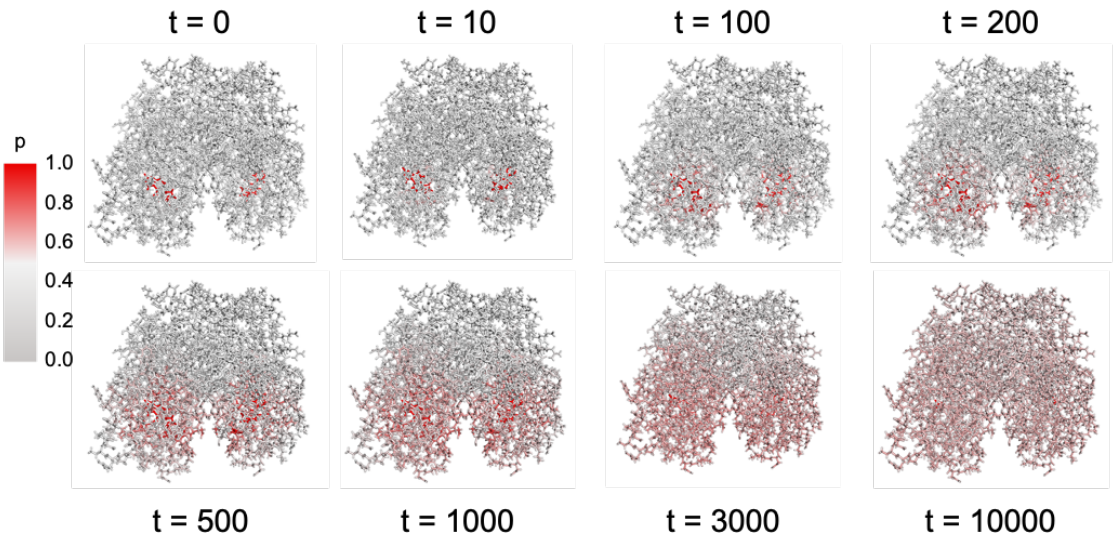


Figure 4.4: MT time steps in the ERα LBD when sourced from EST. The ERα LBD (PDB id: 1G50[227]) is shown in an all-atom stick representation. Atoms are coloured by probability (0 - grey to 1 - red) of the random walker being at this node at a given Markov time step t. Shown are different time steps in the MT analysis, as indicated.

4.2.1 Connectivity towards H12

BBP analysis was sourced from the bound ligands to search for residues that are strongly connected to the molecules that regulate ERα activity. Figure 4.5 provides an overview of the results of this first analysis which was previously presented by L. Strömich[202] and hence is only shortly summarised here. We provide the full data distribution of the residue propensities Π_R over the distance from the source residues for both agonist and antagonist-bound conformations. In the case of the agonist-bound conformation we find high propensity connectivities between the bound EST molecules and the residues of H12 (Fig. 4.5A, C). On a residue level, we find that primarily hydrophobic amino acids in H12 like leucine and methionine score highly in the agonist-bound conformation (Fig. 4.5E).

For the antagonist-bound structure on the other hand, bond-to-bond propensities did not have high values for the connectivity of OHT and H12 (Fig. 4.5B, D). On a single residue level, all residues that are part of H12 have a low quantile score with the highest one of 0.54 for LEU^{544} (Fig. 4.5F) . We need to keep in mind here that the rearrangement of H12 between the two conformations also leads to a shift in residues that form H12 from 539-547 in the agonist-

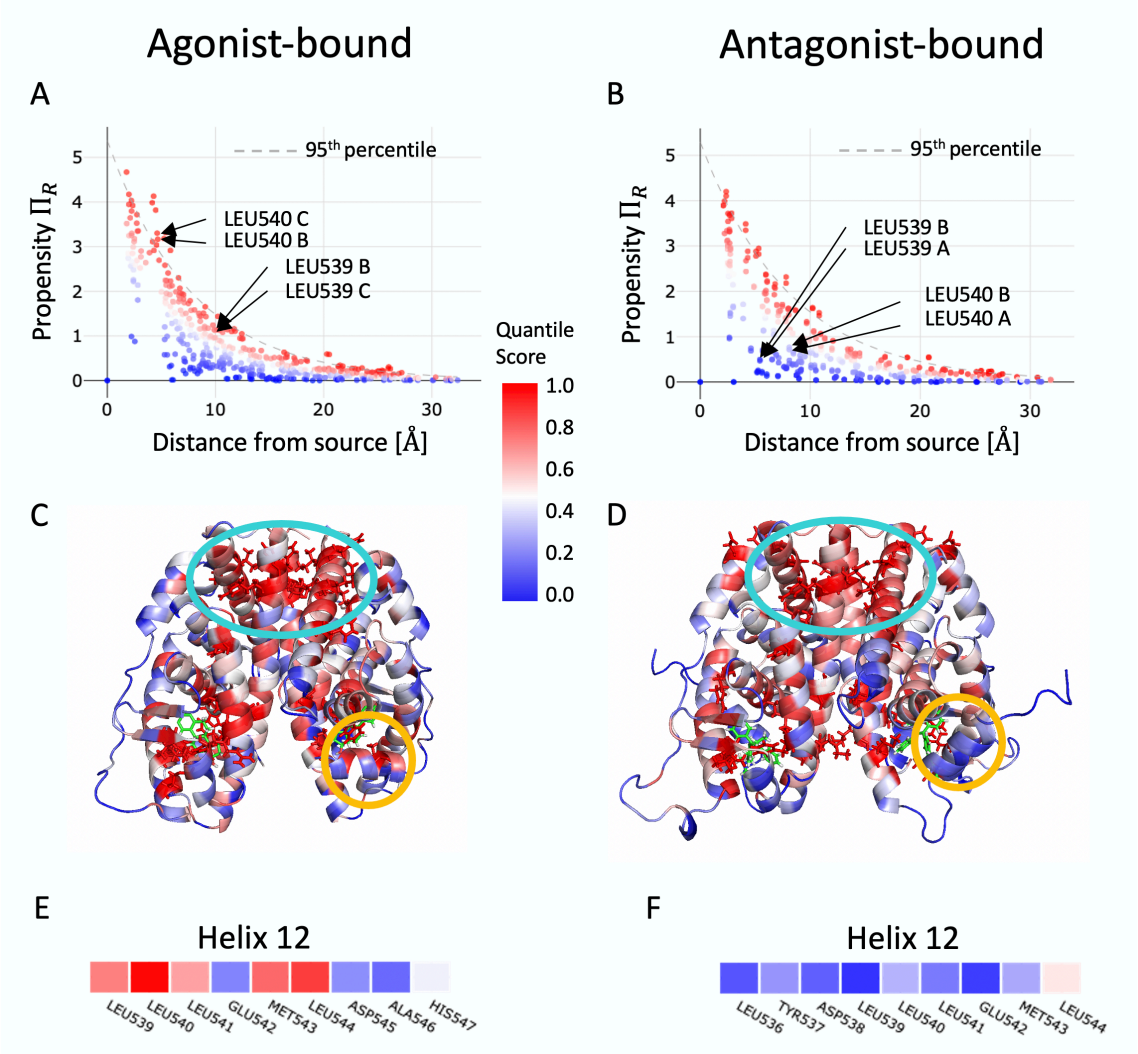


Figure 4.5: BBP analysis across agonist and antagonist-bound structures of ERα when sourced from binding site ligands. A) and B) Propensity Π_R over distance from source for each residue in ERα LBD in agonist and antagonist-bound conformations. H12 residues are highlighted for comparison. Coloured by QS 0 - blue to 1 - red. **C) and D)** Residue QS mapped onto structures of agonist (PBD id: 1G50[227]) and antagonist-bound conformations (PDB id: 3ERT[235]). Ligands are shown as green sticks, and H12 is highlighted with an orange circle. Highlighted in cyan is the upper part of the protein that contributes to the dimer interface. **E) and F)** Detailed sequence for H12 residues coloured by QS. Figure adapted from L. Strömich[202].

bound conformation to 536-544 in the antagonist-bound conformation. This structural shift complicates comparing the bond-to-bond propensity results of H12 on the single residue level.

Scoring the whole helix instead (as described in Sec. 3.1.5), allowed us to compare the helices between conformations. In the agonist-bound conformation, H12 has an average QS of 0.59. However, in the antagonist-bound conformation H12 scores with an average QS of 0.26. These results provided a first indication that bond-to-bond propensities can be a useful tool in

validating the molecular mechanism of ERα.

The previous work by the author of this Thesis[202] also investigated whether this connectivity is bi-directional and looked at whether H12 also influences the binding of estrogens and anti-estrogens. To this end, BBP analysis was sourced from H12 residues in the agonist and antagonist-bound conformations of the ERα LBD. Figure 4.6 summarises the results of this comparative analysis. It was found that the binding site ligands are differentially connected to H12. For the agonist-bound conformation, the bound EST molecules are amongst the highest scoring residues with a QS of 1.0 (Fig. 4.6A). In the antagonist-bound conformation, on the other hand, the bound OHT compounds are less connected to H12 and found to have a QS of 0.76. This lowered connectivity of H12 to the bound anti-estrogens in comparison to the bound estrogens indicates that a bi-directional signal perturbation is at place between the ligands and H12. Taken together with the results presented above, bond-to-bond propensities successfully validate the molecular mechanism in the ERα LBD where the ligand binding and subsequent positioning of H12 is crucial for AF-2 of the protein[225].

4.2.2 Importance of dimer interface connectivity

Lastly, our previously reported studies[202] included observations on the dimer interface of the ERα LBD. As described above, ERα function is dependent on dimerisation of two identical monomers into a homodimer[218–220]. When comparing the BBP results in the two conformations (Fig. 4.5C, D) we detected an area of high QSs in the upper part of the proteins (cyan circle). These areas located in the upper part of the structures contribute to the dimerisation of the protein[97] and score highly in both the agonist and antagonist-bound conformations, as shown in Figure 4.7.

From these first BBP analyses in the ERα LBD we learned that the dimerisation site plays an important role in the intra-structural connectivity of the protein, which is in line with the observation that ERα functions as an obligate homodimer[218–220]. Taken together, we found these first results to validate the molecular mechanism in ERα as described in literature[226]. These results gave us confidence that BBP analysis can be fruitfully applied in the ERα system,

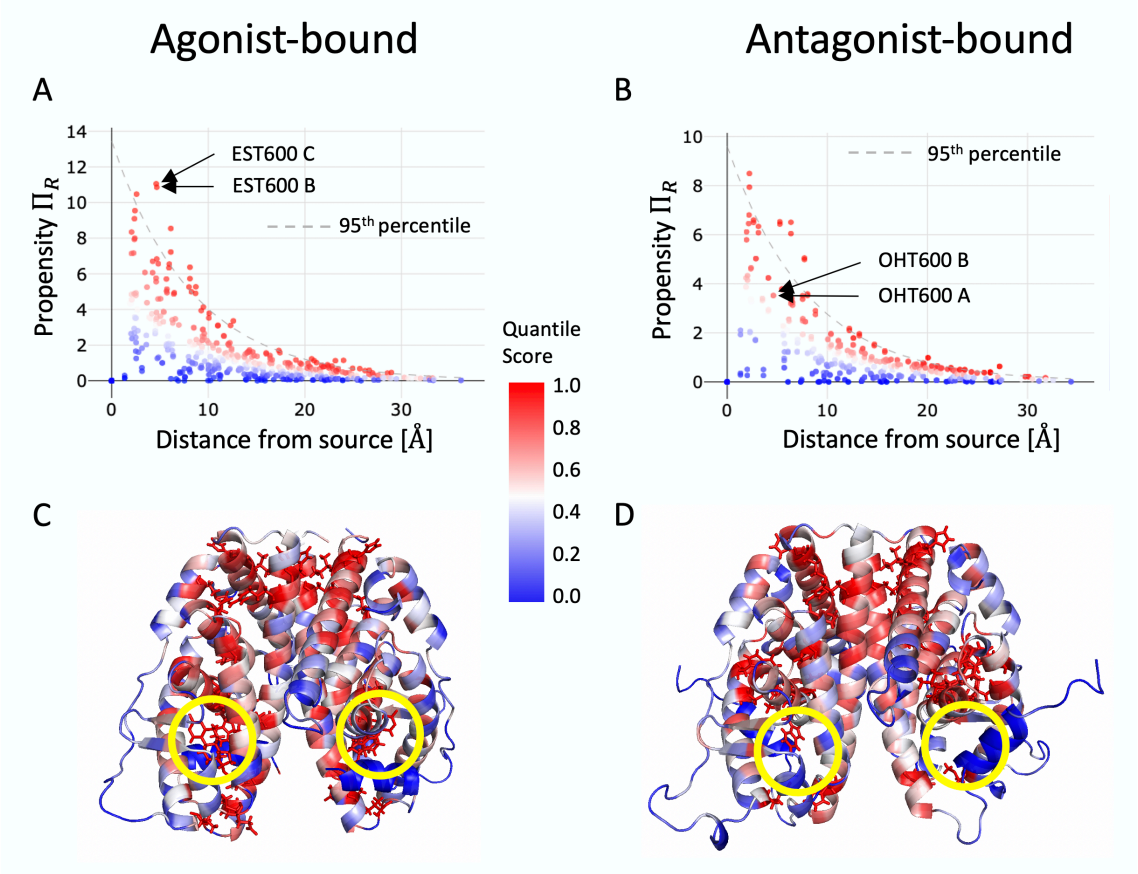


Figure 4.6: BBP analysis across agonist and antagonist-bound structures of ERα when sourced from H12. A) and B) Propensity Π_R over distance from source for each residue in ERα LBD in agonist and antagonist-bound conformations. Ligands are highlighted for comparison (EST - 17β-estradiol, OHT - 4-hyrdoxytamoxifen). Coloured by QS 0 - blue to 1 - red. **C) and D)** Residue QS mapped onto structures of agonist (PBD id: 1G50[227]) and antagonist-bound conformations (PDB id: 3ERT[235]). Ligand-binding site is highlighted in yellow. Figure adapted from L. Strömich[202].

and in this Thesis, the methodology was applied in further analyses to reveal in-depth details about the dimer interface and how it might contribute to resistance in cancer mutants.

4.3 Signal connectivity in the structural features of the dimer interface

Based on our first observations, which suggested a role of the dimer interface in signalling connectivity within the ERα LBD, we investigated the dimer residues in more detail. Figure 4.8A summarises the dimer interface (as defined by PDBePisa[78]) in the agonist and

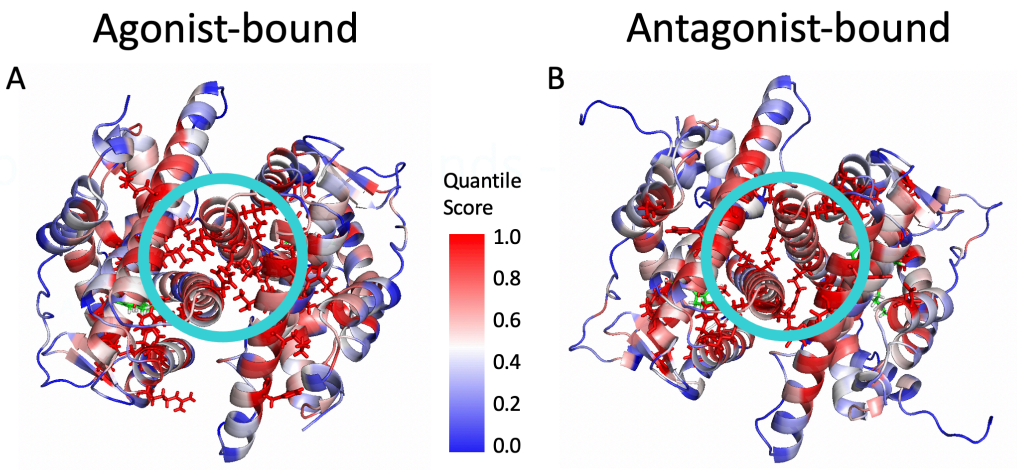


Figure 4.7: The dimer interface in ERα is highlighted by BBP analysis sourced from binding site ligands. A) and B) Residue QS mapped onto structures of agonist (PBD id: 1G50[227]) and antagonist-bound conformations (PDB id: 3ERT[235]) viewed from the top. Ligands are shown as green sticks, and the interface is circled in cyan. Figure adapted from L. Strömich[202].

antagonist-bound conformations. The dimer interface (orange in Fig. 4.8B, C) is larger in the agonist-bound conformation with 47/46 residues per monomer* vs 2x 38 residues in the antagonist-bound conformation. Helices 5/6, 8, 9, 10 and 11 contribute to the interface for both conformations. Appendix A.1.1 contains the definition of the structural features in the interface and Tables C.1 and C.2 list the interface residues in both conformations. In the agonist-bound conformation PDBePisa also detects that the C-terminal residues of H12 (HIS^{547} onwards) are part of the interface. These residues might aid in the stabilisation of H12 in the agonist-bound conformation and further strengthen the interface.

We chose to investigate the interface in relation to the bound ligands to determine how ligand binding is coupled with dimer connectivity. To this end, we sourced BBP analysis from the EST and OHT molecules in the agonist and antagonist-bound conformations, respectively.

We then determined the average QS of the interface over all residues and only over the residues that form hydrogen bonds across the interface as determined by PDBePisa[78]. In the agonist-bound conformation the whole interface scores with an average QS of 0.61, while the residues involved in hydrogen bonding (n=21) have an average QS of 0.60. In the antagonist-bound

*The difference here can be explained due to the fully solved structure which amongst others contained two missing C-terminal residues for chain C.

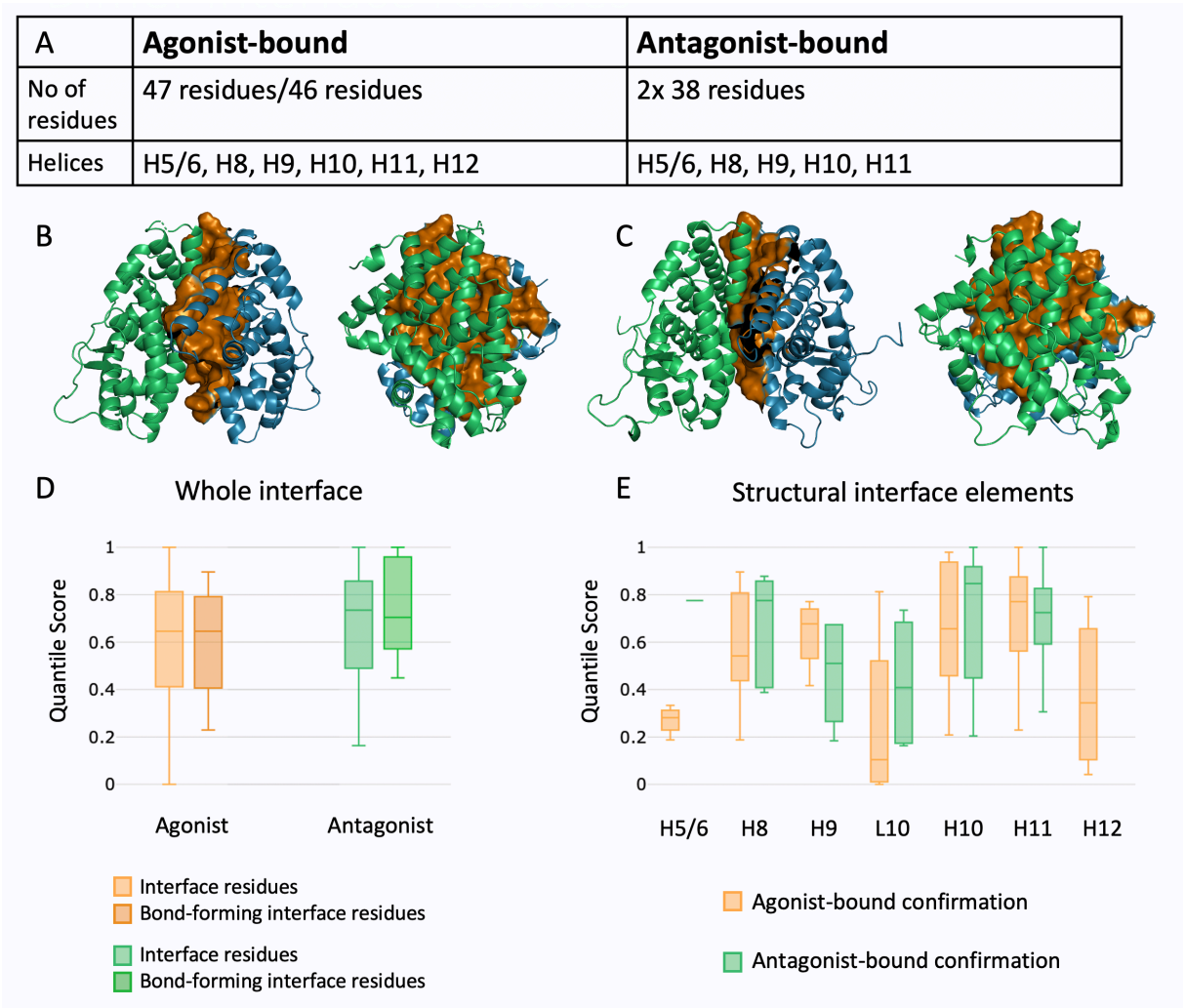


Figure 4.8: The ERα LBD dimer interface and BBP QS results in different structural features. **A)** An overview of the dimer interface in agonist and antagonist-bound conformations as determined by PDBePisa[78], residue details can be found in Tables C.1 and C.2. **B) and C)** The LDB in agonist (PDB id: 1G50[227]) and antagonist-bound (PDB id: 3ERT[235]) conformations with the two monomers shown in green and blue and the interface highlighted in orange. **D)** QS distribution of all interface residues in the agonist (yellow) and antagonist-bound (green) conformations. Residues that form hydrogen bonds or salt bridges are shown as a subcategory of all interface residues (bond-forming). **E)** QS distribution of all interface residues split into different structural features in the agonist (yellow) and antagonist-bound (green) conformations. H - helix, L – loop.

conformation we detect slightly higher average QSs of 0.67 for all residues and 0.76 for the hydrogen bond-forming ones (n=18). These results indicate that hydrogen bonds play a more important role in the dimerisation of the antagonist-bound LBD than in the agonist-bound one.

However, looking at the whole data distribution in Figure 4.8D we found that the residue QSs

are overall similarly distributed. Hence, we decided to zoom more closely into the interface and investigate the QS distribution for the individual structural elements, as shown in Figure 4.8E. It becomes apparent that some helices are more critical in the dimer connectivity than others, with H8, H10 and H11 being the highest-scoring ones overall.

One step further takes us to the single residue level, where we investigated the highest-scoring residues as listed in Table 4.1. As can be seen from the colour code, almost all residues are located on H10 and H11 in both the agonist and antagonist-bound conformation. We propose targeting these areas might have the highest potential for disrupting ERα activity. This could be done in two ways: either by targeting the highest scoring residues with a small molecule inhibitor, thereby disrupting ligand binding, or by introducing a peptide binder onto the high scoring helices. This latter strategy has indeed been proposed for H10 and 11 residues in work by Chakraborty et al.[97].

Table 4.1: Top-scoring residues in the ERα dimer interface. Top 10 residues with the highest QS in BBP analysis when sourced from different elements in the agonist and antagonist-bound conformations. EST - 17β-estradiol; OHT - 4-hydroxytamoxifen; light red - H10 residues; orange - H11 residues.

Agonist-bound		Antagonist-bound	
EST	H12	OHT	H12
HIS501 B	HIS501 B	ARG515 A	ASP484 B
HIS501 C	HIS501 C	ASP484 B	ILE487 B
THR483 C	ILE487 C	ASP484 A	LEU479 B
LYS481 C	MET437 B	ARG515 B	ILE487 A
ILE487 C	ILE487 B	ASP480 B	LEU479 A
LEU504 C	MET522 B	ASP480 A	ILE510 A
LYS481 B	LEU504 C	LYS520 B	ASP480 B
MET522 C	THR483 C	LYS520 A	THR483 B
THR483 B	MET437 C	GLN502 A	ASP484 A
ILE487 B	LEU504 B	GLN502 B	ASP480 A

The results we presented and discussed above show that the signalling that involves the dimerisation site is not conferred over all residues in the interface but rather over singular critical residues that can be determined with bond-to-bond propensities. This observation is in line with how F. Vianello[163] established the applicability of the methodology to explore protein-protein interaction sites. We propose that a change in these high scoring residues between

different conformations or ligand binding events could indicate alternate signalling paths in the protein. We explore this notion further in the following Section, which investigates the binding of different chemotherapeutics in the ERα LBD cancer mutant L536R.

4.4 Conferring resistance in cancer mutations over the dimer interface

Many mutations in the LBD of ERα have been described in the BC context and are thought to confer resistance against chemotherapeutics. The most common ones are located in the hinge region just before H12 at residues L^{536},Y^{537} and D^{538}. We investigated these cancer mutations and the cellular response to different chemotherapeutics alongside our collaborators. The experimental work was done by Fui Lai in the group of Simak Ali and provided insights into the effect of different drugs on cells expressing mutant ERα*. To investigate the inhibitory potential of different classes of anti-estrogens, the group established breast cancer cell lines (MCF-7 Luc cells) carrying an altered ESR1 gene which led to the expression of several ERα mutants at positions 536, 537 and 538. The cell cultures were then treated with increasing concentrations of chemotherapeutics over several days and cell growth was monitored. These experiments allowed to determine the half maximal inhibitory concentration (IC50) of each anti-estrogen on the respective cell type. By comparing these IC50 values with wild type MCF-7 Luc cells, the group of Simak Ali established which mutants showed an unexpected resistance pattern, as summarised in Figure 4.9A for a range of SERMs and SERDs in L536R mutant cells.

As highlighted in Figure 4.9B, Fui Lai and Simak Ali specifically detected an interesting pattern in inhibition between different SERDs. For Faslodex and AZD9496 cell growth of L536R ERα expressing cells was equally as inhibited as in the cells expressing wild type ERα. However, for AZD9833, GDC-0810 and RAD1901 cell growth was uninhibited compared to the wild type

*This work was part of a large screen of several cancer mutants. More details can be found in Section A.2 and in Figure B.3

A

	IC50 (nM)	MCF-7 Luc	L536R CL1	L536R CL2	L536R CL3
SERM	Tamoxifen	0.34	4.61	2.21	3.72
	Raloxifene	0.24	7.03	3.79	3.76
	Bazedoxifene	0.15	2.05	1.51	1.39
	Lasofoxifene	0.08	1.36	0.71	1.13
SERD	Faslodex	0.18	0.91	0.91	1.00
	AZD9496	0.13	0.65	0.53	0.52
	AZD9833	0.11	7.72	2.96	3.69
	GDC-0810	0.04	0.59	0.55	0.52
	RAD1901	0.46	17.65	8.17	10.21

	Fold Mutant clones vs WT	MCF-7 Luc	L536R CL1	L536R CL2	L536R CL3
SERM	Tamoxifen	1.0	13.5	6.5	10.9
	Raloxifene	1.0	28.8	15.6	15.4
	Bazedoxifene	1.0	13.7	10.1	9.3
	Lasofoxifene	1.0	17.7	9.3	14.7
SERD	Faslodex	1.0	5.1	5.1	5.6
	AZD9496	1.0	4.8	4.0	3.9
	AZD9833	1.0	67.9	26.1	32.4
	GDC-0810	1.0	14.6	13.6	12.8
	RAD1901	1.0	38.1	17.7	22.1

B

	Fold Mutant clones vs WT	MCF-7 Luc	L536R CL1	L536R CL2	L536R CL3
SERM	Tamoxifen	1.0	13.5	6.5	10.9
	Raloxifene	1.0	28.8	15.6	15.4
	Bazedoxifene	1.0	13.7	10.1	9.3
	Lasofoxifene	1.0	17.7	9.3	14.7
SERD	Faslodex	1.0	5.1	5.1	5.6
	AZD9496	1.0	4.8	4.0	3.9
	AZD9833	1.0	67.9	26.1	32.4
	GDC-0810	1.0	14.6	13.6	12.8
	RAD1901	1.0	38.1	17.7	22.1

C

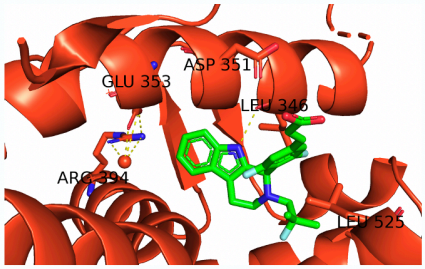
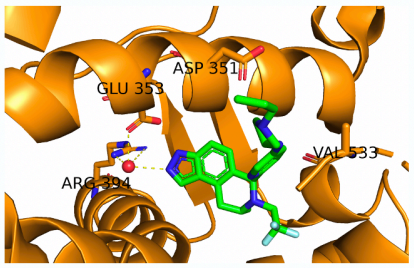
	AZD9496	AZD9833
Interactions		
Binding site	LEU346, ASP351, GLU353, ARG394, LEU525	ASP351, GLU353, ARG394, VAL533
Cancer mutant **L536R**	sensitive	resistant

Figure 4.9: Effect of chemotherapeutics on L536R ERα mutant. **A)** Two classes of chemotherapeutics were investigated regarding their inhibitory effects on breast cancer cells: selective estrogen receptor modulators (SERMs) and selective estrogen receptor degraders/downregulators (SERDs). Their half maximal inhibitory concentration (IC50) on wild type (MCF-7 Luc) cells was measured by increasing the concentration of the shown chemotherapeutics in the cell culture over 6 days. The IC50 values were also determined for three clones (CL1 - 3) carrying the L536R ERα mutant. The lower panel shows the subsequent fold differences in IC50 between WT and mutant clones. Values highlighted in orange show >10-fold difference from the IC50 value determined for MCF7 cells, with the cells in red identifying >20-fold difference in sensitivity to the drugs. **B)** Zoom into the fold changes between mutant clones and WT cells. Highlighted in green is the section that shows that L536R ERα is sensitive for treatment with some SERDs like AZD9496 but resistant to others like AZD9496. **C)** Overview of the binding characteristics of two differentially acting SERDs. Shown is the binding mode in the ERα LBD for AZD9496 (PBD id: 5ACC[255]) and AZD9833 (PDB id: 6ZOR[254]), and binding site residues are listed. Cells carrying the L536R ERα mutant were found to be sensitive to inhibition with AZD9496 but resistant to AZD9833, as shown in **(A)** and **(B)**. Experimental data was provided by Fui Lai and Simak Ali as described in detail in Appendix A.2.

cells suggesting that the cancer mutant L536R confers resistance against these chemotherapeutics. Investigating the underlying molecular mechanism of this differential resistance might aid in developing future drugs that are robust to cancer mutations.

Based on this experimental data and available structures, we decided to investigate the apparent difference between SERDs by comparing AZD9496[251] and AZD9833[254]. Figure 4.9C describes the differences between the two drugs. The binding site interactions between the drugs and the ERα LBD were determined from literature[254,255]. For AZD9496 these are five residues: L^{346}, D^{351}, E^{353}, R^{394}, L^{525}. For AZD9833 the binding mode is slightly different with only four contributing residues: D^{351}, E^{353}, R^{394}, V^{533}. Both binding modes also contain a water molecule that aids in hydrogen bonding contacts between the GLU^{353} and ARG^{294} residues. From the data provided by Fui Lai and Simak Ali, we can deduce that L536R ERα is sensitive towards inhibition with AZD9496 but shows resistance for treatment with AZD9833.

As there is no structure available for L536R, we mutated the ERα LBD (PDB id: 1G50[227]) *in silico* using PyRosetta[263]. In the absence of a bound ligand, we source BBP analysis from the binding site residues which has been found to be an adequate alternative approach[54]. Using the binding site residues as listed in Figure 4.9C allows us to simulate a binding event of these drugs to the L536R mutant ERα LBD. Figure 4.10 shows the results of this analysis where we colour the structures by residues QS. We detect a "bridge" of high scoring residues from the ligand-binding sites over the dimer interface for the run sourced from the AZD9833 binding site residues (Fig. 4.10B). This area is less connected in the analysis sourced from the AZD9496 binding site residues (Fig. 4.10A).

Table 4.2 details which residues are involved in this interface "bridge" hotspot. We list residues that are high scoring with a QS >0.95 in the AZD9833 binding site run. Towards the interface, these residues are all located on H11, where especially ASN^{519} and LYS^{520} score much lower in the AZD9496 binding site sourced analysis. The residues oriented towards the binding site are located in H5/6, H7 and H8 and the results show PHE^{425} as the residue with the highest discrepancy between the AZD9496 analysis and the AZD9833 one.

Given the experimental observations shown in Figure 4.9B we propose the detected interface

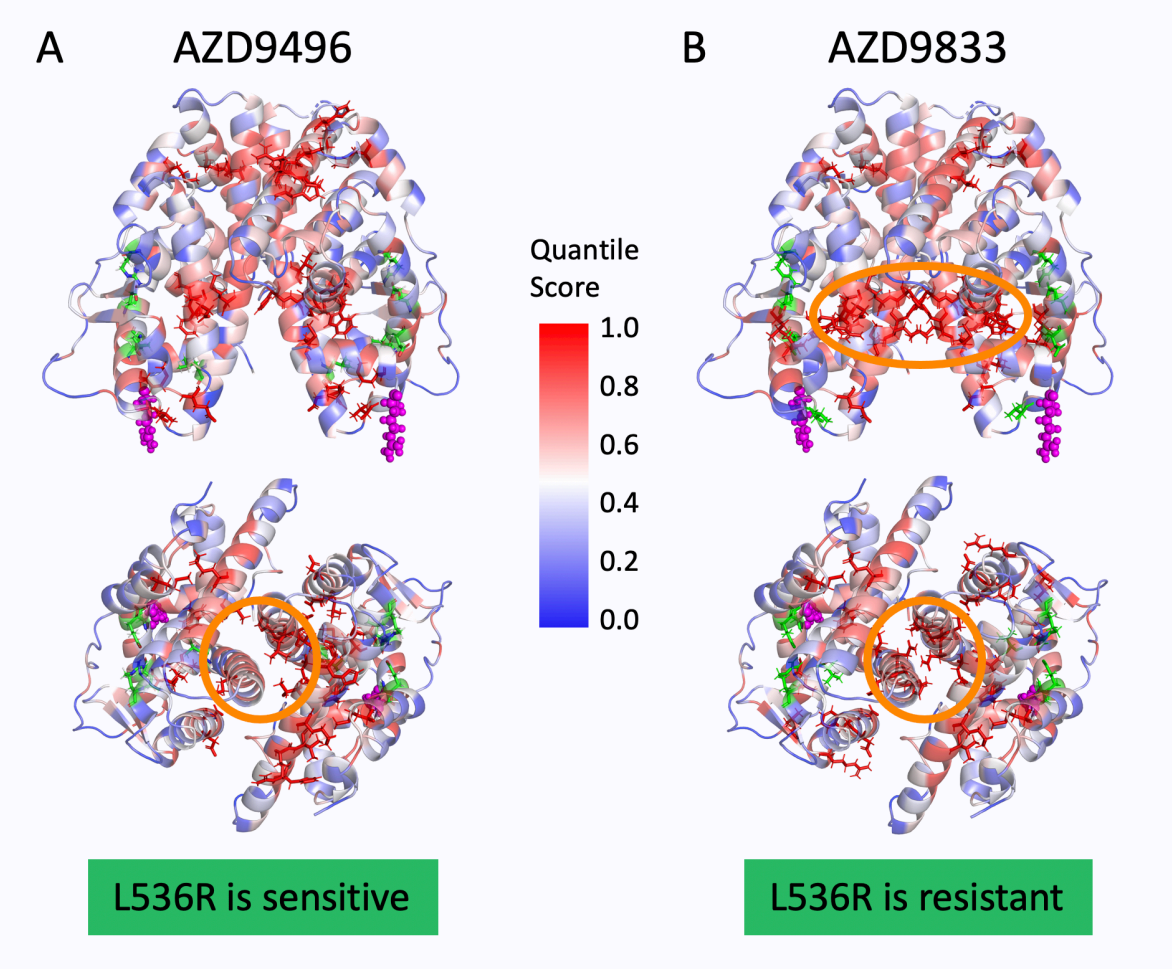


Figure 4.10: BBP analysis of the L536R ERα LBD mutant sourced from two drug binding sites. The L536R mutation (shown as pink spheres and sticks) was introduced into the ERα LBD (PDB id: 1G50[227]). The source residues are shown in green and were chosen as the binding site residues of AZD9496 (**A**) and AZD9833 (**B**). QS results are mapped onto the structures from 0 - blue to 1 - red. Highlighted in orange is the identified interface "bridge" as identified in the AZD9833 run, which is absent in the AZD9496 run. Drug response is indicated in a green box below.

"bridge" to be involved in L536R ERα resistance against AZD9833 but not AZD9496. The resistance mechanism here might be based on increased dimer stability. Our analysis further allows us to highlight single residues that might be of particular interest in modulating this dimer stability. We propose ASN^{519}, LYS^{520} and PHE^{425} as the first target points in overcoming ongoing resistance in BC tumours that show L536R mutated ERα.

Table 4.2: Interface "bridge" residues in L536R ERα LBD. QS of residues with a QS >0.95 in the BBP analysis sourced from AZD9833 binding site are listed with the equivalent values in AZD9496 analysis. The structural features in which the residues are located are indicated. The residues with the largest difference between the two runs are highlighted in blue.

Source		AZD9496 binding site	AZD9833 binding site
Towards interface H11	His516	0.93	0.99
	Met517	0.98	0.97
	Asn519	0.72	0.99
	Lys520	0.73	0.98
	Met522	0.95	1
Towards binding site H5/6, H7, H8	Trp383	0.96	0.99
	Arg412	0.91	0.96
	Phe425	0.83	1
	Asp426	1	0.99

4.5 Conclusions

The work presented here is a continuation of previous studies by the author of this Thesis[202]. One of the primary outcomes of this earlier work, was the validation that our methods can be used to investigate the LBD of ERα. Bond-to-bond propensities had confirmed the molecular mechanism of AF-2 of ERα in three parts:

- In the case of the agonist-bound conformation a connectivity between estradiol and H12 was detected which is absent in the antagonist-bound conformation (Fig. 4.5). When comparing H12 connectivity between the conformations, the results showed that H12 in the agonist-bound conformation scores more than twice as high as H12 in the antagonist-bound conformation.

- The analyses sourced from H12 detected a bi-directionality that highlighted the estradiol residues in the agonist-bound conformation which was less apparent for the tamoxifen residues in the antagonist-bound conformation (Fig. 4.6).

- A first exploration of the connection between ligand binding and dimer formation was done to explore the obligatory dimerisation in ERα[226]. To this end, BBP analysis

detected a connectivity between ligand binding and the dimer interface that was mainly present in the upper part of the protein (Fig. 4.7).

In this Thesis, we followed up on these preliminary results and characterised and investigated the dimer interface in more detail. We were interested to see whether we would detect any differences between the two conformations. The calculated QSs that were obtained from BBP analysis allowed us to investigate the dimer interface at different resolutions. We zoomed into the different structural features starting from the whole interface and ultimately onto the single residue level. In line with what is described in literature[226], we saw that both agonist and antagonist binding are connected to a strong interface connectivity in the ERα LBD. We found that overall H10 and H11 score the highest in the dimer interface (Fig. 4.8) and these areas have been previously studied for peptide inhibitors of the dimerisation process[97,98]. On the single residue level, we do detect slightly different high scoring residues between agonist and antagonist-bound conformations (Tbl. 4.1). These residues might serve as targets for selective small molecule inhibitors against the agonist-bound conformation.

On the other hand, we did not detect any paths of significant fast signalling within the protein in either conformation when employing Markov transient analysis. The signal, which is modelled by a random walker on the atomistic protein graph, is rather evenly diffusing away from the ligands. As ERα is a nuclear hormone receptor that does not have enzymatic activity, these results tie in with previous use cases of Markov Transients in the group. So far, MT analysis has been shown to be successful in detecting allosteric sites and signalling pathways primarily in catalytically active proteins, as is further shown in Chapters 5 and 6.

Lastly, we showed that our methodologies can be fruitfully applied to shed light on the molecular details of resistance mechanisms. We studied the cancer mutant L536R and found motivation in the inhibitor patterns detected by experimental work done by our collaborators Fui Lai and Simak Ali. We propose that the differential effect that certain SERDs have on the L536R ERα mutant is due to a change in dimer interface connectivities. L536R is resistant to AZD9833, and our results suggest that is due to dimer stabilisation over residues in the lower part of the interface (Fig. 4.10). Exploring this avenue further provides prospects for targeting strate-

gies that disrupt the dimer interface, and hence, overcome acquired resistance against current therapies.

Following on from here, we demonstrate the application of Markov Transients and bond-to-bond propensities in conjunction to study a highly relevant protein in the context of COVID-19 in Chapter 5. By investigating the molecular signalling within the homodimeric main protease of SARS-CoV-2, we can provide insights into alternative targeting approaches that involve the dimerisation site as well as putative allosteric hotspots. In Chapter 6, we go one step further and explore a heterodimeric protein system that is crucial for cell cycle regulation. The cyclin-dependent kinases 4 and 6 are important drug targets in BC, and their interactions with D-type cyclins provide a crucial protein interaction that can be exploited in inhibition approaches.

Chapter 5

The main protease of SARS-CoV-2

The work in this Chapter was submitted to the *Journal of Molecular Biology* and is available as a preprint with DOI: 10.1101/2020.11.06.369439. Parts of the work in this Chapter were done by Nan Wu and this is indicated throughout.

This Chapter describes the case study of another homodimeric protein that has implications for a highly topical disease. The main protease (M^{pro}) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an important drug target to combat coronavirus disease 2019 (COVID-19). We applied our atomistic graph analysis to study the protease's molecular mechanism which involves the dimer interface and identify putative allosteric hotspots. As M^{pro} is a catalytically active protein, we applied both Markov transient (MT) and bond-to-bond propensity (BBP) analyses.

5.1 A virus causing a global pandemic

Since the end of 2019, the world has been greatly impacted by a novel multi-organ disease called COVID-19. The ongoing global pandemic has seen over 285 million cases of infection

and almost 5.5 million deaths by the end of 2021*. COVID-19 is caused by SARS-CoV-2† which was first identified in patients with pneumonia in late 2019 in China[264–267].

SARS-CoV-2 belongs to the coronavirus family, members of which are responsible for a wide range of respiratory tract diseases in humans and animals. Given the close phylogeny amongst them and their ability to infect animals and humans, the potential for a zoonotic disease event has been noted before[268]. Two notable coronaviruses that have caused large outbreaks in the past two decades are severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). The severe acute respiratory syndrome (SARS) pandemic happened in 2002/2003 and resulted in approximately 8000 cases and 770 deaths‡,[269]. The Middle East respiratory syndrome (MERS) pandemic was primarily observed in 2013-2015 but smaller outbreaks have been continuously reported since[270]. Until 2021, the WHO has recorded roughly 2500 MERS infections and 880 deaths§. Although the WHO reported their status as a global threat, no vaccine or drug is available against SARS or MERS[271]. The ongoing global health emergency caused by COVID-19 renewed the interest in finding therapeutics that target coronaviruses in a safe and efficient manner.

5.1.1 Proteolytic cleavage is essential for viral replication

Coronaviruses belong to the family of RNA viruses with a positive sense, single-stranded RNA ($^+$ssRNA) genome that is enveloped in a viral capsid. The most prominent feature of the capsid is the spike protein that sticks out from its surface and prompts the name coronavirus. Figure 5.1 provides an overview of the coronavirus life cycle. The spike proteins bind to receptors on human cells and allow viruses to enter the cells where the viral genome is released. The $^+$ssRNA is in the same orientation as human mRNA, and thus, the host cell ribosomes can directly translate the viral genome into a long polypeptide. After cleavage of the polypeptide into viral proteins (discussed in more detail below), the RNA-dependent RNA-polymerase (RdRp)

*Data obtained from the official World Health Organization (WHO) COVID-19 dashboard: covid19.who.int (Accessed: 31.12.2021)

†The virus was originally named 2019-nCov for 2019 novel coronavirus but was later classified and renamed SARS-CoV-2[264].

‡WHO details on SARS: who.int/health-topics/severe-acute-respiratory-syndrome

§WHO details on MERS: who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers

complex is translocated into the endoplasmic reticulum (ER). In the next step, the transcription of viral genomic and subgenomic RNA is facilitated by the RdRp complex. The subgenomic

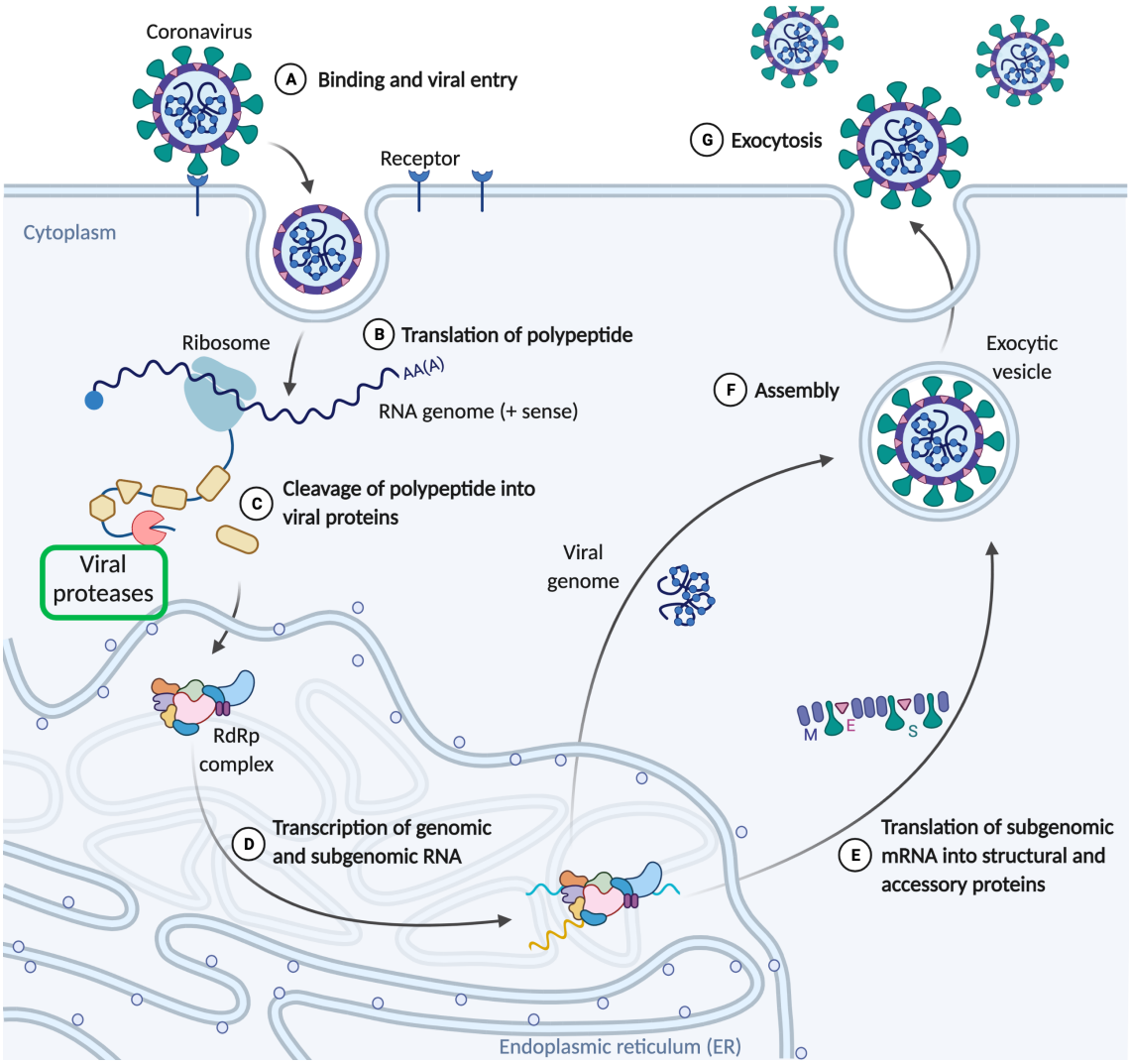


Figure 5.1: The coronavirus life cycle. A) The spike protein (S) on the virus surface allows binding to receptors on a human cell. **B)** Upon entry to the cell, the viral genome, which comes as a $^{+}$sense RNA molecule, is translated into a polypeptide by host cell ribosomes. **C)** This polypeptide is cleaved into viral proteins by two viral proteases (green highlight): papain-like protease, and the main protease which is the focus of this Chapter. **D)** The RdRp complex is assembled from viral proteins and locates into the ER. The complex is then involved in the transcription of viral genomic and subgenomic RNA. **E)** The subgenomic RNA gets translated into structural and accessory proteins. M - membrane glycoprotein; E - envelope protein; S - spike protein. **F)** During assembly, the viral RNA genome is encapsulated within structural proteins. **G)** The mature viruses are released from the host cell via exocytosis. Adapted from "Life Cycle of Coronavirus", by BioRender.com (2022)*.

RNA is subsequently translated into structural and accessory proteins that are needed for the assembly of new virions. Once enough viral material is produced, the assembly step encapsulates

*Retrieved from app.biorender.com/biorender-templates

the viral genomic RNA stabilised by nucleocapsid phosphoproteins. The outer virus shell is composed of three structural proteins: the spike protein, the envelope protein and the membrane glycoprotein. Fully assembled viruses are then released from the host cell via exocytosis[272].

One of the crucial steps of this viral replication process is the initial cleavage of the polyprotein encoded by the viral genome into functional proteins (Figure 5.1 green box). The main protease (also known as 3CL protease) cuts the polyproteins at 11 cleavage points, making it an essential protein for viral replication. Figure 5.2 shows the structure of the SARS-CoV-2 M^{pro} (PDB id: 6Y2E[5]), which follows a chymotrypsin-like fold for domains I and II and has an extra domain III that is involved in regulating dimerisation[273]. The active centre for catalysis consist of a histidine and a cysteine residue at positions 41 and 145, respectively. This catalytic dyad can be extended to a triad over a catalytically important water molecule (Fig. 5.2B). When research activity around SARS-CoV-2 picked up in the first quarter of 2020, structures for the M^{pro} were amongst the first ones to be deposited. It quickly became apparent that the SARS-CoV-2 M^{pro} is closely related to the earlier isoform of SARS-CoV. The proteins share a 96 % sequence identity as well as a close structural alignment[5] (Fig. 5.2C,E), and most residues involved in catalysis, substrate recognition and dimerisation are conserved[274]. Hence, many of the insights that were gained in the structure of the SARS-CoV M^{pro} can be transferred into the context of the new coronavirus.

Learnings from SARS-CoV M^{pro}

The first structure of the SARS-CoV M^{pro} was deposited in 2003 (PDB id: 1UJ1[275])*, and the protein has been studied since. The active site consists of HIS^{41} and CYS^{145}†, which are needed for proteolytic cleavage. The substrate binding site, which is located between domains I and II, facilitates peptide binding for which selectivity for the cleavage pattern is achieved over sub-pockets. One of these sub-pockets is formed by the interaction of GLU^{166} with the N-terminal residue SER^{1}. The interaction of the binding site with the N-terminal region, the so-called N-finger (residues 1-8), is facilitated by positioning of the N-finger between the dimer

*The first general coronavirus M^{pro} structure was released just a year earlier for the transmissible gastroenteritis (corona)virus[276].

†The residue numbering is consistent between the SARS-CoV and SARS-CoV-2 M^{pro}.

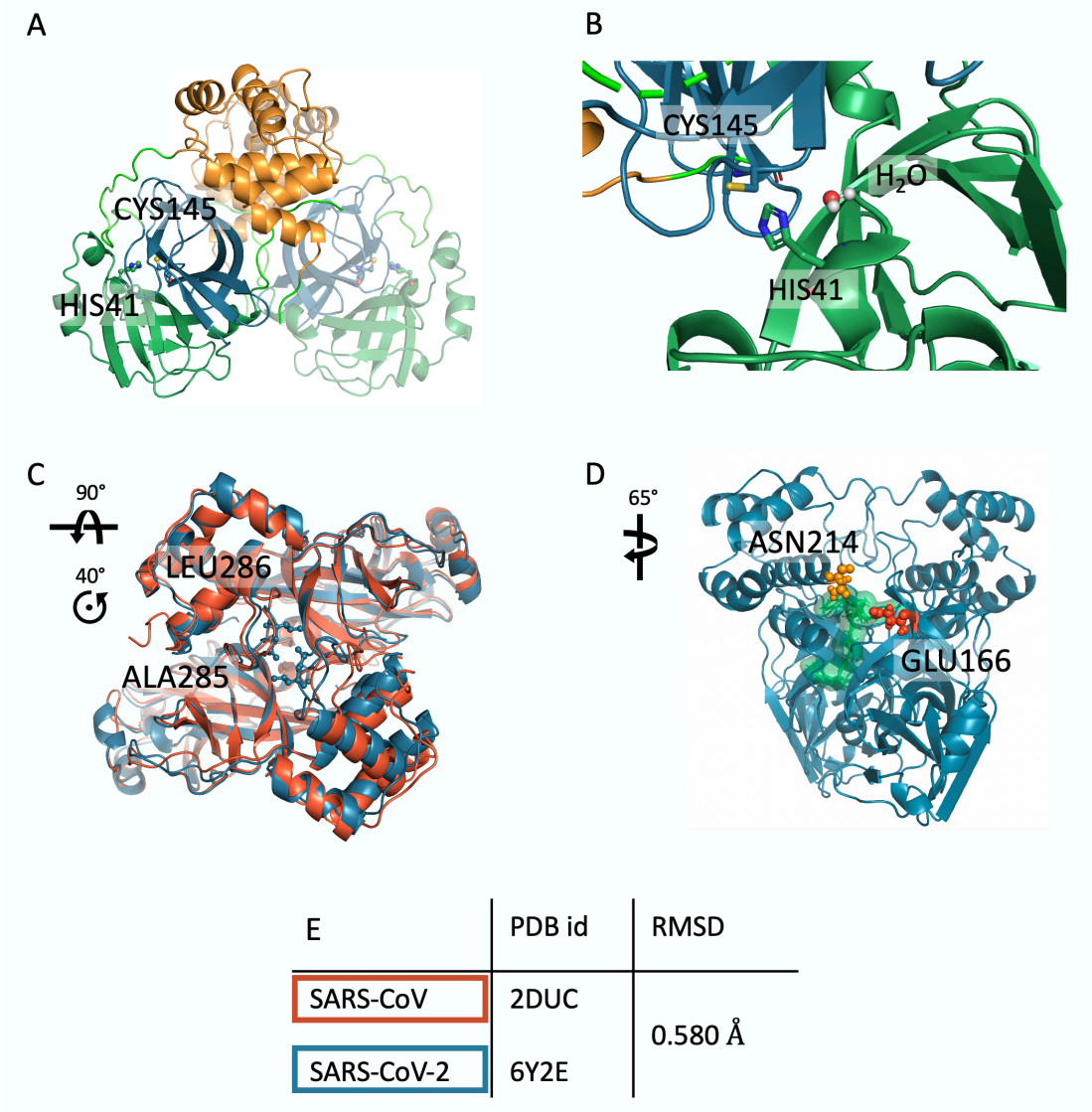


Figure 5.2: The structure of the SARS-CoV-2 M^{pro} dimer. A) The M^{pro} dimer (PDB id: 6Y2E[5]) with the active site residues on both monomers shown as sticks and spheres. One monomer is shown more transparent to visualise how the monomers come together to form a dimer. Colours are according to domain: Domain I residues 10 to 99 - dark green, domain II residues 100 to 182 - dark blue, domain III residues 198 to 303 - orange, loops in light green. **B)** Zoom-in of the active site with histidine 41 and cysteine 145 forming a catalytic dyad which is extended to a triad by a water molecule nearby. **C)** Structural alignment of SARS-CoV (red; PDB id: 2DUC[277]) and SARS-CoV-2 (blue) with two residues that are mutated between the viruses highlighted. **D)** Important structural features are highlighted on the structure of SARS-CoV-2 M^{pro}: green - N-finger (residue 1-8); orange - ASN^{214}; red - GLU^{166}. **E)** Structural details and root mean square deviation (RMSD) as calculated in PyMol[191]. Adapted from Strömich et al.[55].

halves[278], as shown in Figure 5.2D. This positioning explains the importance of dimer formation in achieving M^{pro} activity[275,279]. Dimer formation is facilitated in large parts over the extra domain (domain III, orange in Fig. 5.2A) and residues in this domain have been found to regulate dimerisation and activity of M^{pro}[273,280]. Chou et al.[280] highlighted residues ARG^{4} and GLU^{290} as forming an essential salt bridge that stabilises the dimer interface. Shi and Song[281] extended the list of functionally important residues in mutational studies to include the regions 288-290, 298-300 and ASN^{214}, of which mutations to alanine all decreased M^{pro} activity. ASN^{214} was also subject of investigation to determine its crucial role in achieving M^{pro} activity by interacting with N-finger residues[282] (Fig. 5.2D).

Another important region at the interface contains positions 284/285/286 which have been proposed to play a role in conferring dimerisation enhanced activity when mutated to alanine[281,283]. Notably, two of these positions are mutated to smaller residues from SARS-CoV to SARS-COV-2: a threonine to an alanine at position 285 and an isoleucine to a leucine at position 286 (Fig. 5.2C). These smaller residues lead to a closer dimer packing[5]. However, contrary to what was found in alanisation studies of these positions in SARS-CoV[281,283], the SARS-CoV-2 M^{pro} does not show an increased activity[5].

5.1.2 Inhibiting M^{pro} to tackle COVID-19

The main protease of SARS-CoV-2 is one of the most prominent drug targets in the fight against COVID-19. Indeed, the first United States Food and Drug Administration (FDA) approved SARS-CoV-2 specific oral medication against COVID-19 targets M^{pro}*. Developed by Pfizer under the trade name Paxlovid, nirmatrelvir binds covalently to the active site residue CYS^{145} of the SARS-CoV-2 M^{pro}[284]. Furthermore, M^{pro} is an attractive drug target due to its unique substrate cleavage pattern resulting in less side effects[285]. Given the acute motivation to find inhibitors against the SARS-CoV-2 M^{pro}, we have seen many studies, which often incorporated computational methodologies, aiming to tackle the active site (reviewed in Yang and Yang[286] and Macip et al.[287]).

*The FDA approval was announced on 22.12.2021: www.fda.gov/media/155049/download

New targeting approaches outside of the active site

In this work, we want to focus on alternative targeting approaches of the SARS-CoV-2 M^{pro}. We hope to provide an additional perspective to drug design strategies, allowing for more specificity and robustness to resistance, while widening the chemical search space.

Allostery in the previous SARS-CoV M^{pro} has been described mainly in the context of domain III impacting dimerisation and subsequently the catalytic activity[273,281–283] as detailed above. Using a computational approach, Kidera et al.[288] observed similar allosteric behaviour between domain III and catalytic activity in SARS-CoV-2 in a structural ensemble study. Work by another group[289,290] further studied the long-range impact of distal residues onto the dimerisation process and on SARS-CoV M^{pro} activity. However, to the best of our knowledge there is no description of an allosteric site in the SARS-CoV M^{pro}.

With the recent surge in research focusing on the SARS-CoV-2 M^{pro}, more attention has been given to distal regions of the protein that might impact proteolytic activity. Some distal regions have been identified in large-scale fragment and drug screens with crystallographic approaches. El-baba et al.[291] made use of an extensive data set from an X-ray crystallographic fragment screen that was published in the first quarter of 2020[292]. They focused on binding events distal from the active site and used mass spectrometry based kinetic essay to confirm allosteric inhibition of M^{pro}. They chose candidate compounds from the fragment screen which were readily available and binding non-covalently in distance from the active site. The impact of these compounds was investigated in two dimensions: disruption of the proteolytic activity and perturbation of the equilibrium between monomeric and dimeric states of the apo protein. One of the fragments they found to be allosterically active was located in the dimer interface close to the N-finger residues MET^6 and PHE^8. Their results suggested that the fragment could work in two ways: by disrupting dimerisation or by impacting activity over interactions with the N-finger[291]. In another recent fragment-based screen, which used nuclear magnetic resonance (NMR) spectroscopy, Cantrelle et al.[293] also identified a binding event at the dimer interface, which they confirmed to resemble the binding pose described by El-baba et al.[291].

Another study that used a crystallographic screening approach was done by Günther et al.[294].

The authors used a chemically more complex library of drug compounds to give a more refined insight into M^{pro} targetability. They identified two allosteric sites, one directly adjacent to the substrate-binding site, but the binding of 5 compound hits that they identified is oriented away from the active site residues towards the dimer interface. The second allosteric site is located between domains II and III, and binding here leads to a conformational change of ASP^{153} and TYR^{154}. They further validated the antiviral potential of the proposed allosteric compounds in cell-based assays[294]. For this purpose, the authors used Vero E6 cells and infected them with SARS-CoV-2. They then measured the load of viral particles to determine whether a compound could disrupt viral replication. They found that one of the compounds binding to allosteric site 1 (pelitinib) had a high antiviral activity, while the compound binding to allosteric site 2 (AT7519) was moderately active. Independently, Du et al.[295] proposed allosteric inhibition of the SARS-CoV-2 M^{pro} with repurposed drugs and used docking to model a binding pose between domain II and III.

Computational approaches to study allosteric sites in SARS-CoV-2 M^{pro}

Due to the relative novelty of the field, experimental literature on allosteric regions and regulation in the SARS-CoV-2 M^{pro} is still limited. But that has been more than compensated by the wealth of computational studies aiming at providing insights into allosteric behaviours of M^{pro}. As described in the previous Section, preliminary studies have detected allosteric sites are primarily located between domains II and III or on the dimer interface. The methodologies used to identify allosteric sites range from molecular dynamics (MD)[296–299] to normal modes in elastic network models[300,301]. Furthermore, repurposing studies using molecular docking and subsequent MD on these allosteric regions have been published[302–309]. However, the surge of docking approaches, in particular, should be taken with a grain of salt as they were recently found unreliable in many studies against the SARS-CoV-2 M^{pro}[62].

Objective

This Chapter demonstrates the application of our methodologies to a topical study case in the context of a highly relevant disease. As the main protease is essential for replication of SARS-

CoV-2, a detailed understanding of intra-protein connectivity might shed light on new targeting approaches. Given the catalytic activity of coronavirus proteases, we also see grounds to apply Markov Transients in addition to bond-to-bond propensities to study the protein structure.

Similar to what was shown in Chapter 4, we explored the interface connectivities in the obligate homodimeric protease. We built upon the knowledge provided by studies of the earlier SARS-CoV protein isoform and focus on mutated residues that impact the dimerisation. We aimed to provide insights into the differences between the two isoforms and how these differences are related to the dimer interface.

Furthermore, we demonstrated another feature of our methodologies by exploring allosteric signalling in the SARS-CoV-2 M^{pro}. Exploiting allosteric behaviours is a valuable alternative drug targeting approach as discussed in Section 1.3.1, and we decided to elucidate these concepts in the setting of a viral protease. This approach is in line with work done by our group in various study cases and allosteric benchmarking sets over the years[51,54,149,159,161,162]. By applying BBP and MT analyses on M^{pro} we aimed to identify allosteric sites that allow modulation of the proteolytic activity and could be a strategy for a drug against COVID-19.

5.2 Insights into the molecular mechanism of the SARS-CoV-2 M^{pro} dimer

We chose to investigate the apo form of the SARS-CoV-2 M^{pro} to get an idea of the molecular mechanisms of the protein that are intrinsically encoded in the native structure. M^{pro} is a functionally obligate dimer, and thus available structures are of the protein in dimeric form, as shown in Figure 5.2A. For the study of the SARS-CoV-2 M^{pro}, we considered one of the first solved structures in apo form, which was deposited in March 2020 with the PDB id 6Y2E[5] at a high resolution of 1.75 Å. The dimer was modelled using the symmetry information in the `.pdb` file from a monomeric protein chain. Further into this Chapter we also present results on the apo form of the SARS-CoV M^{pro} for which the PDB structure 2DUC[277] with a resolution

of 1.70 Å was chosen. This entry contained the fully solved dimeric protein which resulted in slight structural differences between the dimer halves and all results presented here are of the average between the two chains. For further details on how the PDB structures were processed, see Appendix A.1.2.

In a first explorative step, we sourced our analysis from the active site residues HIS^{41} and CYS^{145} in both dimer halves (Fig. 5.2B). Bond-to-bond propensities then allow us to find areas of the protein that are strongly coupled to the active site and might provide starting points for regulation of M^{pro}. Figure 5.3 summarises the BBP analysis results.

We identified two main areas of interest, by investigating residues with a high residue QS. The first protein area is located towards the back of the protein with respect to the active site and stretches over domains I and II (Fig. 5.3A). With a QS of 0.98 or above, LYS^{100}, TYR^{101} and PHE^{103} are especially high scoring residues. We discuss this highly connected area in more detail in the following Sections.

The second area showing high scoring residues is located at the dimer interface and contains residues from both monomers. SER^{1} and ARG^{4} from one monomer and HIS^{172} and GLU^{290} from the other monomer are scoring highly with a QS of 0.97. SER^{1} and HIS^{172}, as well as ARG^{4} and GLU^{290}, form interface bridges that are important for the catalytic activity of M^{pro} [280]. SER^{1} and ARG^{4} are further part of the N-finger, which is packed between the dimer halves and contributes to the peptide binding area [276].

These first results indicated that the dimer interface was picked up as an area of interest for the activity of the SARS-CoV-2 M^{pro} with our methodologies. We found further indications for the importance of the dimerisation interface by investigating mutations between SARS-CoV and SARS-CoV-2, as described below.

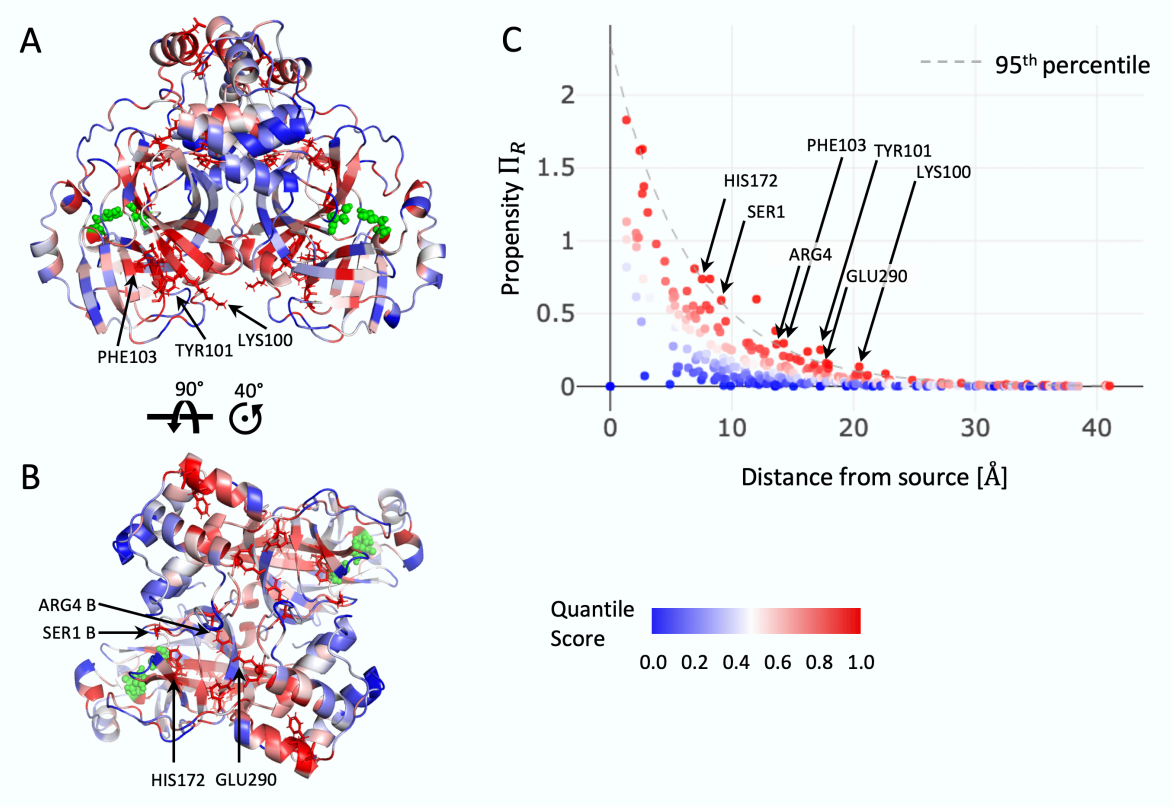


Figure 5.3: BBP analysis of M^{pro} sourced from active site residues. A) and B) Residue QS from BBP analysis mapped onto the structure of SARS-CoV-2 M^{pro} (PDB id: 6Y2E[5]) in two orientations. Source residues shown in green, and high scoring residues (QS > 0.95) shown as sticks. Two areas of interest were identified for which residues are indicated. **C)** Residue wise data distribution of BBP over distance from the source. The same residues are indicated. Adapted from Strömich et al.[55].

5.3 The dimer interface under the regulation of mutations

SARS-CoV is the coronavirus that was responsible for the 2003 SARS pandemic and has been studied since. At the beginning of the pandemic in early 2020, drawing on the knowledge that was gathered around structures of SARS-CoV M^{pro} was a valuable approach as only few structures were solved for the SARS-CoV-2 isoform. The two proteins are also closely related with 96 % sequence similarity[5] and close structural alignment (RMSD of 0.58 Å) as shown in Figure 5.2C and E. Taken together, the conserved sequence and structure draw particular attention to residues that are mutated between the isoforms as they might exhibit functional relevance. Several of these mutations are located at the dimer interface and have been picked up as high scoring in the BBP analysis described above. ALA^{285} (QS: 0.77) and LEU^{286} (QS: 0.77) are two residues of interest as they have been described to be responsible for closer dimer packing[5]. The equivalent residues in SARS-CoV (THR^{285} and ILE^{286}) have also been shown to have an impact on catalytic activity when mutated alongside position 284[283].

Given the importance of positions 285 and 286 and their mutational change from SARS-CoV to SARS-CoV-2, we chose to investigate their impact on the dimer interface in M^{pro}. The dimer interface, as shown in Figures 5.4A, B, was defined with PDBePisa[78] and was found to be smaller in SARS-CoV (41 residues per monomer, Tbl. C.4) than in SARS-CoV-2 (52 residues per monomer, Tbl. C.3). To understand the impact of the mutated residues, we ran BBP analysis from positions 285/286 in both dimer halves. Figures 5.4C, D show the residue QS results mapped onto the structures and provide a first comparison between the two structures. Although, we detect large overlaps of the hot and cold scoring regions, we find differences in the dimer interface connectivities.

Similar to what we found in Chapter 4, we do not detect a difference on the whole dimer interface level (Fig. B.5) but instead find shifts in high scoring residues. As shown in Figure 5.4E we detect a higher proportion of dimer interface residues amongst subsets of the highest scoring residues for SARS-CoV-2 than for SARS-CoV. When looking at the top 100 highest

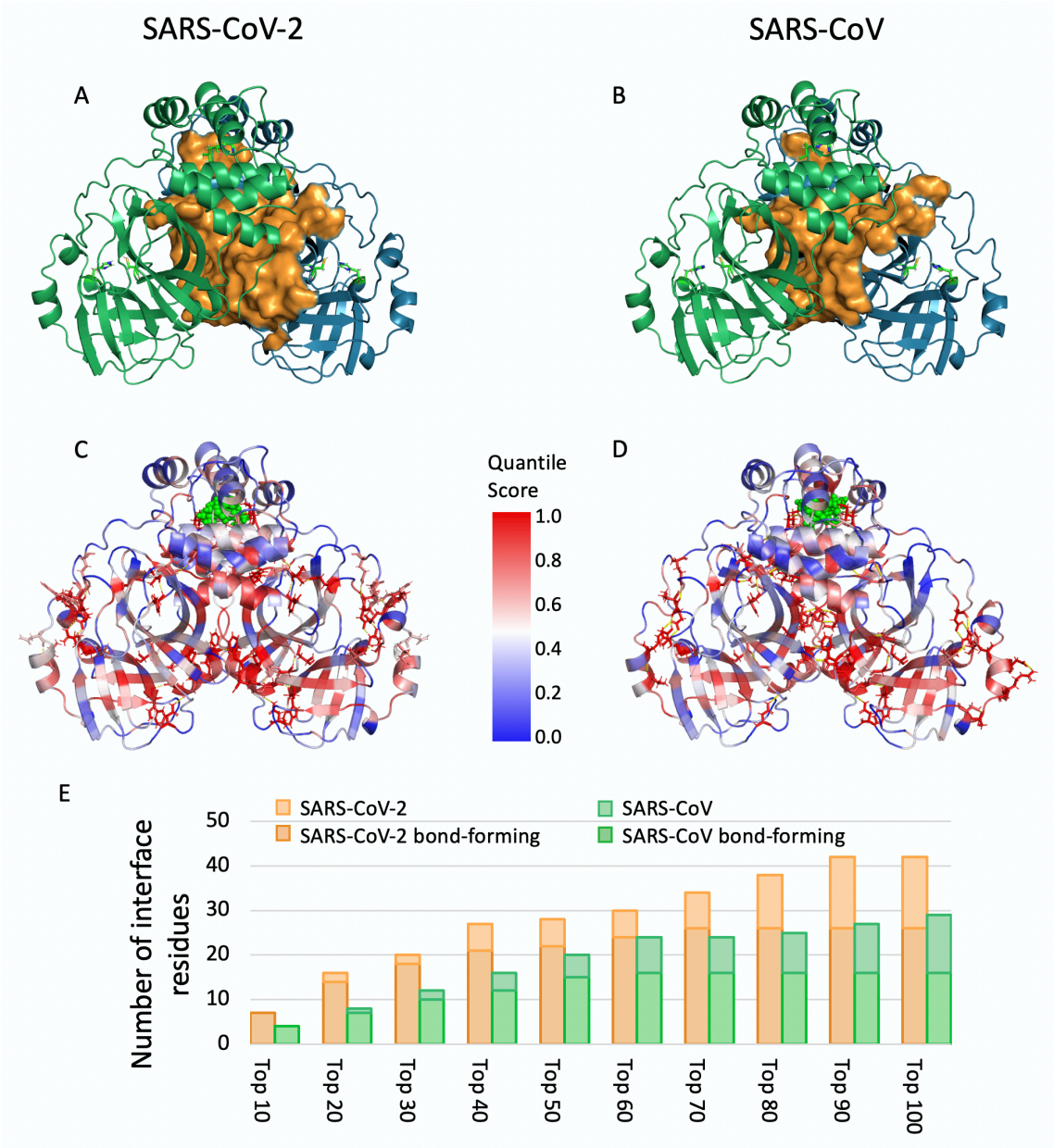


Figure 5.4: Differences in dimer interface between SARS-CoV-2 and SARS-CoV. A) and B) Dimer interface (orange) is formed between two identical monomers (green and blue) in the structures of SARS-CoV-2 (PDB id: 6Y2E[5]) and SARS-CoV (PDB id: 2DUC[277]). Dimer interfaces were calculated with PDBePisa[78]. **C) and D)** BBP analyses sourced from positions 285 and 286, which are mutated between the two viruses. Structures are coloured by residue quantile score; source residues are shown as green spheres and high scoring residues (QS >0.95) as sticks. **E)** Number of interface residues in the top 10 to top 100 highest scoring residues in SARS-CoV-2 (orange) and SARS-CoV (green). Residues that form a bond in the interface are a subset of all interface residues.

scoring residues, we find 42 are interface residues in SARS-CoV-2, while in the SARS-CoV M^{pro} only 30 interface residues are amongst the top 100. We distinguish between residues involved in forming bonds (hydrogen bonds or salt bridges) and ones that are not as defined by

PDBePisa[78]. Our results suggest that the smaller residues ALA and LEU at position 285/286 for SARS-CoV-2 in contrast to THR and ILE for SARS-CoV, lead to a strengthened interface connectivity. These smaller residues were proposed to lead to a higher activity of the protease in mutational studies of the SARS-CoV M^{pro}[283]. However, a comparison of the wild type protein activity between SARS-CoV-2 and SARS-CoV did not confirm a heightened activity in the new viral protease[5].

In line with what we proposed in Chapter 4, our methodologies allow us to pick out particularly high scoring residues in the interface that might be of importance. Table 5.1 lists the top 10 scoring residues in the BBP analysis sourced from residues 285/286. We find that for SARS-CoV-2 and SARS-CoV, some of the N-finger residues appear in this list: SER^{1}, PHE^{3} and ARG^{4}. These residues are critical in modulating the catalytic activity, as their interface packing brings them into contact with GLU^{166} in the opposite monomer, which is also picked up as high scoring in our results. The interaction between the N-finger (especially SER^{1}) and GLU^{166} leads to the formation of an extended binding pocket for substrate recognition[276], and this connection was revealed for both SARS-CoV-2 as well as SARS-CoV with our methodologies.

Table 5.1: Top scoring residues in the M^{pro} dimer interface. Top 10 residues with the highest QS in BBP analysis when sourced from residues 285/286 in the SARS-CoV-2 and SARS-CoV M^{pro}. Orange - N-finger residues.

SARS-CoV-2	SARS-CoV
ARG4 B	GLU166 A
ARG4 A	PHE305 A
SER1 A	PRO122 B
SER1 B	SER123 A
PRO122 B	ARG4 B
PRO122 A	PHE3 B
GLN306 B	SER1 B
PHE3 A	SER10 B
PHE3 B	GLN299 B
GLU166 A	VAL125 B

Another residue that has been identified in SARS-CoV to be of high importance for the catalytic machinery of M^{pro} is ASN^{214}[282]. The MD studies by Shi et al.[282] showed that the catalytic impact of ASN^{214} is conferred over the N-finger residues. We pick up ASN^{214} as high scoring in SARS-CoV-2 (QS: 0.85) as well as SARS-CoV (QS: 0.73) and further identified the N-finger

as important for connectivity in the dimer interface as shown in Table 5.1.

Taken together, our results suggest that the dimer interface plays a vital role in conferring M^{pro} activity. Our analysis sourced from the active site residues picks up many of the residues in the dimer interface that have been described before to play a role in modulating the catalytic activity. However, from the opposite direction when sourcing from interface residues ALA^{285} and LEU^{286}, we did not detect an immediate link to the active site residues HIS^{41} (QS = 0.22) and CYS^{145} (QS = 0.35). We instead describe a two-step connectivity from the dimer interface residues 285/286 over the N-finger residues towards the extended peptide-binding pocket that seems to be at play here. Although we see a strengthened dimer interface connectivity in SARS-CoV-2, we propose the overall activation dynamics that have been studied at length in SARS-CoV[281–283] are transferable onto the SARS-CoV-2 M^{pro}.

5.4 Identification and scoring of putative allosteric sites

The previous Section discussed the connectivity towards and within the dimerisation interface and bond-to-bond propensities revealed important residues that are involved in conferring the catalytic activity of the SARS-CoV-2 M^{pro}. Targeting these residues or disrupting the dimerisation process might be fruitful targeting strategies for drugs against COVID-19. Another avenue of drug targeting is the modulation of protein activity over binding at allosteric sites. We have shown the application of our methodologies for the purpose of allosteric site prediction in many study systems[54,149,159,161] and established a web server to allow the study of allosteric behaviour by the community[51]. We predict allosteric hotspots in the main protease of SARS-CoV-2 with BBP and MT analyses.

5.4.1 Bond-to-bond propensities identify a hotspot in the dimer interface

In Section 5.2, we identified two main areas of interest which contain high scoring residues. Investigating these areas in more detail and taking into consideration the surface exposure allowed us to identify allosteric hotspots. We further applied a scoring step with a structural bootstrap as described in Section 3.1.5 to assess the statistical significance of identified hotspots. This first identification step and hotspot scoring was done by Nan Wu. Downstream scoring steps of the active site and all visualisations, were performed by the author of this work.

Allosteric hotspot 1 is located at the back of the monomer in relation to the active site and contains residues from domains I and II as shown in Figure 5.5A. Hotspot 1 (highlighted in yellow in the figure) contains nine residues that are listed in Table 5.2 with corresponding QS. We also list the solvent-accessible surface area (SASA) of each residue to provide an indication of their exposure and targetability. Overall, hotspot 1 has an average QS of 0.97, which is significantly higher than an average QS of 0.53 (95 % CI: [0.53-0.54]) that a random site of the same size would have.

Allosteric hotspot 2 stretches over both monomers as it contains residues that are located at the dimer interface (shown in pink, Fig. 5.5B). Amongst other residues (listed in Tbl. 5.2), hotspot 2 contains ARG^{4} in contact with GLU^{290} of the respective second monomer. The salt bridge that is formed between these residues is important for the dimerisation and activity of M^{pro} [280]. Hotspot 2 has an average QS of 0.96 in comparison to a random site score of 0.52 for a site of the same size (95 % CI: [0.51-0.53]).

We further investigated whether we detect a bi-directionality in connectivity between these identified hotspots and the active site. For this purpose, we performed BBP analyses when sourced from the hotspot residues. When sourcing the analysis from the hotspot 1 residues, we detect an average QS of 0.64 for the active site*. This is above a random site score of 0.47

*As a scoring of only two residues provides a very narrow picture of the connectivity, we used a definition as described in Section A.1.2 for the active site for all site scoring analysis.

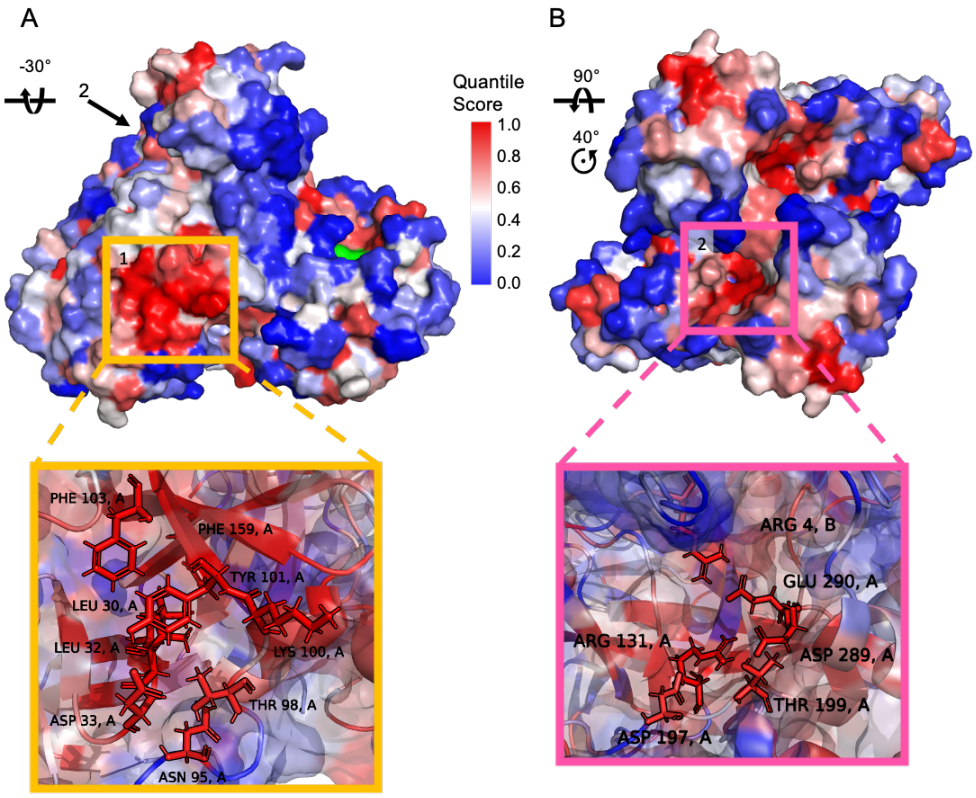


Figure 5.5: Allosteric hotspots in SARS-CoV-2 M^{pro} identified with BBP analysis. Surface representation of SARS-CoV-2 M^{pro} (PDB id: 6Y2E[5]) in two orientations and coloured by QS from BBP analysis sourced from active site residues (green). **A)** Hotspot 1 (yellow) is located on the back of the monomer in relation to the active site in domain I and II. **B)** Hotspot 2 (pink) is located in the dimer interface bridging both monomers. We provide a zoom into the hotspots with a transparent surface to highlight important residues for which full details are given in Table 5.2. Adapted from Strömich et al.[55].

(95 % CI: [0.47-0.48]) and lets us conclude that the bi-directional coupling between hotspot 1 and the active site is significant.

The same analysis for hotspot 2 results in an average QS of 0.49 for the active site which is only slightly above a random site score of 0.48 (95 % CI: [0.47-0.48]). We conclude that there is no straight coupling between hotspot 2 in the dimer interface and the active site. This is conclusive with our findings about the dimer interface reported above, where we show that the coupling between the interface and the active site is conferred over the N-finger residues. Hence, we think that hotspot 2 might still be an interesting target point for drug design as it provides the possibility to indirectly impact the active site catalysis. Furthermore, this hotspot might provide scope for disrupting the dimerisation, which is essential for M^{pro} activity.

Table 5.2: Allosteric hotspots in M^{pro} as determined with BBP analysis. QSs are given for each residue and solvent-accessible surface area (SASA) was determined in PyMol[191].

Hotspot	Residue	QS	SASA [Å^2]
Hotspot 1	LEU30	0.98	0.00
	LEU32	1.00	0.00
	ASP33	0.96	137.86
	ASN95	0.92	13.79
	THR98	0.92	72.01
	LYS100	1.00	293.05
	TYR101	1.00	118.61
	PHE103	0.98	128.62
	PHE159	1.00	0.00
Hotspot 2	ARG4	0.97	68.49
	ARG131	0.99	36.50
	ASP197	0.89	101.57
	THR199	0.93	65.26
	ASP289	0.98	27.98
	GLU290	0.97	25.86

5.4.2 Markov transient analysis identifies two more hotspots

Motivated by the catalytic activity in M^{pro} and the complex communication patterns that we detected between the active site, the dimer interface and the N-finger, we chose to apply Markov Transient analysis to the structure of the SARS-CoV-2 M^{pro}. We sourced our analysis from the active site residues HIS^{41} and CYS^{145} to obtain complementary insights to the BBP analysis discussed above.

Figure 5.6 summarises the results that were obtained by MT analysis. Markov Transients highlighted primarily one area of interest, which stretches over domains I and II in the back of the monomer with respect to the active site (Figure 5.6A). In domain I, we detect VAL^{35} and ASP^{92} with a QS of 0.95 as the highest scoring residues. We find CYS^{156} with a QS of 1.0 and ASP^{153} with a QS of 0.98 amongst the top scoring residues for domain II.

When investigating this large area in more detail, we identified two allosteric hotspots that are shown in Figure 5.6C.

Allosteric hotspot 3 is shown in cyan in Figure 5.6C is located on the back of the monomer

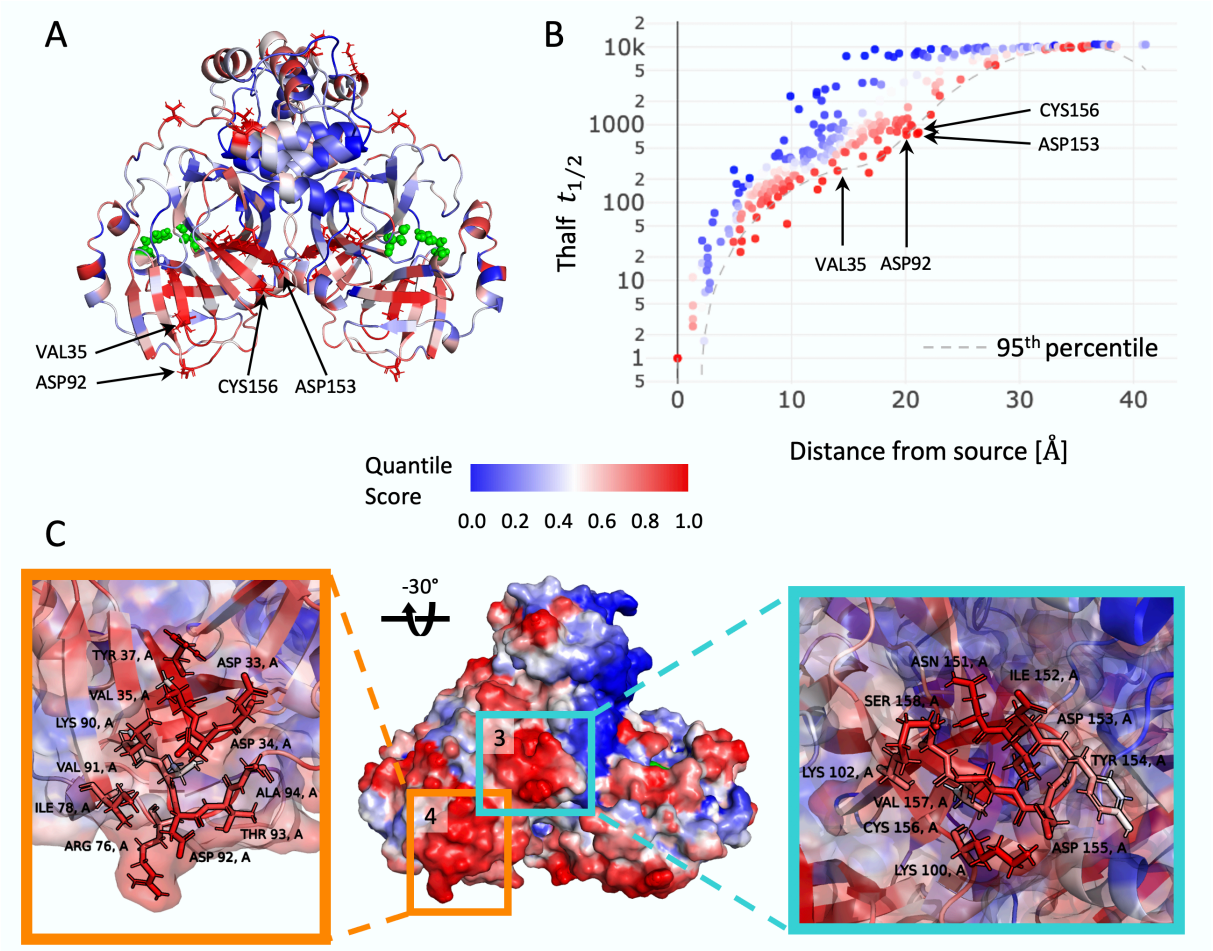


Figure 5.6: MT analysis of M^{pro} sourced from active site residues and identification of allosteric hotspots. A) Residue QS from MT analysis mapped onto the structure of SARS-CoV-2 M^{pro} (PDB id: 6Y2E[5]). Source residues shown in green, and high scoring residues (QS > 0.95) shown as sticks. Two hotspot areas of interest were identified for which residues are indicated. **B)** Residue wise logarithmic data distribution of $t_{1/2}$ over distance from the source. The same residues are indicated. **C)** Surface representation of the structure coloured by atom QS. Both hotspots are located on the back of the monomer in relation to the active site. Hotspot 3 (cyan) is located in domain II. Hotspot 4 (orange) is located in domain I. We provide a zoom into the hotspots with a transparent surface to highlight important residues for which full details are given in Table 5.3. Adapted from Strömich et al.[55].

with respect to the active site and is formed exclusively by domain II residues as listed in Table 5.3. Interestingly, the highest scoring residue at position 156 is a cysteine which could be a valuable target point for covalent drug binding. Overall, hotspot 3 has an average QS of 0.87, significantly higher than a random site score of 0.50 (95 % CI: [0.49-0.50]).

Allosteric hotspot 4 is formed by 11 residues in domain I and is highlighted in orange in Figure 5.6C. Table 5.3 lists all residues in hotspot 4, which average to a site QS of 0.87. Again, this hotspot scores significantly higher than a random site of the same size would with a QS of

0.49 (95 % CI: [0.49-0.50]).

Table 5.3: Allosteric hotspots in M^{pro} as determined with MT analysis. QSs are given for each residue and solvent-accessible surface area (SASA) was determined in PyMol[191]. Highlighted in blue is a cysteine residue that can be targeted for covalent binding.

Hotspot	Residue	QS	SASA [Å^2]
Hotspot 3	LYS100	0.89	145.96
	LYS102	0.75	113.04
	ASN151	0.97	31.34
	ILE152	0.93	5.95
	ASP153	0.98	113.04
	TYR154	0.59	132.10
	ASP155	0.92	25.18
	CYS156	1.00	24.76
	VAL157	0.76	0.00
	SER158	0.89	15.97
Hotspot 4	ASP33	0.93	68.93
	ASP34	0.93	45.79
	VAL35	0.95	19.56
	TYR37	0.85	21.65
	ARG76	0.83	170.23
	ILE78	0.85	93.58
	LYS90	0.82	89.29
	VAL91	0.64	0.71
	ASP92	0.95	80.04
	THR93	0.87	60.17
	ALA94	0.90	63.23

For hotspot 3 and 4 we follow the same reverse approach as we did for hotspot 1 and 2 and source an MT analysis from the hotspot residues. The data we obtain from these runs is subsequently used to score the active site to investigate bi-directional connectivity. When we source MT analysis at hotspot 3, we obtain an active site score of 0.66, which is above a random site score of 0.53 (95 % CI: [0.52-0.53]). These results indicate a reciprocal connectivity between hotspot 3 and the extended active site (defined in Sec. A.1.2). In the case of hotspot 4 as the source, the active site scores with an average QS of 0.52, almost the same as a random site score of 0.50 (95% CI: [0.50-0.51]). We follow that there is no immediate link between hotspot 4 and the active site. However, previous studies in multimeric proteins by members of our group[162,163] suggest that there might be another dynamic or structural element at work that is yet to be uncovered.

Although studied for a much longer time, to the best of our knowledge there have been no reports of allosteric sites in the M^{pro} of SARS-CoV. Hence, we chose to cross examine our findings in the structure of the SARS-CoV M^{pro}. Apart from the differences discussed in the context of the dimer interface in Section 5.3, we do not detect major shifts in the overall connectivity patterns found by BBP and MT analyses. Indeed, the detected allosteric hotspots seem to be consistent between the two isoforms, as shown in Tables C.5 and C.6 for all hotspot residues and as a visual comparison in Figure B.4. These results provide preliminary indications that these identified hotspots might find applications in more than one disease caused by coronaviruses.

5.4.3 Indications for hotspot targetability

The identified allosteric hotspots might provide valuable starting points to develop drugs for the treatment of COVID-19. We take our results one step further and try to provide first insights into the targetability of the found putative allosteric sites. To this end, we use data produced in the first half of 2020 by the Diamond Light Source in Oxford*. The project produced an extensive PDB data set of small fragments bound to M^{pro} and we made use of this structural data set as described below. We further cross-check our results with recent studies of the SARS-CoV-2 M^{pro} that found indications of allosteric sites in crystallographic drug screens[294] and mass spectrometry activity experiments[291].

The Diamond Light Source X-ray crystallographic fragment screen led to the deposition of 96 PDB structures with small fragments bound to the SARS-CoV-2 M^{pro} [292]. Out of these 96 fragments, 48 were covalently bound to the active site, and 23 were non-covalent active site binders. However, for this work we are more interested in fragments that bind distal from the active site. Nan Wu identified 25 of such distal binding hits. Subsequently, these were further narrowed down to 15 fragments with atoms within 4 Å of any of the identified hotspots. Figure 5.7 provides an overview of where these fragments bind relative to the identified hotspots. To model the effect of a binding event at our identified hotspots, we use the structures of M^{pro}

*The main protease project at the Diamond Light Source: www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html

bound to relevant small fragments as a proxy. We then sourced BBP and MT analyses from the small fragments and evaluated their connectivity towards the active site.

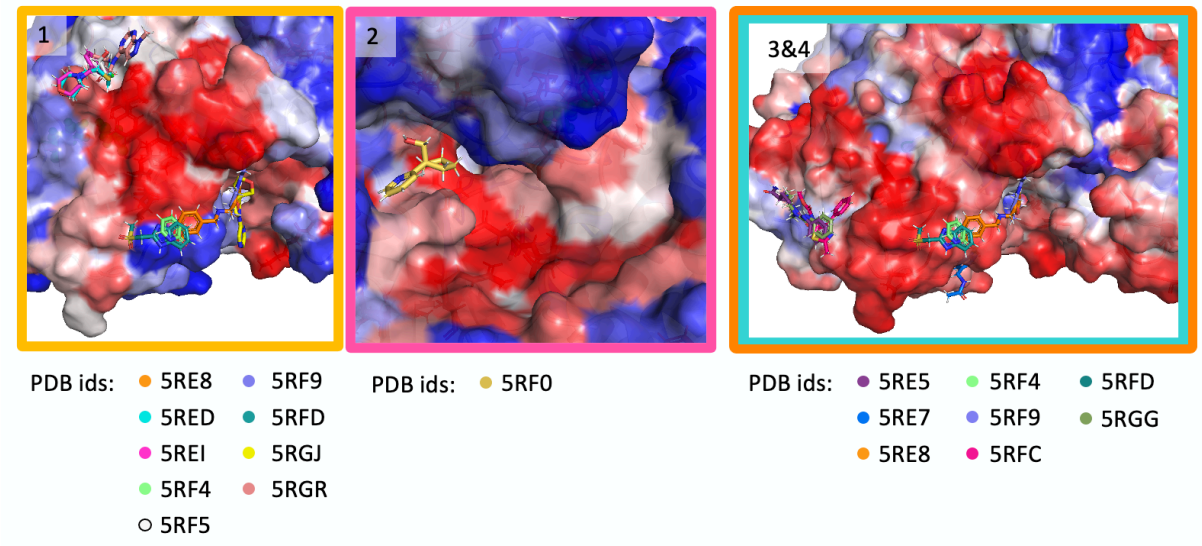


Figure 5.7: Fragments in proximity to identified allosteric hotspots. Close-up of the 4 hotspots with the colour code as in the figures above: yellow - hotspot 1, pink - hotspot 2, cyan - hotspot 3, orange - hotspot 4. Compounds from the X-ray crystallographic fragment screen[292] within 4 Åof the hotspots are shown in different colours and their PDB ids are indicated.

Table 5.4 presents the scoring of the active site in BBP and MT analyses sourced from each small fragment as indicated with their respective PDB id. We also indicate which fragment is in proximity to which hotspot and provide a random site score for each scoring to allow a significance assessment. The active site scores significantly higher than a random site would for several fragments. We identified the fragment with the PDB id 5RE8 to be of particular interest as it shows a high connectivity to the active site in both BBP and MT analyses. This fragment is bound in proximity to hotspots 1 and 4 and indicates that this extended region has an allosteric link to the active site.

El-baba et al.[291] took the same X-ray crystallographic fragment screen data as a basis for mass spectrometry experiments with distal binding fragments, and their results provide further support for our analysis. One result they reported is for fragment 5RGJ, which they found to slow substrate turnover rates of M^{pro}. This fragment binds close to hotspot 1 (yellow in Fig. 5.7) and is found to have a significant connectivity to the active site in both BBP and MT analysis (Tbl. 5.4).

Table 5.4: Scoring of the active site when BBP and MT analyses are sourced from small fragments bound to M^{pro}. Small fragments identified to bind close (<4 Å) to identified hotspots are listed with their PDB id[292]. Colours for hotspots are indicated as used in the plots above. Highlighted in orange are fragments mentioned in the text. Active site score $\overline{p_{R,\text{active site}}}$ as defined in Eq. 3.11, random site score $\langle\overline{p_{R,surr}}\rangle_E$ as in Eq. 3.12.

Source	BBP $\overline{p_{R,\text{active site}}}$	BBP $\langle\overline{p_{R,surr}}\rangle_E$ [95 % CI]	MT $\overline{p_{R,\text{active site}}}$	MT $\langle\overline{p_{R,surr}}\rangle_E$ [95 % CI]
5RED (1)	0.63	0.49 [0.48, 0.49]	0.65	0.54 [0.54, 0.55]
5REI (1)	0.73	0.48 [0.47, 0.49]	0.57	0.56 [0.55, 0.56]
5RGJ (1)	0.65	0.46 [0.46, 0.47]	0.65	0.53 [0.53, 0.54]
5RGR (1)	0.67	0.49 [0.48, 0.50]	0.67	0.54 [0.54, 0.55]
5RF5 (1)	0.65	0.48 [0.47, 0.49]	0.62	0.54 [0.53, 0.54]
5RF0 (2)	0.44	0.52 [0.51, 0.53]	0.56	0.55 [0.54, 0.56]
5RGQ (2)	0.35	0.49 [0.48, 0.50]	0.52	0.52 [0.51, 0.52]
5RE5 (4)	0.40	0.49 [0.48, 0.49]	0.51	0.53 [0.53, 0.54]
5RE7 (4)	0.53	0.45 [0.44, 0.46]	0.58	0.50 [0.49, 0.51]
5RFC (4)	0.42	0.48 [0.47, 0.48]	0.49	0.52 [0.52, 0.53]
5RGG (4)	0.24	0.51 [0.50, 0.52]	0.41	0.53 [0.53, 0.54]
5RF9 (1) (3)	0.69	0.50 [0.49, 0.51]	0.69	0.54 [0.53, 0.54]
5RE8 (1) (4)	0.70	0.47 [0.46, 0.48]	0.74	0.50 [0.49, 0.51]
5RF4 (1) (4)	0.64	0.45 [0.45, 0.46]	0.67	0.51 [0.50, 0.51]
5RFD (1) (4)	0.64	0.47 [0.46, 0.47]	0.58	0.52 [0.52, 0.53]

Furthermore, El-baba et al.[291] found that one of the dimer interface binding fragments (PDB id: 5RFA[292]) destabilises dimerisation and can act as an inhibitor of SARS-CoV-2 M^{pro}. Fragment 5RFA is located at 5.8 Å from hotspot 2 and overlaps spatially with fragment 5RGQ which is within our proximity cutoff of <4 Å to hotspot 2. This dimer interface binding pose has been confirmed by another recent fragment screen study[293]. 5RGQ as well as the other interface binding fragment 5RF0 show no direct connectivity to the active site in neither BBP nor MT analyses (Tbl. 5.4). These results are in line with our observations presented above, that the dimer interface residues show no immediate link to the active site but might impact protease activity over the N-finger residues. Hence, we propose that targeting the dimer interface hotspot starting from fragments like 5RGQ and 5RF0 would be a fruitful approach for disrupting dimerisation and M^{pro} activity.

In a similar approach to the X-ray crystallographic fragment screen described above, Günther et al.[294] used two repurposing drug libraries to perform a crystallographic screen of the

M^{pro}. Using a library of chemically more complex molecules, this study goes a step further towards predicting drugability than a fragment library. They describe two allosteric sites and subsequently investigated the antiviral activity of successfully binding agents in cell-based assays. Again, we can find overlap between our results and what was reported in the study by Günther et al.[294]. Their allosteric site 2 is located between the catalytic domain II and the dimerisation domain III and contains the following residues: the loop 107-110, residues ASN^{151}, ASP^{153}, TYR^{154}, VAL^{202}, ILE^{249}, THR^{292}, PHE^{294} and ARG^{298}. With regard to our results, this allosteric site 2 overlaps with our hotspot 3 at residues ASN^{151}, ASP^{153} and TYR^{154}.

Taken together, these studies provide preliminary validation of our findings and strengthen the confidence in our predictions. The combination of small fragment data and our allosteric hotspot locations can be used as a starting point for designing allosteric modulators against the SARS-CoV-2 M^{pro}. Although our approach allows us to predict allosteric perturbations, we cannot say whether a binding at our hotspots will lead to up or down regulation. However, the experimental results discussed above[291,294] suggest that binding in proximity to hotspots 1, 2 and 3 would lead to a decreased viral activity.

5.5 Conclusions

We presented the application of our graph analytical methods onto a topical study system in the virus responsible for the global COVID-19 pandemic. By studying the SARS-CoV-2 M^{pro} with BBP and MT analyses, we provided insights into the molecular mechanism underlying its activation dynamics. We further describe two approaches for the targeted inhibition of the protein independent of active site binding.

We validated activation mechanisms proposed in the previously studied SARS-CoV M^{pro} and show that the same concepts apply to the new isoform in SARS-CoV-2. Mutations between the two isoforms informed our approach and we detected a related strengthening in dimer interface connectivities from SARS-CoV to SARS-CoV-2 (Fig. 5.4E). The general dynamics of M^{pro} which involve mandatory dimerisation and signalling over the N-finger residues towards

the binding pocket, were detected for both SARS-CoV and SARS-CoV-2 (Tbl. 5.1). These results strengthened the notion that the dimerisation process represents a viable approach when trying to disrupt M^{pro} to inhibit viral replication. This could be achieved over designing peptide binders against the dimer interface as recently proposed by ElSawy et al.[310] in a computational study of the SARS-CoV M^{pro}.

The importance of the dimer interface as a targeting approach was further strengthened when we detected an allosteric hotspot (hotspot 2, Fig. 5.5B) in that location. Using BBP and MT analyses in a complementary approach, we detected a total of four allosteric hotspots that are highly connected to the active site. Especially for hotspot 1 (Fig. 5.5A) and 3 (Fig. 5.6C), we found a bi-directional connectivity towards the active site. Notably, we found indications in experimental work based on crystallographic screens that binding in proximity to these hotspots might lower proteolytic activity and viral replication[291,294].

We hope that building on these results will lead to the development of small compounds which can allosterically regulate the main protease of SARS-CoV-2. This will have major implications for the development of a drug against COVID-19. Targeting these sites might be transferable to other coronaviruses and provide an even larger therapeutic potential. Ultimately, the suitability of our allosteric hotspots as drug binding sites and whether a binding event achieves allosteric modulation needs to be confirmed in experimental studies. Given the acute threat by COVID-19, the research field is progressing rapidly, and we can hope for further studies in the direction of allosteric modulation in the near future.

This Chapter ties in with what we presented in Chapter 4 in the approach to investigate dimer interfaces. Again, we were able to demonstrate that not the whole interface but rather particularly high scoring residues are involved in the molecular mechanism of the dimeric SARS-CoV-2 M^{pro}. We also introduced a more classic approach of applying our methodologies by predicting four allosteric hotspots that hold potential for the modulation of M^{pro} activity. In the next step, we expand these concepts onto a lesser studied system in Chapter 6 and demonstrate the whole range of insights that we can gain from atomistic graph analyses.

Chapter 6

Cyclin-dependent kinases 4 and 6

This Chapter builds on the work presented in Chapters 4 and 5 in that we study another dimeric protein complex. However, we went one step further and applied our atomistic graph analysis approach to a heterodimeric protein system. Cyclin dependent kinases (CDKs) 4 and 6 are essential regulators of the cell cycle and function together with D-type cyclins. This Chapter shows how our methodologies can be applied in a less-studied system to deliver predictive results. We investigated the dimer interface and highlight differential signalling clusters that Markov transient (MT) and bond-to-bond propensity (BBP) analyses detect. Furthermore, we explored the effect of different chemotherapeutics and highlight how our approach can aid in understanding differential inhibitor behaviour.

6.1 The cell cycle regulators

Nuclear hormone receptors which we encountered in Chapter 4, are one big part of the signalling pathways in our cells. However, they are mainly interacting with our DNA and initiating gene expression. There is no way around another major protein family on the level of controlling proteins: protein kinases. Their main function is that of a switch where they activate or inhibit their substrates by phosphorylating them. On the genomic level, we talk about the so-called 'kinome' which encompasses 518 putative protein kinase genes[311]. Given the ubiquitous nature

of kinase regulation, it follows that these proteins are implicated in a wide range of diseases, and in cancer in particular[312]. This, together with the fact that kinases constitute up to 22 % of the druggable genome[313], make them a major area of research in drug discovery[314].

Kinases are often classified by their phosphorylation mechanism, and one such class comprises the serine/threonine kinases, named after the residues that they target on their substrates. Amongst those, we find the CDKs. This family consists of 20 members, which are characterised by their need to bind to a cyclin protein for activity[315]. CDKs are commonly divided into two groups: the ones that fulfil a function in transcription regulation, and cell cycle CDKs. Figure 6.1 shows the cell cycle and highlights the CDKs involved in regulating the different phases. CDKs 1, 2, 4 and 6 directly regulate the cell cycle, and CDK7 is indirectly involved by acting as a CDK-activating kinase (CAK) for them[6]. The essential cyclin protein partners are different for each CDK and play an important role in substrate selectivity, which ties into how the cell cycle is regulated[316]. While CDK2 and 1 are widely studied proteins, and their activation mechanism is well understood, CDK4/6 have been much less in the centre of attention, and hence this work focusses on elucidating the mechanisms that underly their activity and inhibition.

6.1.1 CDK4/6 - drivers of the G1 phase

Oscillating activities of CDKs drive the cell cycle. In the first growing phase (G1 phase), the entry into a new cell cycle is stimulated by mitogenic signals that initiate the gene expression of CDK4/6 and D-type cyclins. Once they reach certain concentrations and become active, the G1-phase progresses, and they trigger the activity of CDK2 until the G1/S checkpoint is reached[6]. Although CDK4/6 seem to be functional homologs in the cell cycle and can compensate for each other[317], we can assume that they have diverging roles in different contexts. Indeed it has been found that CDK6 fulfils a role in differentiation in different tissues[318].

To this day, it is not fully understood how CDK4/6 are activated, and Figure 6.2 attempts to provide an overview of what is known to date. What is clear is that a multi-layered input is at play which reflects the needs of a tightly regulated cell cycle progression under the

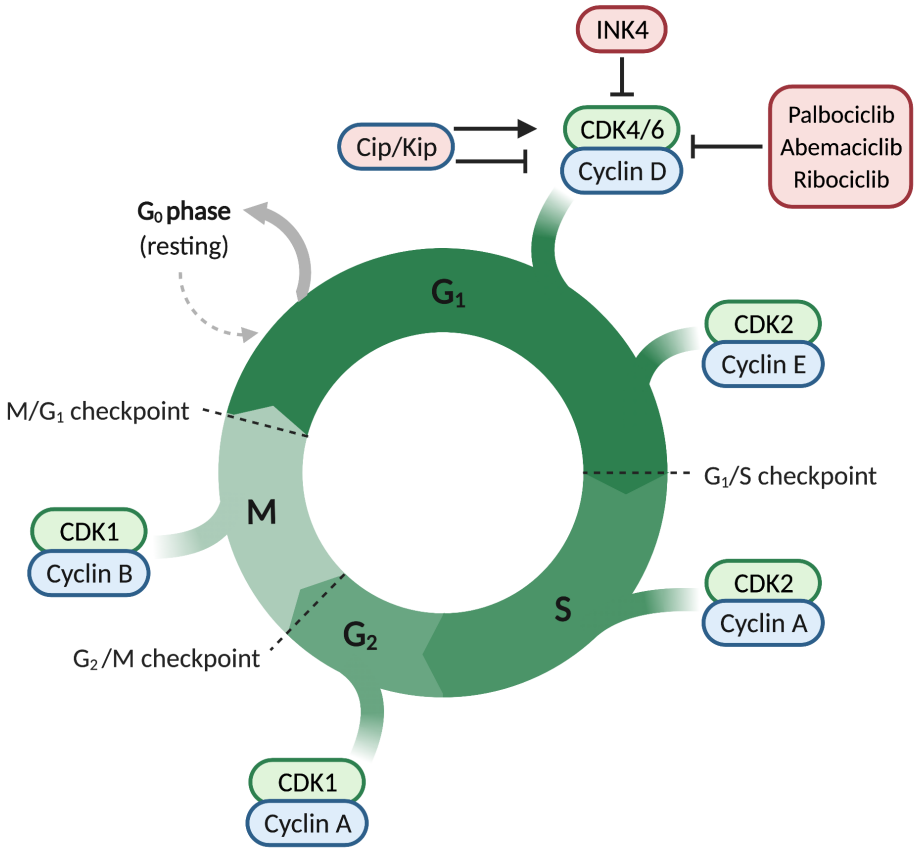


Figure 6.1: The human cell cycle. Shown are the different phases of the cell cycle. In the G1 stage the cell increases in size, in the S stage the DNA is copied, in the G2 stage the cell prepares for division which happens in the M stage. The respective CDKs and corresponding cyclins are also indicated. In the case of CDK4/6, inhibitors of CDK4 (INK4) and the cancer therapeutics palbociclib, abemaciclib and ribociclib act as inhibitors. CDK interacting protein (Cip)/kinase inhibitory protein (Kip) proteins can have both an activating and an inhibiting effect on CDK4/6. Adapted from "Cell Cycle Deregulation in Cancer", by BioRender.com (2022)[†].

influence of a variety of internal and external growth factors[315]. In CDK2, we see a much broader field of research, and the activation cycle has been studied extensively. To achieve full activity, CDK2 requires cyclin binding and phosphorylation of a threonine in the activation loop (T160) by CDK-activating kinases (CAKs)[6]. Although it was originally proposed that cyclin binding would be followed by phosphorylation[319], recent literature seems to argue on the side of a phosphorylation event, followed by cyclin binding[320–322]. However, a semi-active state of the CDK2 - cyclin A/E complex is possible where cyclin binding triggers structural rearrangements[19].

[†]Retrieved from app.biorender.com/biorender-templates

For CDK2 these steps are under inhibition of CDK interacting proteins (Cips) and kinase inhibitory proteins (Kips)[323,324].

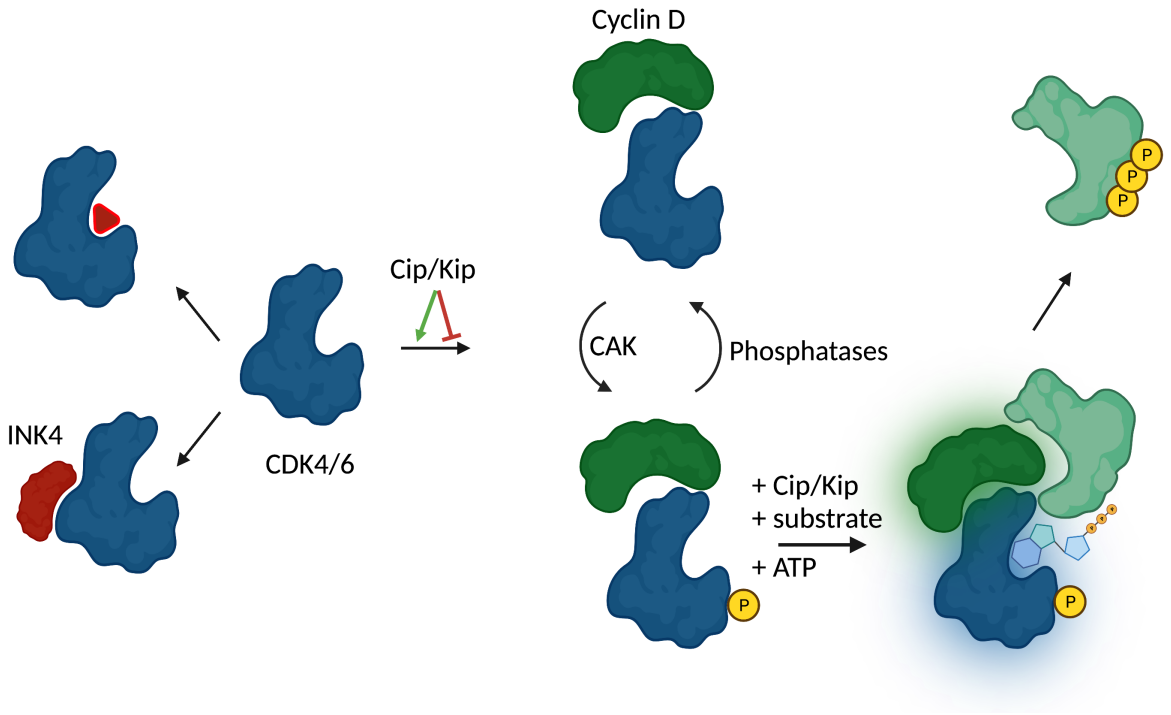


Figure 6.2: Activation pathway of CDK4/6. The CDK4/6 activation pathway is depicted with CDKs in dark blue and cyclins in dark green. Monomeric CDK4/6 can be inhibited by small-molecule inhibitors or protein partners like INK4 (depicted in red). Several signals are required for activation of CDK4/6. In a first step, D-type cyclins are recruited, a process which is facilitated or interrupted by Cip/Kip interactions. CDK-activating kinases (CAK) lead to phosphorylation on CDK4/6, a flexible process in the inactive state of cyclin bound CDK4/6. After the recruitment of substrate proteins (light green) which is again facilitated by Cip/Kip proteins, the complex becomes active (halo around proteins), ATP (shown as simplified chemical structure) is recruited, and the substrate is phosphorylated. Phosphate depicted in yellow[‡].

Contrastingly, these same inhibitory proteins (Cip/Kip) fulfil a stabilising and activating role for CDK4/6 and D-type cyclins under certain conditions[324]. The inhibitor role for CDK4/6 is fulfilled by inhibitors of CDK4 (INK4) proteins, which prevent cyclin binding (reviewed in Sherr and Roberts[323] and Pavletich[319]). It is proposed that this allosteric inhibition is counteracted by the Cip/Kip proteins p21 and p27. Although p21 and p27 can act as inhibitors for CDK4/6[325,326], they are also required to assemble an active CDK4/6 - cyclin D complex[325,327].

Another difference that becomes quite apparent when comparing the structures of CDK2 and

[‡]Created with biorender.com

CDK4 bound to their cyclin partners is the much smaller dimer interface[19], as discussed in Section 6.1.2 below. In CDK4, the interaction with D-type cyclins occurs at a smaller interface with the cyclin in an 'elevated' orientation[328,329]. The position of the cyclin leaves the complex in an inactive state, with the activation loop in a more flexible conformation. It is proposed that this flexibility is needed to facilitate a fine tuning of the G1 progression which is achieved by continuous phosphorylation by the CAK CDK7[329]. Indeed it has been shown that CDK7 needs to be continuously active to keep CDK4/6 phosphorylated[330] (Figure 6.2 middle).

To push the complex to a fully active state, the help of Cip/Kip proteins p21 and p27 is required again. It has also been proposed that the binding of substrate proteins of the complex is required for final structural rearrangements that allow activity[329]. Whether there is a sequential order to the above-described steps or whether it is more of a dynamic coming together of the various input signals and binding partners in the CDK4/6 - cyclin D1 complex remains elusive.

After full activation is achieved, CDK4/6 phosphorylate a comparatively small substrate set of transcription factors and the main substrates retinoblastoma proteins (pRBs)[331]. pRBs are essential regulators of the cell cycle and important tumour suppressor proteins[332]. Hence, a CDK4/6 induced dysregulation of their function has implications for almost all cancer types[333].

CDK4/6 as targets in breast cancer

In general, dysregulations of the cell cycle are a hallmark of cancer growth and tumour progression[6,334]. To a large part, this dysregulation is connected to a malfunctioning of cell cycle kinases[335], which makes them an attractive drug target[336]. For this work, we want to focus on the effects that CDK4 and 6 have on cancer development and progression, and on the need for selective inhibitors against them. The general role of CDK4/6 in the cancer context is the inactivation of the tumour suppressor pRB through phosphorylation. pRB then in turn no longer inhibits the function of E2F transcription factors that stimulate gene expression of a wide range of target genes, amongst other CDK2 and cyclin E, which are required for cell cycle progression[337]. D-type cyclins and CDK4/6 have also been shown to have further cell cycle-independent roles in cancer cells (reviewed in Gao et al.[53]), strengthening the need for

effective inhibitors.

Connecting to the work we presented in Chapter 4, the aforementioned cyclin D - CDK4/6 - pRB pathway has been shown to be abnormally activated in breast cancer (BC). Cyclin D1, cyclin D3 and CDK4 amplifications and mutations are commonly observed in BC subtypes[338] and are proposed to further play a role in conferring resistance against anti-estrogen therapies[339]. Moreover, it is known that CDK4/6 and cyclin D1 are involved in BC metastasis either as single agents[340] or as CDK-cyclin complex[341]. Taken together, this makes targeting CDK4/6 and D-type cyclins a prominent therapeutic route in BC patients[339]. Three widely studied examples that have been approved for hormone receptor positive/HER2-negative BC treatment are abemaciclib, palbociclib and ribociclib. These small compounds bind competitively to the ATP binding site and exclusively inhibit CDK4/6 but not other CDKs[342]. They are commonly prescribed as combination therapy with tamoxifen or fulvestrant in hormone receptor-positive BC patients (reviewed in Gao et al.[53] and Susanti and Tjahjono[343]). However, as seen for other chemotherapeutic agents, BC cells have shown intrinsic and acquired resistance mechanisms against CDK4/6 inhibitors[344]. In line with our work in Chapter 4, we aim to understand the inhibitor mechanisms and how they interrupt CDK activation on an atomistic level. In doing so, we provide scope for alternative targeting mechanisms once current options in recurrent tumours are exhausted.

6.1.2 Structural features of CDK4/6

Like most protein kinases, CDKs follow the highly conserved 'kinase fold', which is made up of two principal regions: the N- and C-lobe[345]. Figure 6.3 highlights the important structural elements in CDK2, 4 and 6 that have been extensively reviewed by Wood and Endicott[19]. The binding site, located between the N- and C-lobe, is where adenosine triphosphate (ATP) binds, which is supported by magnesium ions. The binding site (light orange in Fig. 6.3) consists of the glycine-rich loop (G-loop), which contributes to the binding of the phosphate moieties. The hinge region bridges between the N- and C-lobe and binds to the adenosine binding site. The DFG motif (dark orange in Fig. 6.3) is also involved in binding the phosphates and sits at

the start of the activation loop. For a detailed rationale behind choosing binding site residues used for this study, see Appendix A.1.3. The activation loop (red in Fig. 6.3) is highly flexible

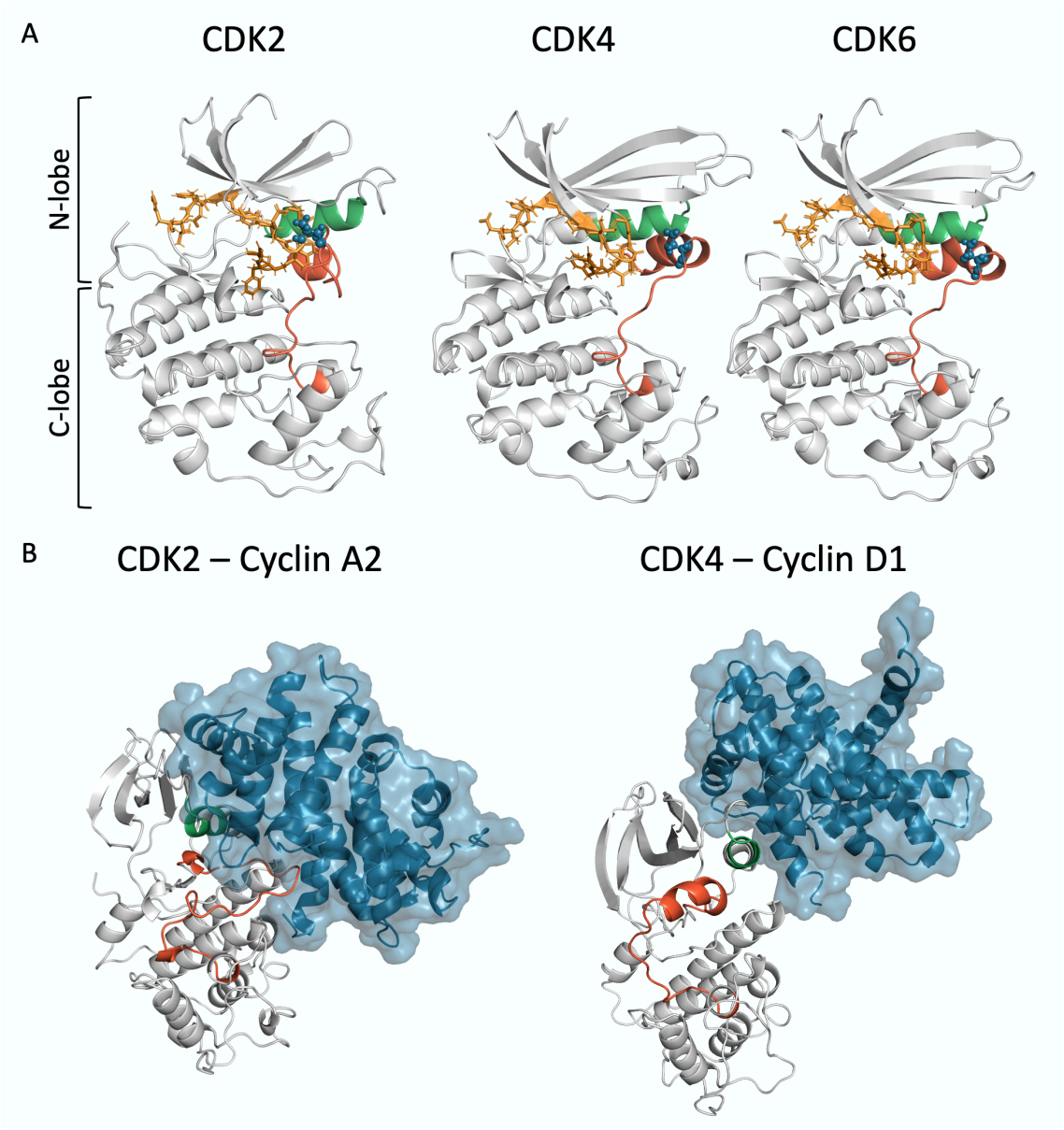


Figure 6.3: Structural features in CDKs and complexes. A) Monomeric structures for CDK2 (PDB id: 1HCL[346]) and CDK4 and 6 (both modelled with AlphaFold[49]). Highlighted in colour are the following elements: light orange - binding site residues (excluding DFG motif); dark orange - DFG motif; red - activation loop; blue - phosphorylation site; green - Cα helix. For a complete list of the residues constituting these structural features, see Table C.7. **B)** CDK2 in complex with cyclin A2 (PDB id: 1FIN[347]) and CDK4 in complex with cyclin D1 (PDB id: 2W9Z[328]). The cyclin binding partner is shown in blue, the activation loop and Cα helix are coloured as above.

and located between the highly conserved DFG and APE motifs. This loop blocks the ATP binding site in an inactive (DFG-'out') conformation until binding partners stabilise it in a

DFG-'in' conformation where it serves as an assembly platform for substrate peptides. This loop also contains the phosphorylation site (blue in Fig. 6.3), which is a threonine at the position 160 in CDK2, 172 in CDK4 and 177 in CDK6. The Cα helix (green in Fig. 6.3) is part of the interaction site where cyclin binds. It also contains the structurally important PSTAIRE motif[§], which relocates into the binding site upon activation.

These elements are rearranged upon binding of cyclin proteins to CDK2 but not to CDK4/6. This differential mode of action can be partially explained by the different binding modes that lead to a larger binding interface in CDK2 than in CDK4[316] where the cyclin is in an 'elevated' position (Fig. 6.3B). This latter binding mode is also proposed in CDK6. Unfortunately, there is no structure available for CDK6 in complex with a human D-type cyclin as of today. However, structures of CDK6 with a viral cyclin are available (Fig. A.1) and show the complex in an active conformation similar to that of the CDK2 - cyclin A/E complex[348]. This complex does not contradict the assumption that CDK6 - cyclin D would be in an inactive conformation, as the viral mechanism overcomes the normal cell cycle regulation in the G1 phase and is resistant to inhibition[348].

The cyclin parts in the complexes belong to a 30-member family of proteins that share little sequence homology (reviewed in Tatum and Endicott[316]). Instead, they are structurally defined by the cyclin-box motif, which can be present in either one or more copies in the cyclin. In the cell cycle, CDK4/6 bind exclusively to D-type cyclins, which contain one cyclin-box motif[349]. Another relevant structural feature is the RXL binding site (highlighted in Fig. 6.5 and Tbl. 6.1) which forms a hydrophobic pocket where substrates and co-factors can assemble. It was first described in 1996 by Russo et al.[350] and structurally confirmed in CDKs for co-factors like Cip/Kip proteins[327,350] and substrates[351]. This protein-protein interaction (PPI) site is essential for the activity of the complex, as is the RXL motif on the substrate side[352].

From a structural perspective, CDK4/6 are much less studied than CDK2. This is also mirrored in how many structures are available in the protein data bank (PDB): 426 structures for CDK2 (UniProt id: P24941), whereas only 13 structures for CDK4 (UniProt id: P11802) and 18

[§]PSTAIRE in CDK2, PISTVRE in CDK4 and PLSTIRE in CDK6

structures for CDK6 (UniProt id: Q00534) are available.[¶] An overview of available structures and which ones were used in this project is given in Appendix A.1.3 and Figure A.1. To make basic comparisons between CDK2 and CDK4/6, we chose to use AlphaFold, a deep learning structure prediction tool[49]. In a recent effort in collaboration with the EBI, a publicly accessible database for modelled protein structures was established. We used this resource to obtain the monomeric structures of CDK4[‖] and CDK6[**]. More details can be found in Appendix A.1.3 and Figure A.2.

Objective

In this Chapter we demonstrate the application of Markov transients in conjunction with bond-to-bond propensities to explore the connectivities and dimer interactions within CDK - cyclin complexes. The overarching aim of this Chapter is to provide detailed insights into the activation mechanism of CDK4/6 and D-type cyclins as well as investigate the inhibition patterns of chemotherapeutics.

First, we contrasted the connectivities in monomeric CDK4 with CDK2 to identify general differences between the two kinases. Next, we investigated the heterodimers that are formed between CDK4 and cyclin D1 and D3, and described how the activation of CDK4/6 might be achieved in a stepwise manner. We also took this opportunity to identify extended hotspots (similar to what we presented in Chapter 5) on the dimer that might hint towards PPI sites as assembly points for substrates and co-factors. Finally, we extended our analysis onto structures that contain approved cancer therapeutic molecules. We aimed to shed light onto the inhibitory mechanisms in CDK6 and provide first insights into the atomistic signalling differences between three approved chemotherapeutics. We extended this analysis on an inhibited CDK2 structure to compare the mechanisms. This ties in with Chapter 4, where we demonstrate that our methodologies allow us to explore the mechanisms that underly inhibitory molecules.

[¶] As of 28.11.2021
[‖] Entry accessible at: alphafold.ebi.ac.uk/entry/P11802
[**] Entry accessible at: alphafold.ebi.ac.uk/entry/Q00534

6.2 Differences in cyclin binding site are revealed in monomeric CDK2 and 4

In a first step, we were interested to see whether we could detect differences in the connectivities in the monomeric forms of CDK4 and CDK2. To perform this comparison, we had to use a modelled structure for CDK4, which we obtained from AlphaFold[49]. For CDK2, we used the apo monomeric form in inactive conformation at 1.8 Å[346]. One of the activating factors for CDKs is a phosphorylation event in the activation loop. To mimic this signal, we sourced our methodologies from this phosphorylation site: THR^{172} and THR^{160} for CDK4 and CDK2, respectively.

In a first step, we focused on Markov Transients as we gained from the work in Chapters 4 and 5 that this method is particularly powerful in catalytically active proteins, which is the case for CDKs. In general, we see the results from BBP analyses support the same patterns, hence they will only be shortly summarised in the following Sections with full results in the Appendix. Where the results of the two methodologies show diverging patterns, they will be discussed in detail.

We evaluate the MT analysis results across the different structural features that convey activity in CDKs (see Fig. 6.3). Figure 6.4 gives an overview of the results where we highlight different aspects. As detected by MT analysis, the most apparent difference between CDK4 and CDK2 is the signal transmission towards the Cα helix. For CDK4, the average QS of the Cα helix is 0.75 as opposed to 0.58 for CDK2. These results suggest a faster signalling towards the interaction site where the cyclin partners bind. Within the Cα helix, the PISTVRE/PSTAIRE motif is of particular interest as it contributes to the structural rearrangements required for CDK activity[19]. The single residue level (Fig. B.7) shows that this motif is more of a hotspot for CDK4 (average QS = 0.71) than for CDK2 (average QS = 0.51). Figure 6.4B and C allow a direct comparison of the QS results across the structures and highlight the high scoring Cα helix (green circle in the Figure). The scatterplots in Fig. 6.4B and C provide an overview of the data distribution for all residues which shows that the Cα helices are reached faster in

CDK4 than in CDK2 with an average $t_{1/2}$ of 1076.07 in comparison to 2004.93, respectively. This lowered connectivity towards the Cα helix can also be seen when looking at the BBP results (Fig. B.6). Here, we detect a decrease in average QS from 0.57 in CDK4 to 0.37 in CDK2.

On the other hand, we find a common feature in both CDK4 and CDK2 in the fast signal propagation towards the activation loop with average QSs of 0.73 and 0.77, respectively. The activation loop is highlighted in orange on the protein structures in Figures 6.4B and C. Here, we show that our results are in line with the observation that the activation loop is an essential structural element of the catalytic activity of CDKs. Again, this trend is supported by the BBP analysis (see Fig. B.6 and B.7) where the activation loops have an average QS of 0.71 (CDK4) and 0.63 (CDK2).

Taken together, this first investigation of the monomeric from of CDK4 provided us with valuable insights into the signalling within the protein when sourced from the phosphorylation site. The high scoring Cα helix is in line with the activation mechanism of CDK4 that requires cyclin D to bind to the kinase at this site. Moreover, we detect a high signal connectivity towards the activation loop, which is an essential element of catalytic activity. When contrasting these results with CDK2, the main point that stands out is the lowered connectivity towards the Cα helix. This changed communication pattern might be due to a different activation mechanism in CDK2, where a recruitment of cyclin D might be less dependent on a phosphorylation event at this position.

6.3 Signalling and interactions in the CDK4 and D-type cyclin complexes

To further elucidate the molecular mechanism of CDK4 activity, we chose to investigate CDK4 bound to its cyclin partner proteins D1 and D3. This approach also ties into the overall scope of this work to investigate dimeric proteins in disease. We apply our methodologies to study

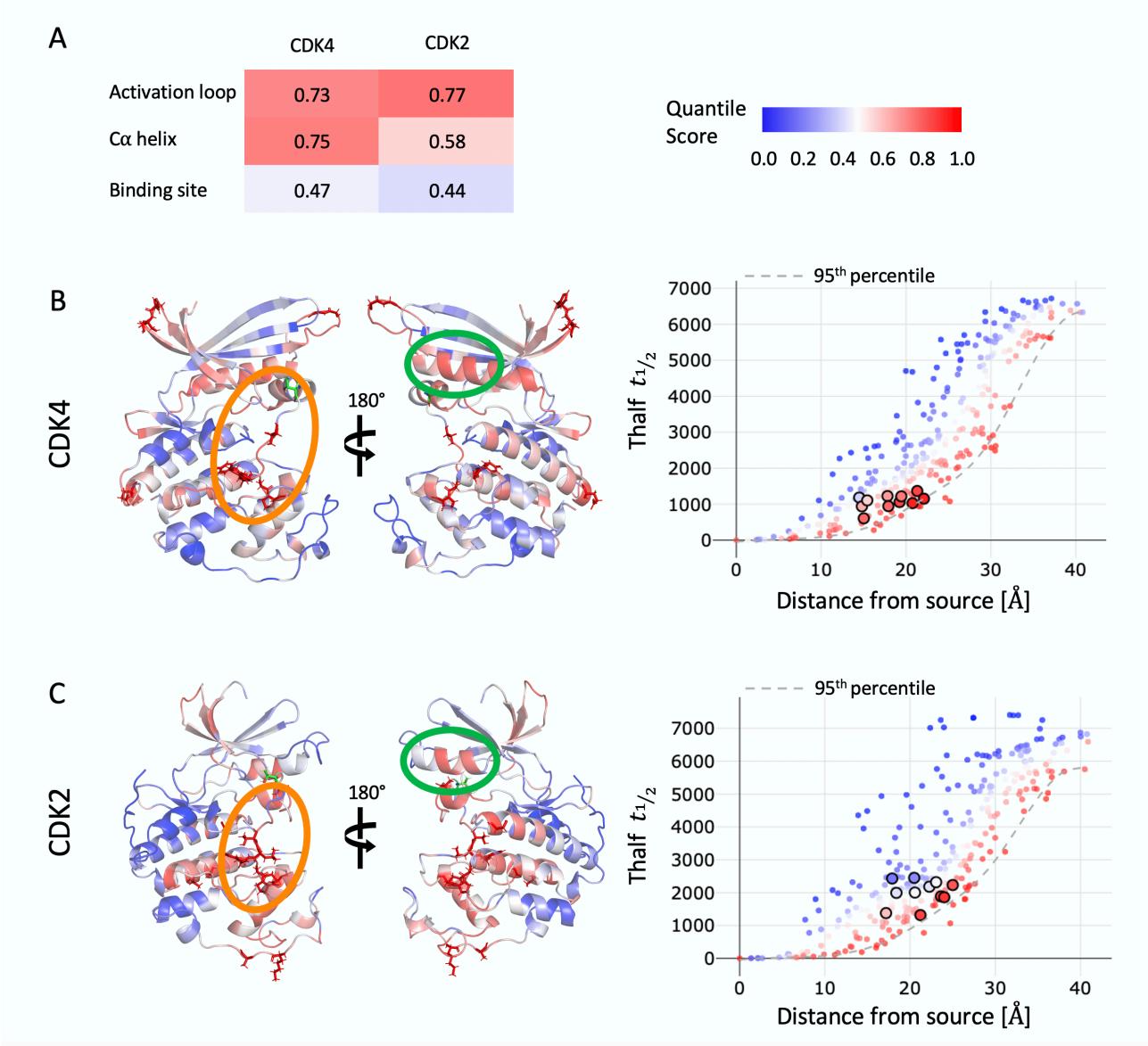


Figure 6.4: Markov transient analysis of monomeric CDK4 and 2 when sourced from the phosphorylation site. A) Average QS results for Markov Transients are shown for each structural element in CDK4 and CDK2. **B) and C)** The structures of CDK4 (AlphaFold model[49]) and CDK2 (PDB id: 1HCL[346]) are shown in two orientations with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a green circle are the Cα helices, and with an orange circle the activation loops. The scatterplots show $t_{1/2}$ values over the distance from the source for each residue in the protein. Cα residues are highlighted as larger dots with a black outline.

a heterodimeric system with CDK4 and the two D-type cyclins (1 and 3). We investigate the different activation signals that are proposed to come together to confer CDK4 - cyclin D activity: the phosphorylation event in the kinase activation loop, binding of ATP at the kinase binding site (Fig. 6.3) and the association with co-factors and substrates which bind at the cyclin RXL site (Fig. 6.5 and Tbl. 6.1). For ease of reading, we present our results on the CDK4

- cyclin D1 complex (PDB id: 2W9Z[328]) in the following Sections. While we found that the CDK4 - cyclin D3 complex (PDB id: 3G33[329]) largely follows the same patterns, we highlight differences where they have been detected. As shown in Figure 6.5, both complexes are solved in the same orientation with the D-type cyclins binding higher than seen in other CDKs as visualised in Figure 6.3B. This cyclin position leaves CDK4 in an inactive conformation, with the activation loop blocking the active site[19].

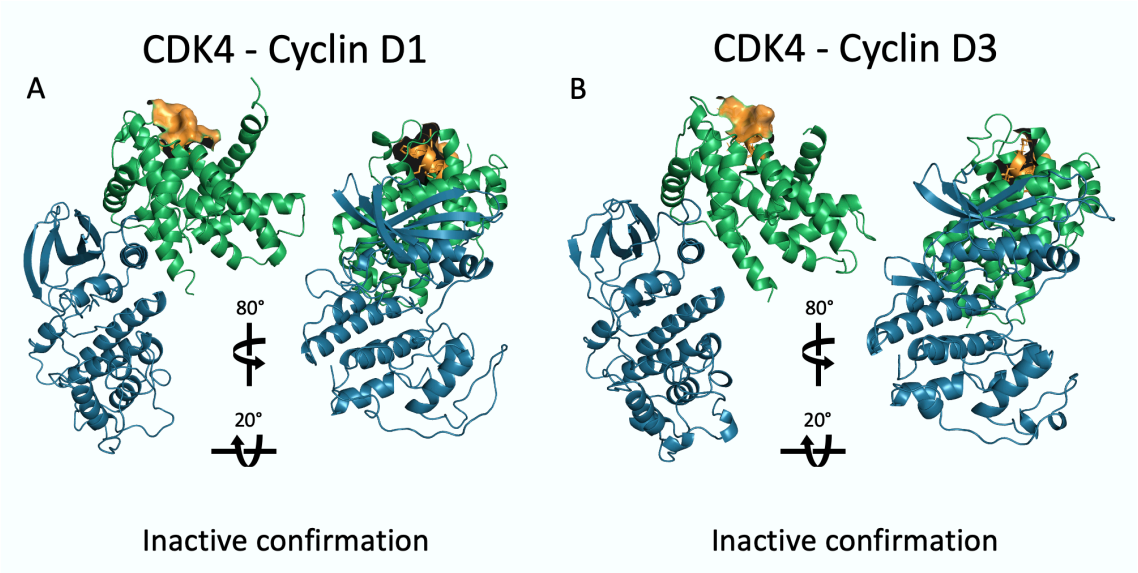


Figure 6.5: CDK4 in complex with two D-type cyclins. A) and B) CDK4 - cyclin D1 (PDB id: 2W9Z[328]) and CDK4 - cyclin D3 (PDB id: 3G33[329]) are shown with CDK4 in blue and the respective cyclins in green. Both complexes are in the inactive conformation, with the activation loop blocking the active site. Highlighted as an orange surface is the RXL site on cyclin D1 and D3, a hydrophobic binding site for substrates and co-factors (residues are listed in Table 6.1).

In a first approach, we mimicked the two activation signals that happen on the kinase side: phosphorylation at position 172 and the binding of ATP. Figure 6.6 shows the results of the MT analysis, which was chosen as the first methodological step here based on the catalytic nature of the protein complex. The MT results allow us to detect areas of the protein which are particularly fast connected to a chosen source. In both scenarios, we detect a fast signal propagation towards the RXL site on cyclin D1 (Fig. 6.6C and D). When we source the signal from the phosphorylation site, we detect the cyclin D1 RXL site with an average QS of 0.83, which is significantly higher than a random site score of 0.48 (95 % confidence interval (CI): [0.46,0.49]). Interestingly, this is one of the only occasions in our study of CDK4 - cyclin D

complexes where we see a difference for the complex bound to cyclin D3. For that complex, the average QS of the RXL site is 0.69 when sourced from the phosphorylation site (Fig. B.8). Although still higher than a random site score of 0.49 (95 % CI: [0.47,0.50]), we think this might hint towards a slightly different prioritisation in input signals between the CDK4 - cyclin D1/3 complexes.

For the run sourced from the ATP binding site residues, we detect a similarly strong connectivity towards the RXL site with an average QS of 0.85 compared to 0.48 (95 % CI: [0.46,0.49]) for a random site score. As further shown in Figure 6.6E and F, residues PRO^{54}, SER^{55}, MET^{56}, LYS^{58} and ILE^{59} in this binding site are particularly high-scoring, each of them scoring an average QS ≥ 0.88. Taken together, we detect a fast connectivity between the kinase activation events and the RXL site on the cyclin partner. These results indicate that a substrate or co-factor binding event at the RXL site is important to achieve catalytic activity of the complex.

Interestingly, we detect a two-step process when investigating the instantaneous communication in the protein with BBP analysis. Figure 6.7 shows the results when mimicking a phosphorylation event and ATP binding. Other than in the MT analysis above, we do not pick up on the cyclin D1 RXL site when sourcing the signal from ALA^{172}. However, we detect a connectivity towards the binding site residues (Fig. 6.7C). A sequence view of the residues that belong to the binding site (Fig. 6.7E) allows us to understand the signal in more detail. Especially high scoring are the hinge residues at position 93 - 97 and the D^{158}-F^{159}-G^{160} motif. When the signal is sourced from these binding site residues in the next step, we detect a high scoring RXL site on cyclin D1, as shown in Figure 6.7D. The RXL site has an average QS of 0.82 in comparison to a random site score of 0.54 (95 % CI: [0.54,0.55]). When looking at the single residue level of the RXL site (Fig. 6.7F), we can see high QSs for ARG^{57} and VAL^{60}, two residues that were not picked up as high scoring in the MT analysis. This shows that Markov Transients and bond-to-bond propensities can be used to highlight different functionally relevant details on the residue level while complementing each other in detecting larger sites of interest.

Taken together, these results indicate a multi-factorial activation process in D-type cyclin controlled CDK4/6. In the monomeric forms of CDK4 and CDK6, we detect a particularly fast

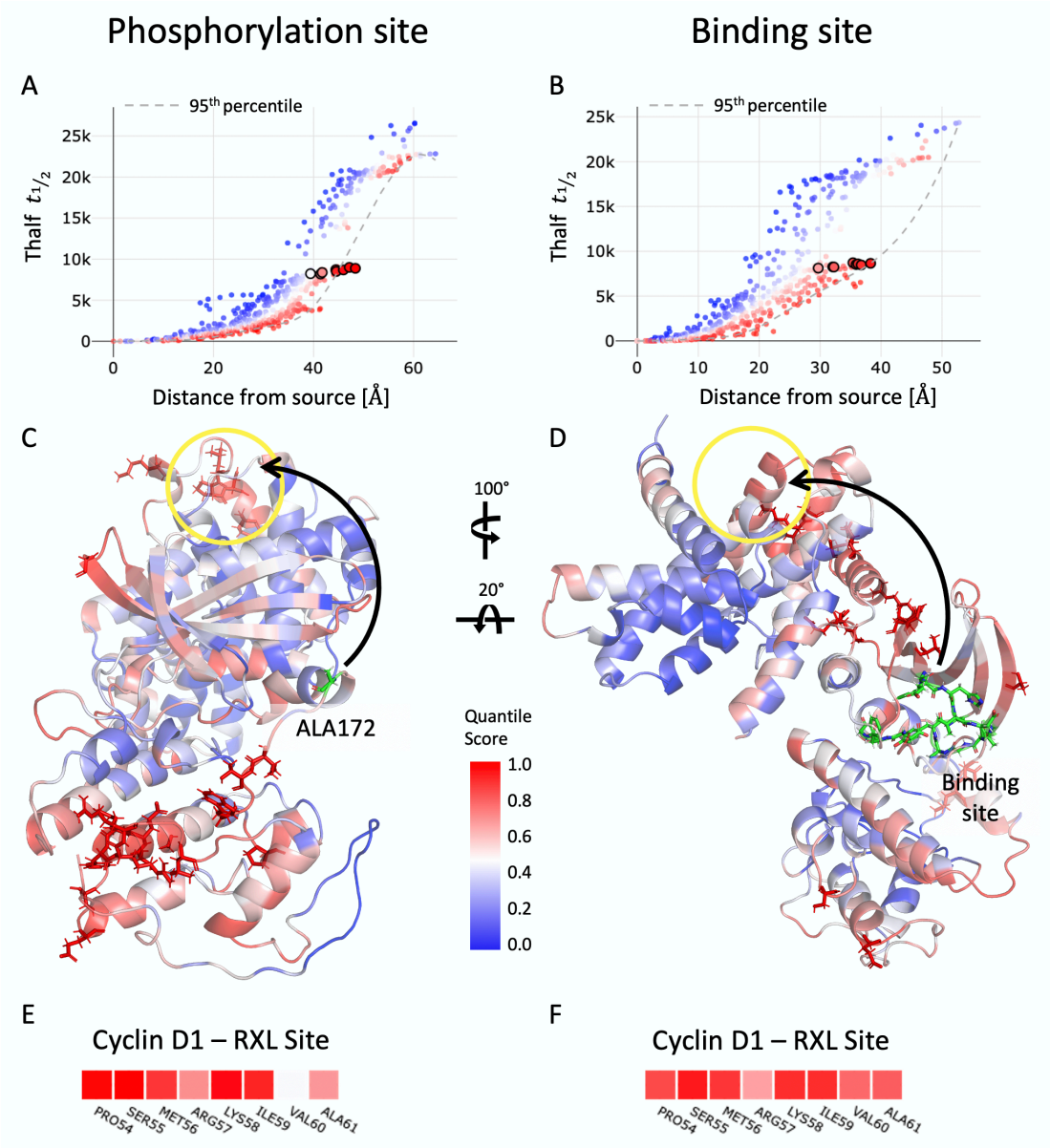


Figure 6.6: Markov transient analysis of CDK4 - cyclin D1. Shown on the left are the results of the analysis sourced from the phosphorylation site ALA^{172} and on the right when sourced from the binding site residues. Colours are according to QS from 0 - blue to 1 - red. **A) and B)** Data distribution of all residues with $t_{1/2}$ values over the distance from the source. RXL site residues are highlighted as larger dots with a black outline. **C) and D)** The complex (PDB id: 2W9Z[328]) is shown in two orientations with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a yellow circle are the RXL sites on cyclin D1. **E) and F)** Detailed sequence for the RXL site residues coloured by QS.

signal propagation towards the Cα helix (Fig. B.10), a binding interface for D-type cyclins. When this cyclin partner is bound, we saw a strong connectivity from the phosphorylation site and the ATP binding site towards the RXL site on the cyclin. As the RXL site is where substrate and co-factor proteins bind, our results indicate that these signals need to come together

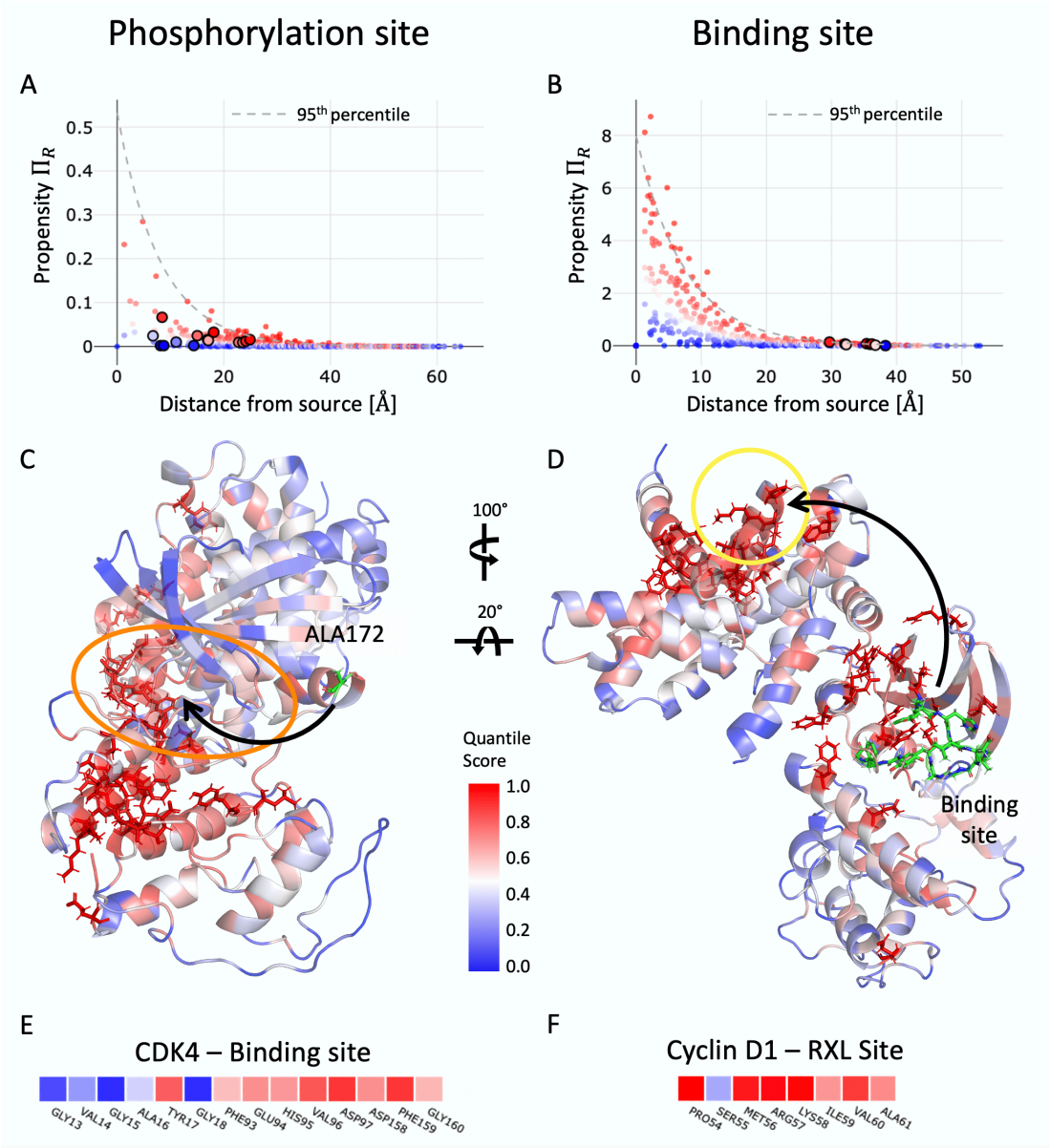


Figure 6.7: Bond-to-bond propensity analysis of CDK4 - cyclin D1. Shown on the left are the results of the analysis sourced from the phosphorylation site ALA^{172} and on the right when sourced from the binding site residues. Colours are according to QS from 0 - blue to 1 - red. **A) + B)** Data distribution of all residues with propensity $\prod_R$ values over the distance from the source. Binding site (A) and RXL site (B) residues are highlighted as larger dots with a black outline. **C) and D)** The complex (PDB id: 2W9Z[328]) is shown in two orientations with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted in orange is the binding site in CDK4 (C) and with a yellow circle the RXL site on cyclin D1 (D). **E)** Detailed sequence of the binding site residues in CDK4 coloured by QS. Especially the hinge region (93-97) and the DFG motif (158-160) are scoring highly. **F)** Detailed sequence for the RXL site residues coloured by QS.

for full activation of the complex. Interestingly, this process seems to be equally fast for the phosphorylation and the ATP binding signal as revealed by Markov Transients. On the other hand, for bond-to-bond propensities, we have a two-step process that suggests a strong connec-

tivity between the phosphorylation site and the binding site and in the next step between the binding site and the RXL site.

6.3.1 Markov Transients reveal protein-protein interaction sites in CDK4 - cyclin D complexes

F. Vianello[163] recently demonstrated in great detail that BBP analysis can be extended from the prediction of allosteric sites onto PPI sites. This stems from the notion that protein-protein interactions are a form of allosteric signalling where the input signal is not given by a small molecule but by a protein binding partner[163]. In a similar line of argumentation, we extend Markov Transients to the study of PPIs.

As demonstrated in Chapter 5, MT analysis is a powerful tool to identify putative allosteric sites in catalytically active proteins. We here demonstrate the use of MT analysis to predict hotspot regions on the CDK4 - cyclin D complex. Again, we use the results that we obtain from runs sourced from the activity triggering input signals, i.e. the phosphorylation site ALA^{172} and the binding site residues. As shown in Figure 6.8 we detect two large hotspot regions on the complex. The first region, which was uncovered in the run sourced from the phosphorylation site, is found on the CDK4 C-lobe as highlighted in cyan in Figure 6.8A.

It is known that the CDK4 - cyclin D complexes require additional protein partners to stabilise an active complex[19,353]. We propose the detected hotspot area might serve as an assembly surface for further protein partners.

In the results we obtained when using the binding site residues as source signal, we detect a large extended hotspot area that spans from CDK4 to cyclin D1 (Fig. 6.8B). Interestingly, this hotspot aligns with the binding site of the Cip protein 21 (p21) and the Kip protein 27 (p27) as solved in the structures by Guiley et al.[327]. Figure 6.8B shows the MT analysis results on the CDK4 - cyclin D1 complex overlaid with p21 in orange (PDB id: 6P8H[327]) and p27 in green (PBD id: 6P8E[327]). Our results align exactly with how the p21 and p27 fragments bridge over CDK4 to cyclin D1 where they bind to the cyclin RXL site with their RXL motives:

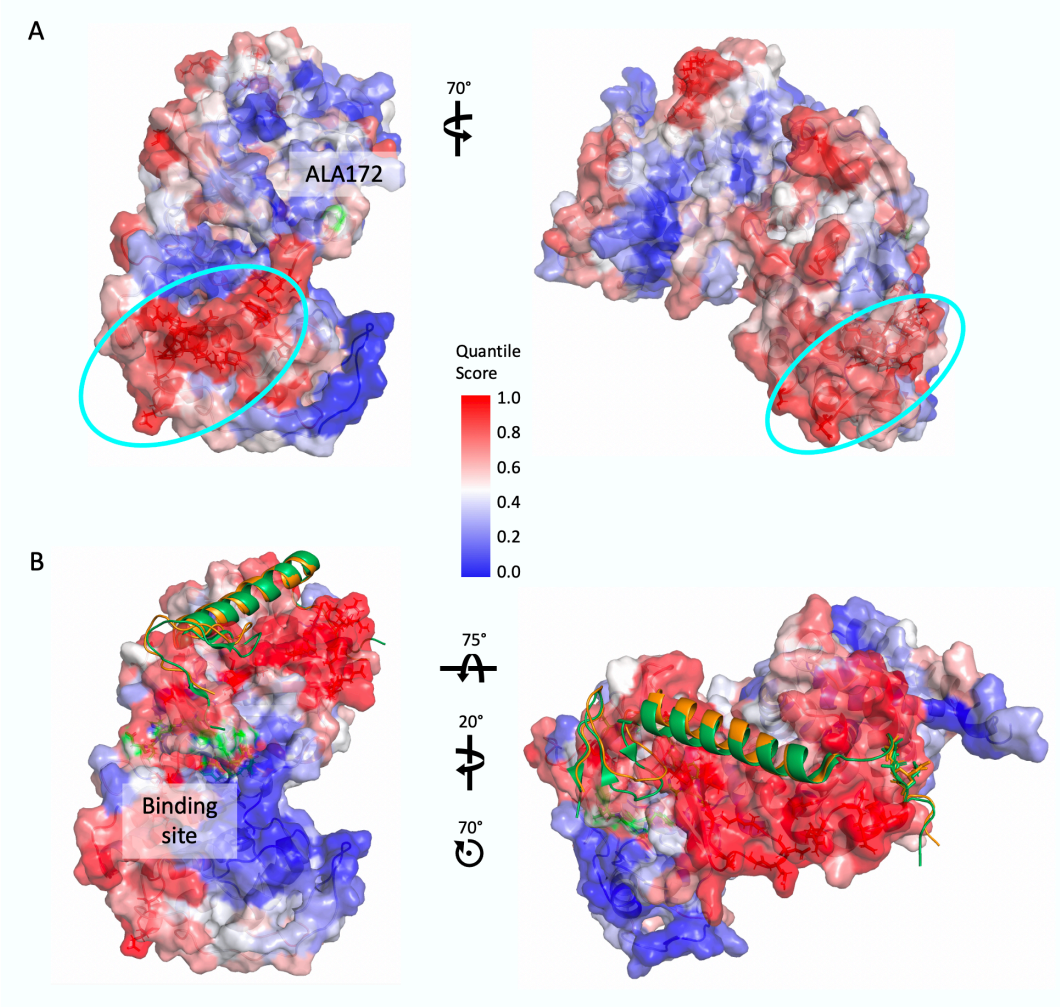


Figure 6.8: Protein-protein interaction sites predicted with Markov transient analysis. A) Kinase-front surface view of the CDK4 - cyclin D1 complex (PDB id: 2W9Z[328]) coloured by QS when sourced from the phosphorylation site ALA172. Highlighted in cyan is an extended hotspot region that we propose to be a PPI site. **B)** Surface view of the same complex coloured by QS when sourced from the binding site residues. Overlaid are the p21 (orange, PDB id: 6P8H[327]) and p27 (green, PDB id: 6P8E[327]) fragments. The RRL/RNL motives are shown as sticks.

R^{19}-R^{20}-L^{21} for p21 and R^{30}-N^{31}-L^{32} for p27. These results further validate that MT analysis is a powerful tool to predict allosteric signalling and hotspots of all sizes and hence can have applications in predicting PPIs.

6.3.2 Bi-directional activity is detected from RXL site

To complement the picture of CDK4 - cyclin D1 activity, we also sourced our methods from the RXL site on the cyclin. This hydrophobic pocket is required for substrate and co-factor

assembly[327,351]. Table 6.1 lists the residues contained in the RXL site for cyclin D1 and D3††.

Table 6.1: RXL site in cyclin D1 and D3.

Cyclin	Residues
D1	PRO54, SER55, MET56, ARG57, LYS58, ILE59, VAL60, ALA61
D3	MET56, ARG57, LYS58, MET59, LEU60, ALA61

MT analysis sourced from the RXL site highlights all structural features discussed above as being part of the activation process, as shown in Figure 6.9. The Cα helix can now be more clearly detected than in the runs sourced from the kinase phosphorylation site and binding site (Fig. 6.8), and it scores highly with an average QS of 0.79. For the binding site, we detect an average QS of 0.72. Interestingly, the signal is now mainly detected in the G-loop residues G^{13}-V^{14}-G^{15}-A^{16}-Y^{17}-G^{18}, as shown in Figure 6.9C. These residues have been identified as binding partners to the phosphate moieties of ATP, and their high QSs might indicate a catalytic signal that is now initiated towards the phospho-groups. The activation loop is scoring moderately high, however the phosphorylation site ALA^{172} is one of the highest scoring residues with a QS of 0.92. We further detect a fast signal propagation towards the area on the C-lobe that we proposed as a potential PPI site in Section 6.3.1. In Figure 6.9B we highlight the hotspot in accordance with what we showed in Figure 6.8A.

Taken together, these results might indicate that once a substrate assembles at the RXL site on the cyclin, signalling steps are initiated within the kinase, which might be supported by further proteins assembling at the highlighted PPI on the C-lobe (Fig. 6.8A and 6.9B). We then also see a high connectivity towards the structural features in the kinase that confer activity and propose that this might be the final step towards complex activity.

††The results in the CDK4 - cyclin D3 complex are in line with the patterns found for the CDK4 - cyclin D1 complex as shown in Figure B.9

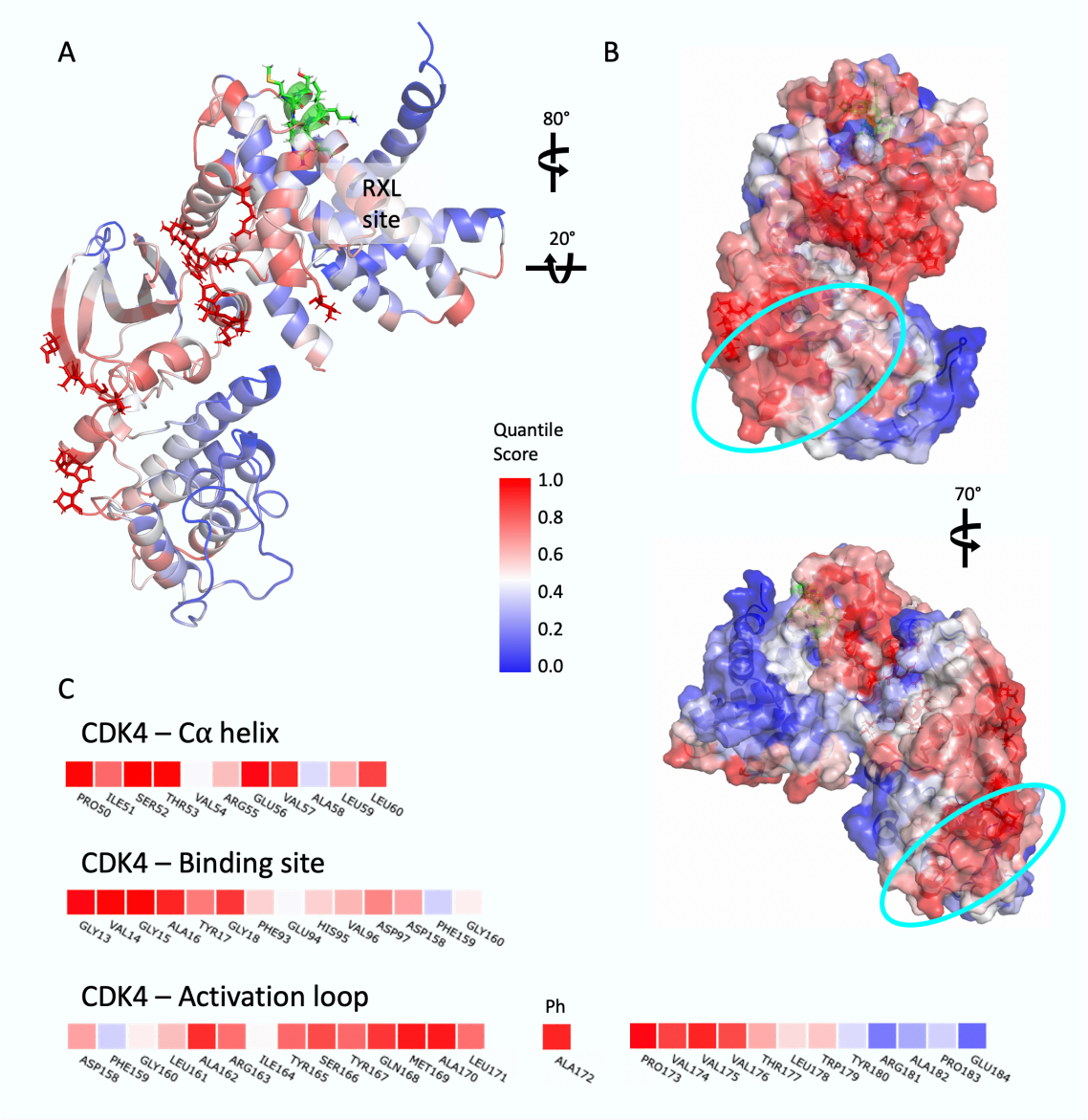


Figure 6.9: The RXL site as a source in MT analysis of CDK4 - cyclin D1. A) The complex (PDB id: 2W9Z[328]) is shown in the front orientations with residues coloured by QS (0 - blue to 1 - red). Residues with a QS > 0.95 are shown as sticks, and source residues are shown as green sticks. **B)** Two surface visualisations of the complex analogous to the results presented in Fig. 6.8. Highlighted in cyan is an extended hotspot region that we propose to be a PPI site. **C)** Detailed sequences for functionally important features on the kinase coloured by QS. Ph - phosphorylation site.

6.3.3 The CDK4 - cyclin D1 interface shows distinct regions for signal transduction

In this work, we are interested to see how dimer interfaces can contribute to activity in protein dimers. We visualise the dimer interface between CDK4 and cyclins D1 and D3 in Figure

6.10. These interfaces were calculated with PDBePisa[78], which distinguishes between interface residues that form bonds (hydrogen bond, disulphide bond or salt bridge) across the interacting protein chains and ones that do not. For the CDK4 - cyclin D1 interface (PDB id: 2W9Z[328]), PDBePisa reports an area of 1138.4 Å^2 for the CDK4 side and 1146.1 Å^2 for the cyclin D1 side. A total of 66 residues are involved in the interaction as listed in Table C.8. For the CDK4 - cyclin D3 interface (PDB id: 3G33[329]) the interface is a little smaller with 63 residues involved (Tbl. C.9), forming an interface area of 1113.4 Å^2 on the CDK4 side and 1085.9 Å^2 on the cyclin D3 side.

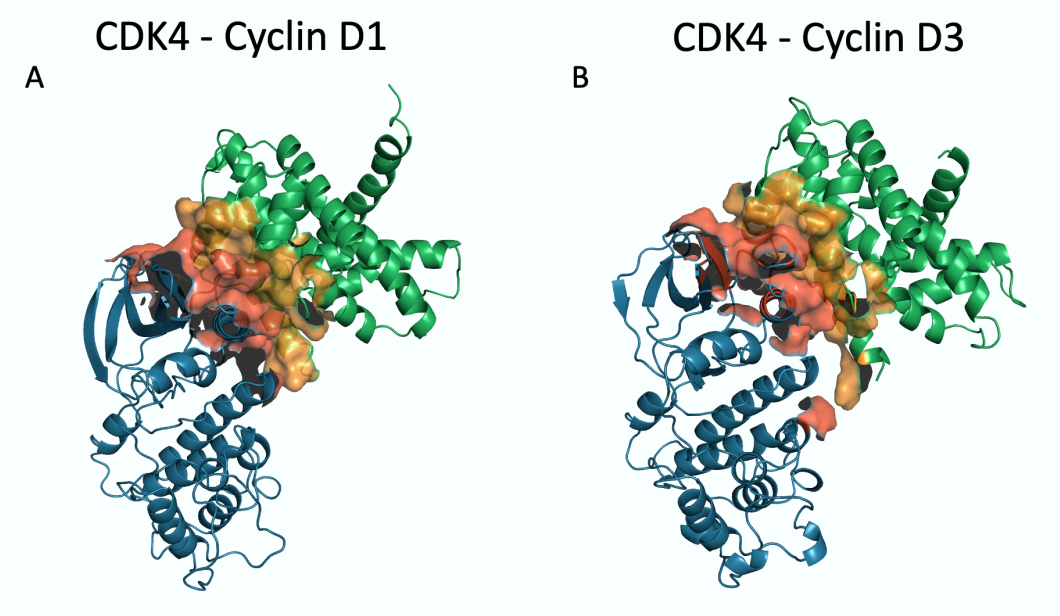


Figure 6.10: The interface between CDK4 and cyclins D1 and D3. A) CDK4 in complex with cyclin D1 (PDB id: 2W9Z[328]). **B)** CDK4 in complex with cyclin D3 (PDB id: 3G33[329]). CDK4 is coloured in blue and the respective cyclin in green. Shown as light red surface is the CDK4 interface contribution and the cyclin D1/D3 one is shown in orange.

We score the interface residues to get an idea of the overall connectivity between the two monomers relative to the source site. We took the same source sites as mentioned in the Sections above to evaluate if and how the different binding events and activation signals are conferred over the dimer interface. We did this analysis for Markov Transients and bond-to-bond propensities to see whether we can distinguish between a catalytically relevant effect and a strong perturbation connectivity. Table 6.2 lists the average QS of the dimerisation site for three different sources: the phosphorylation site on CDK4, the binding site residues on CDK4 and the RXL site on cyclin D1. We find that the average QSs of the interface residues

are higher when the signal is sourced from the binding site or RXL site in comparison to the phosphorylation site. We proposed above that the phosphorylation event might occur at an earlier stage in the activation process of the CDK4 - cyclin D complex, hence at this stage, the dimer interface might not yet contribute to the energetic flow in the complex. When in the next step, a substrate or co-factor is recruited to the RXL site, or ATP binds at the binding site, the interface becomes more important in the signalling process.

The pattern we see in the CDK4 - cyclin D3 complex supports these results as can be seen in Table C.10.

Table 6.2: Average QS of CDK4 - cyclin D1 interface when sourced from different sites. Phosphorylation site is ALA172 and a full list of binding site residues is given in Table C.7. RXL site residues are listed in Table 6.1

	Source site		
Methodology	Phosphorylation site	Binding site	RXL site
Markov Transients Random Site Score [95% CI]	0.44 0.47, [0.47,0.48]	0.59 0.48, [0.48,0.49]	0.59 0.51, [0.50,0.51]
Bond-to-bond propensity Random Site Score [95% CI]	0.47 0.55, [0.54,0.55]	0.55 0.52, [0.51,0.52]	0.56 0.54, [0.54,0.55]

Although we detect differences in the dimer interface QSs depending on where the input signal is injected, the overall scores are low. This becomes apparent when comparing the interface scores to the respective random site scores in Table 6.2. Similarly to what we proposed in Chapters 4 and 5, our methodologies tend to pick up the most important interface residues which confer signalling rather than the whole interaction surface. Investigating the highest scoring residues as predicted by our methods provides a focus lens towards the parts of the dimer interface that can be targeted for disruption of the dimerisation process. Table 6.3 shows the top scoring interface residues according to MT and BBP analyses when sourced from the different sites where activation signals are induced.

The signals stemming from sites on the kinase, *i.e.* the phosphorylation site and the ATP binding site, highlight a cluster of residues towards the front of the interface just above the Cα helix (Fig. 6.11A). When we source our analysis from the RXL site on the protein, the highest scoring residues are mainly on the cyclin site of the dimer apart from two residues which are

Table 6.3: Top scoring residues in the CDK4 - cyclin D1 interface. Top 10 residues with the highest quantile score in MT and BBP analysis when sourced from different elements. Residues in chain A belong to cyclin D1, and chain B constitutes CDK4. Ph - phosphorylation site (ALA^{172}); Green - Cα residues; light red - common residues of MT analysis; orange - common residues of BBP analysis

Markov Transients			Bond-to-bond propensities		
Ph	binding site	RXL site	Ph	binding site	RXL site
GLU43 B	CYS78 B	THR116 A	PHE130 B	ASP76 B	PHE66 B
GLY42 B	ALA79 B	ASN151 A	PHE66 B	PHE130 B	GLU67 B
GLU44 B	GLY42 B	THR53 B	LYS33 A	ARG5 B	PHE130 B
LEU59 B	GLU44 B	ASN146 A	GLU67 B	GLU64 B	LYS112 A
ASN41 B	GLU43 B	LYS149 A	ALA30 A	ARG61 B	TRP150 A
ALA58 B	ARG82 B	TRP150 A	ARG61 B	PHE66 B	VAL57 B
ALA133 B	ARG5 B	ALA153 A	MET31 A	TRP150 A	GLU64 B
ARG55 B	SER81 B	LEU152 A	ARG62 B	ILE87 B	LYS33 A
ALA65 B	ASN41 B	MET113 A	GLU64 B	VAL89 B	ALA30 A
GLU67 B	GLU135 A	VAL57 B	ALA34 A	LYS33 A	ASN151 A

part of the Cα helix (THR^{53} and VAL^{57}).

Interestingly this pattern is shifted towards the back of the interface for the highest scoring residues in the bond-to-bond propensity analysis. Residues LYS^{33} of cyclin D1 and GLU^{64}, PHE^{66} and PHE^{130} of CDK4 are located towards the back of the interface (Fig. 6.11B). These results suggest that the signalling between CDK4 and cyclin D1 is conferred over two distinct clusters at the dimer interface. The front of the interface, near the Cα helix (which includes the PISTVRE motif), might contribute to a fast signal propagation as indicated by Markov Transients. These residues might provide a good starting point for an inhibitory event that disrupts the catalytic machinery of the protein. On the other hand, the back of the interface was especially high scoring in the bond-to-bond propensity analysis, indicating a strong coupling in this part of the interface. This residue cluster could hold potential for binding events that disrupt the dimerisation process.

Given the complementary role of CDK4 and 6 in the cell cycle and their synonymous interaction profile, it is fair to assume that the insights we gathered from the CDK4 - cyclin D complexes are transferable onto CDK6 bound to cyclins. However, final validation of this would require a structure of CDK6 in complex with a D-type cyclin, which has not been solved yet.

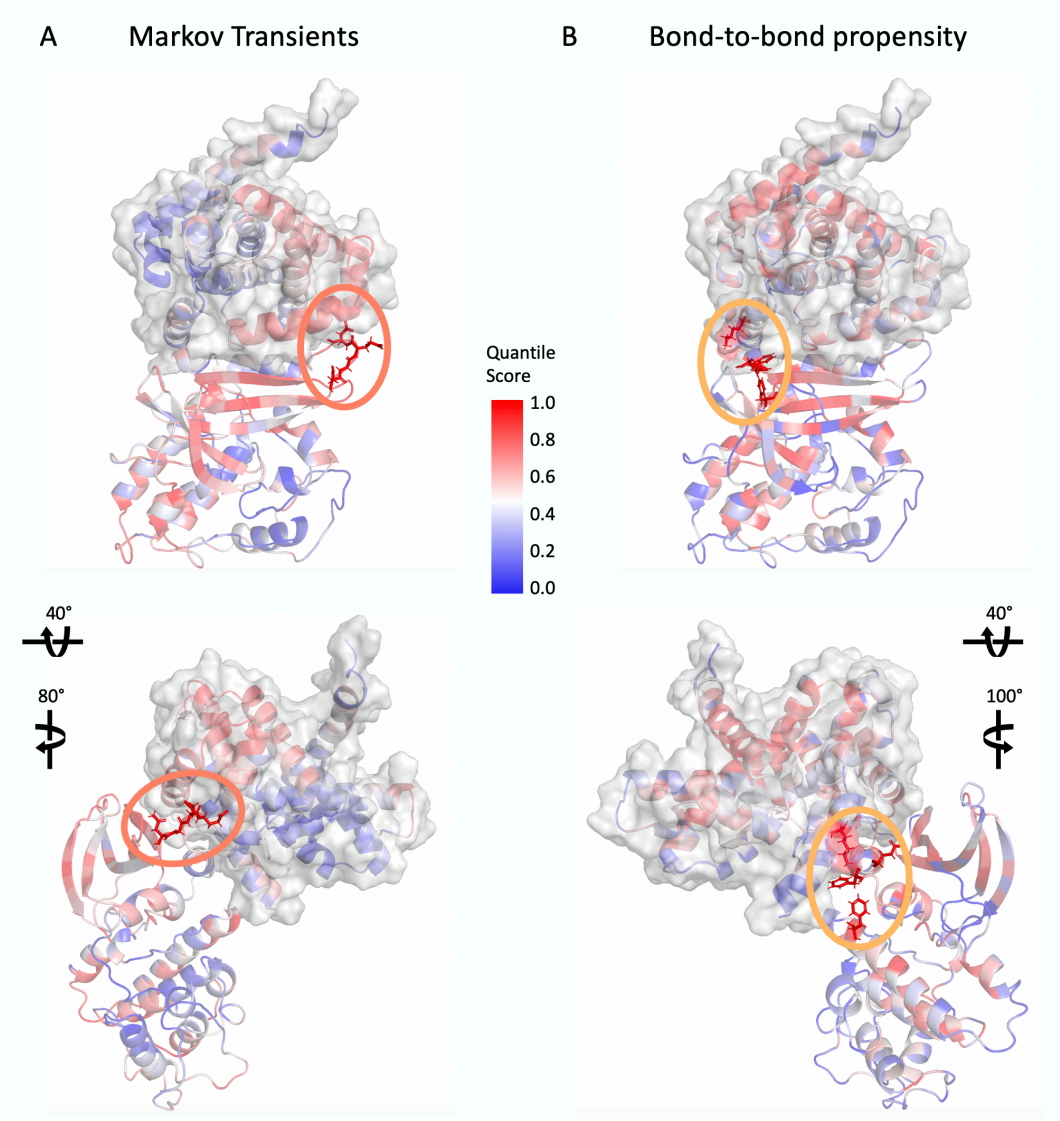


Figure 6.11: Two areas of interest in the CDK4 - cyclin D1 dimer interface were identified in the MT and BBP analyses. The structures of CDK4 - cyclin D1 (PDB id: 2W9Z[328]) are coloured by QS in the MT (**A**) and BBP (**B**) runs sourced from the binding site and shown in two orientations. The residues shown as sticks were identified as important in the interface from MT analysis (light red circle) and BBP analysis (orange circle). Residues as listed in Table 6.3. The cyclin is shown with a grey surface to visualise the interaction site.

6.4 The inhibition of CDK6 with cancer therapeutics

We learned from the monomeric forms that a first step in activating CDK4/6 would be the recruitment of the cyclin partners. We observed this when sourcing our methodologies from the phosphorylation site as demonstrated in Section 6.2 for CDK4. The results for monomeric CDK6 are consistent with those presented above, and the respective data is shown in Figure

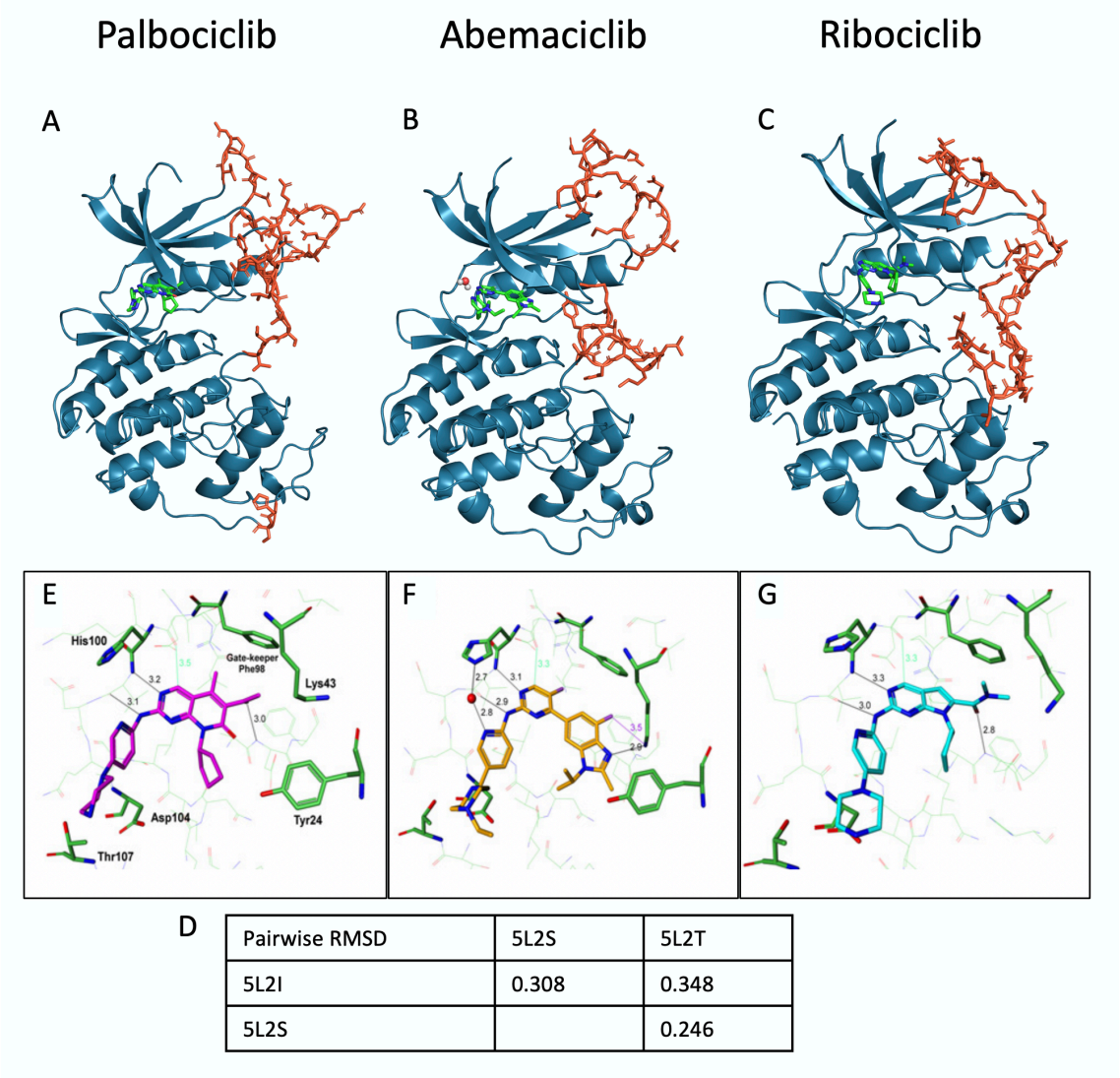
B.10.

In the next step, we are looking at inhibitor patterns and whether we can detect a divergence from what we saw in the monomeric structures. We are particularly interested in CDK4/6-specific inhibitors approved for treatment in BC: palbociclib, abemaciclib and ribociclib[342]. These drugs are active against CDK4/6 *in vivo* and are often used in combination therapies with anti-estrogens[53,343]. However, there are already indications of resistance mechanisms against these inhibitors[344] which stimulates a continuous need for alternative drug targeting strategies. This Section aims to shed light into their inhibitory mechanisms on an atomistic scale.

6.4.1 Chemotherapeutics in CDK6

Chen et al.[354] did a comparative study of CDK inhibiting drugs in which they solved the structure of monomeric CDK6 bound to the approved compounds palbociclib, abemaciclib and ribociclib as shown in Figure 6.12. Unfortunately, these structures have gaps in the loop areas, which are often flexible regions of the proteins that are hard to solve in static X-Ray crystallography[355]. For our purposes, we need a fully connected graph and hence modelled these loops using Chimera[356]. Full details of the loop closing workflow can be found in Appendix A.1.3 and the resulting loops are shown in orange in Figure 6.12. After loop modelling, the structures were close in overall RMSD as listed in Figure 6.12D. We use these structures as the basis for our investigation into the molecular mechanisms of CDK6 inhibition by these compounds.

Figure 6.13 summarises the results of the MT analyses when sourced from the phosphorylation site THR^{177}. We detect the highest variation in connectivity towards the Cα helix, where cyclin binding would happen. As we proposed in Section 6.2, the high scoring Cα helix in monomeric CDK4 and 6 indicates the need for cyclin binding stimulated by the phosphorylation event. We here see an overall much slower signal propagation towards the helix from an average $t_{1/2}$ of 999.72 in apo CDK6 to 2835.83, 6058.13 and 2515.52 in CDK6 bound by palbociclib, abemaciclib and ribociclib, respectively (Fig. 6.13C). However, when we consider the single residue QSs of the Cα helix which consider the distance from the source, we detect some



Pairwise RMSD	5L2S	5L2T
5L2I	0.308	0.348
5L2S		0.246

Figure 6.12: Structures of CDK6 bound to three inhibitors. The structures are shown for CDK6 bound to palbociclib (**A**, PDB id: 5L2I[354]), abemaciclib (**B**, PDB id: 5L2S[354]) and ribociclib (**C**, PDB id: 5L2T[354]). The compounds are shown as green sticks, and modelled loops are shown in orange. The panel below shows the inhibitor binding mode in the binding site for all three ligands (**E, F, G** from Chen et al.[354]). For abemaciclib (**F**) a water molecule that contributes to the binding was kept in the structure for all analyses. Distances are shown in Å. **D)** Pairwise RMSDs for all three structures were calculated with PyMol[191].

intriguing patterns. Although, the inhibitors on average decrease the connectivity towards the helix, there are large differences, especially between abemaciclib and ribociclib. In the case of inhibition by abemaciclib, the Cα helix becomes a cold spot with an average QS of 0.12. However, for ribociclib the Cα helix scores much higher with an average QS of 0.92. Intriguingly, these patterns align with the data presented by Chen et al.[354] where they found abemaciclib > palbociclib >> ribociclib in terms of inhibition of pRB phosphorylation and cell proliferation of BC cells.

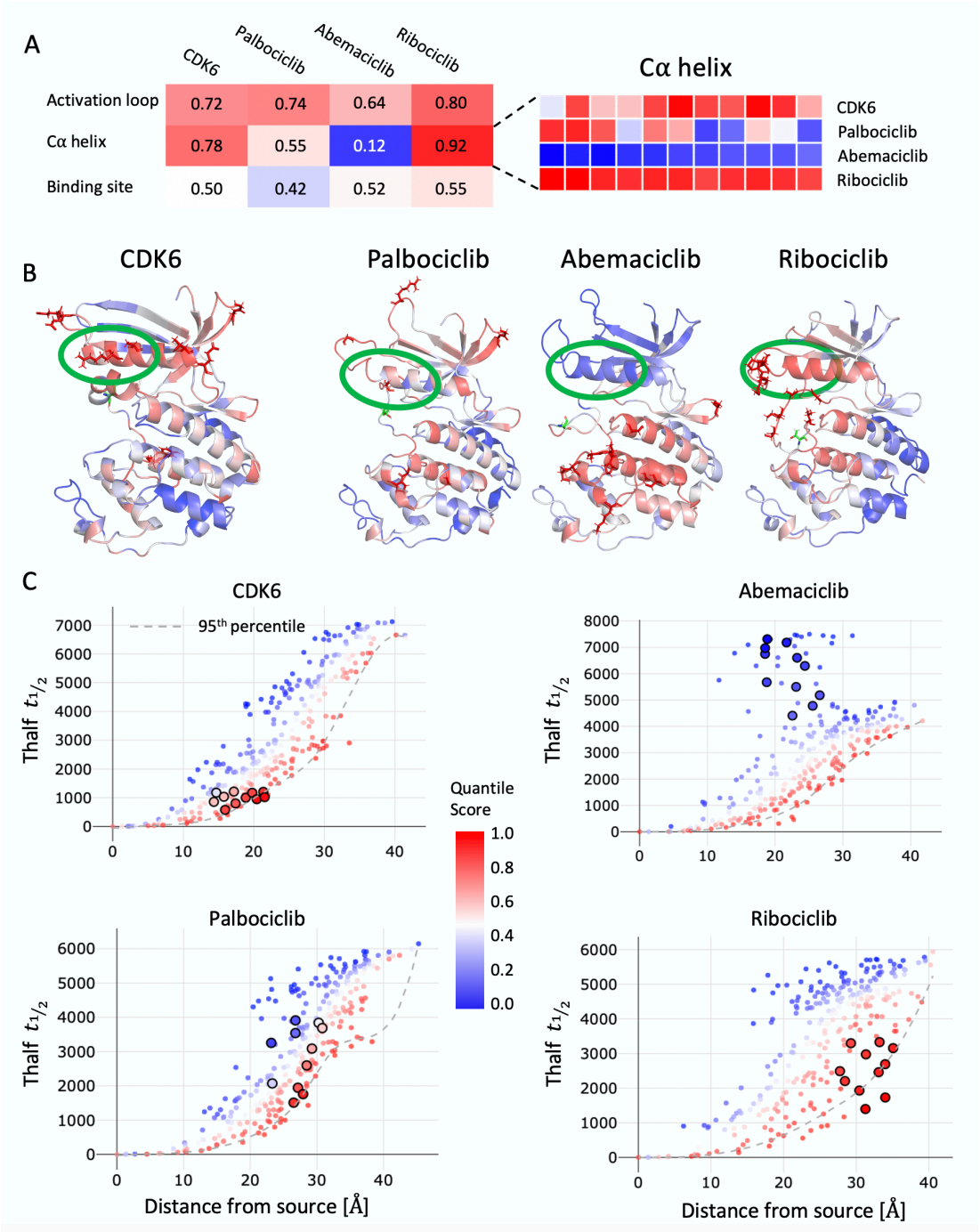


Figure 6.13: Markov transient analysis of monomeric CDK6 in apo and inhibited form when sourced from the phosphorylation site. A) Average quantile score (QS) results for Markov Transients are shown for each structural element in CDK6 in apo and inhibited forms. A zoom into the sequence of the Cα helix is provided to the right. **B)** The structures of monomeric CDK6 (AlphaFold model[49]) and inhibited forms with palbociclib, abemaciclib and ribociclib (PDB ids: 5L2I, 5L2S, 5L2T[354]) are shown from the back view with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a green circle are the Cα helices. **C)** Data distribution in the different structures which show $t_{1/2}$ values over the distance from the source for each residue in the protein. Cα residues are highlighted as larger dots with a black outline.

The effects that might explain CDK6 inhibition seem to be primarily conferred via fast signalling pathways as we can pick them up with Markov Transients. For bond-to-bond propensities, we see similar patterns between apo CDK6 and the inhibited structures but to a lesser extent (see Fig. B.11). For the uninhibited structure of CDK6 the Cα helix scores with an average QS of 0.54. The same score was found for the structure in complex with palbociclib, while the QS was lowered to 0.24 when CDK6 was bound to abemaciclib. While the structure bound to ribociclib had a higher average QS of 0.80 for the Cα helix.

6.4.2 Comparison to inhibition in CDK2

To provide first insights into the mechanisms underlying selective inhibition of the cell cycle kinases, we chose to run our methods on an inhibited structure of CDK2. Figure 6.14 provides an overview of the structure, which was chosen based on the high resolution and because there are no gaps in the structure (PDB id: 2B54[357]). The inhibitor was part of a chemical optimisation series of pyrazolopyrimidines and was found to have an inhibitory effect on the catalytic activity of CDK2 - cyclin E as well as CDK4 - cyclin D1[357]. The structure is in the inactive conformation, with the activation loop blocking the binding site and preventing the assembly of substrate peptides.

Following the approach for the study of monomeric CDK6 with and without inhibitors, we sourced an MT analysis from the phosphorylation site THR^{160}. Figure 6.15 shows the results of the analysis. We do not detect a remarkable difference when comparing the apo results to the inhibitor bound structure. All structural features are equally well reached via fast signal propagation, with the activation loop being highlighted here. As discussed in Section 6.2, the activation loop is an important feature of CDK2 activation as it needs to undergo structural rearrangements to allow the binding of ATP and substrate assembly[19]. When we compare these results to the patterns we have detected in CDK6 (Sec. 6.4.1, Fig. 6.13), we see no disruption to the Cα helix signalling in CDK2. Based on these results, we can assume that the inhibition in CDK2 must be conferred over a different mechanism which is yet to be revealed.

These Sections aimed to provide first insights into the differential inhibition patterns in CDK4/6

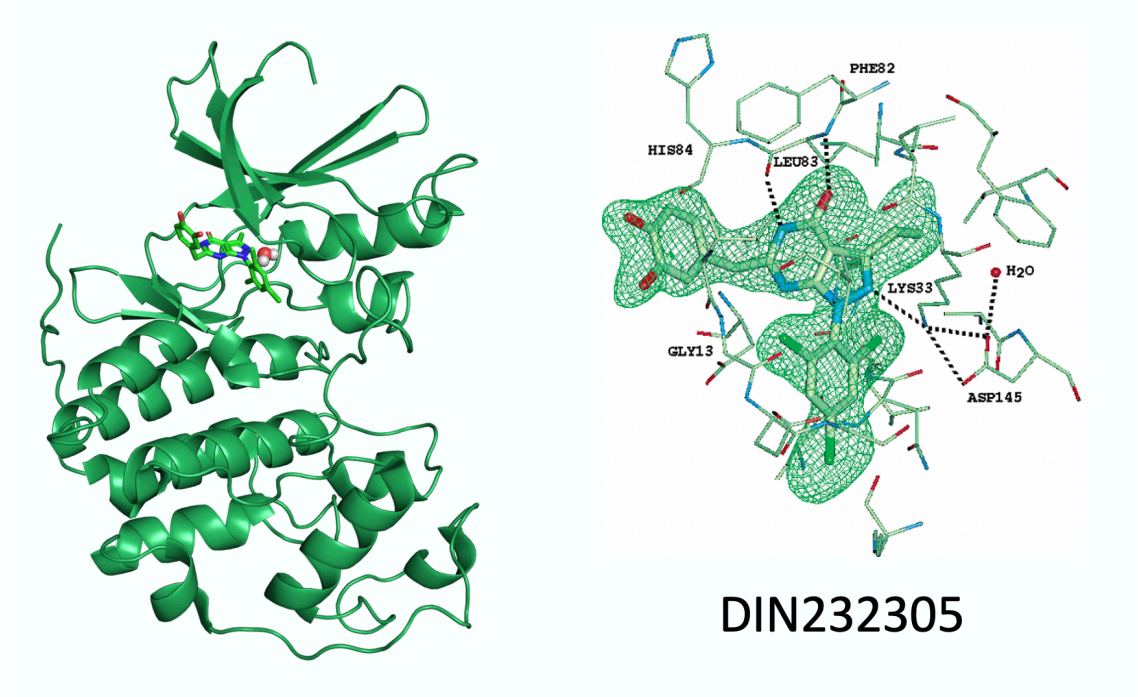


Figure 6.14: CDK2 bound to inhibitor DIN232305. Shown is the structure of CDK2 bound to an inhibitor molecule shown as green sticks (PDB id: 2B54[357]). On the right is a close up of the inhibitor binding mode in the binding site including a water molecule. On the right, from Markwalder et al.[357].

versus CDK2. We showed here that the inhibition in CDK6 is likely connected to a disruption in the signal going towards the Cα helix where the cyclin binding partner binds. These insights were obtained with MT analyses in CDK6 apo and inhibited structures which found a much slower signal propagation in the structures bound to chemotherapeutics. This disruption cannot be detected in the CDK2 structures, where the overall signal patterns are very similar for the apo and inhibited protein.

We would like to point out that it needs to be kept in mind that large loops of the CDK6 inhibited structures were modelled (Fig. 6.12). Among others, the loop where the phosphorylation site is located. While interesting preliminary results that go towards elucidating the inhibitory mechanism in CDK6 were presented here, further investigation in fully solved structures is needed.

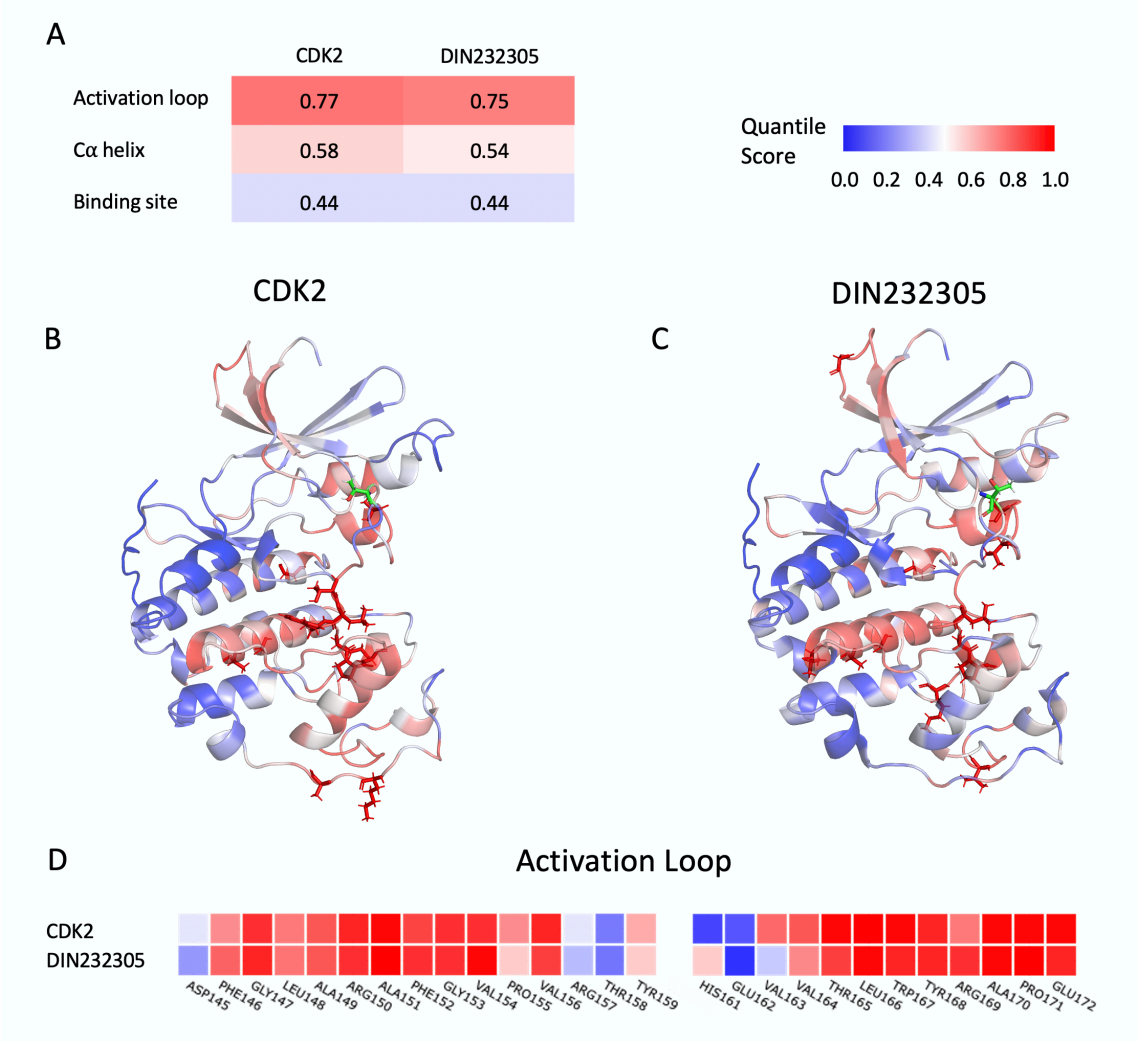


Figure 6.15: Markov transient analysis of monomeric CDK2 in apo and inhibited form when sourced from the phosphorylation site. A) Average quantile score (QS) results for Markov Transients are shown for each structural element in CDK2 in apo and inhibited form. **B) and C)** The structures of monomeric CDK2 (PDB id: 1HCL[346]) and bound to DIN232305 (PDB id: 5L2I[357]) are shown with residues coloured by QS. Residues with a QS $>$ 0.95 are shown as sticks. The source residues are shown as green sticks. **D)** Detailed sequence for the activation loop residues coloured by QS.

6.5 Conclusions

This Chapter saw the application of our atomistic graph analysis in a less-studied system where fewer structural data is available. We detailed the molecular interactions between CDK4/6 and their binding partners, cyclin D1 and D3, showing that the activation mechanism of CDK4 might be more reliant on a cyclin binding event at the Cα helix upon phosphorylation. In contrast, the monomeric form of CDK2 showed a lower connectivity towards the Cα helix (Fig.

6.4), indicating a weaker interdependency between phosphorylation and cyclin binding.

Molecular activation mechanism of the CDK4 - cyclin D complex

We also studied the heterodimeric CDK4 - cyclin D complex. By mimicking the different activation signals on CDK4 and cyclin D, we described the multi-factorial activation process of the complex:

- We detected rapid communication towards the cyclin D RXL site from the phosphorylation site, as well as from the binding site when applying MT analysis (Fig. 6.6). In the BBP analysis where we investigated the instantaneous perturbation connectivity, we detect a two-step process, from the phosphorylation site, over the ATP binding and towards the RXL site on cyclin D1/3 (Fig. 6.7). This Section demonstrated the complementary nature of MT and BBP analyses, where we found that they detect the same functionally important sites while highlighting different aspects on the residue level.

- MT analysis further allowed us to reveal extended regions of high scoring residues. Similar to what was shown by F. Vianello[163] for BBP analysis, we propose these hotspot regions are protein-protein interaction sites. We found preliminary validation for this prediction when showing that the positioning of p21 and p27 is in agreement with an extended hotspot that bridges from CDK4 to cyclin D3 (Fig. 6.8B).

- We further sourced our analyses from the RXL residues on cyclin D1. We found a high connectivity towards the functionally important elements in CDK4, including the proposed PPI for assembly of further co-factors (Fig. 6.9).

Taken together, these observations support the idea that multiple signals are required to facilitate the activity of the CDK4 - cyclin D complex. Our first analysis (Sec. 6.2) suggests a strong connectivity between the phosphorylation and cyclin binding sites. Once cyclin is bound and we investigate the patterns in the heterodimeric complex, we can not deduce a strict sequential order from our results. Instead, we propose that the phosphorylation event and the binding of

ATP come together dynamically with a substrate binding at the RXL site and potential further co-factor assembly to achieve the full complex activity.

Targeting the dimer interface

Tying in with our work from Chapters 4 and 5, we were able to highlight specific residues in the dimer interface between CDK4 and cyclin D1 that could be target points for a drug design distal from the active site. Interestingly, we found two distinct areas of the interface that were important in respective MT and BBP analyses. These differential residues might provide anchor points for two distinct targeting approaches. One approach might lead to disruption of the catalytic machinery (Fig. 6.11A) and another one (Fig. 6.11B) that disrupts the strong coupling between the complex partners and maybe leads to the inhibition of dimerisation.

Cancer therapeutics in CDK6

Section 6.4.1 saw the study of monomeric CDK6 inhibited by three FDA approved cancer therapeutics: palbociclib, abemaciclib and ribociclib. We detected an overall much slower communication towards the Cα helix where a cyclin binding event could be the first signal towards activation (Sec. 6.2). We further showed that the detected connectivity patterns towards the Cα helix (Fig. 6.13) overlap with experimentally found inhibitory activity of the drugs[354]. However, we would like to point out that the three CDK6 structures had missing loop regions that we modelled to run our analysis on full-length structures (Fig. 6.12). While we detect intriguing first results for the inhibition of CDK6, further confirmation in fully solved structures is required.

The results we obtained from studying an inhibited structure of CDK2 provide first indications that a different inhibition mechanism is at play in the monomeric CDK2. Our atomistic graph analysis can detect preliminary differences between the inhibition of CDK4/6 and CDK2, which rely on the connectivity towards the Cα helix where cyclin binds (Fig. 6.15). Notably, this confirms what is found for the approved cancer therapeutics which inhibit CDK4/6 but not CDK2.

Overall, we saw that the main differences between the CDK proteins and their intra-protein connectivity were revealed with MT analysis. This observation aligns with what we observed in Chapters 4 and 5: Markov Transients are a valuable tool in catalytically active proteins where they reveal fast signalling connectivities. Bond-to-bond propensity analyses in CDK4/6 showed weaker tendencies but still contributed to revealing differential concepts in connectivity between the sites of activation. Especially regarding the dimer interface, the complementary usage of both methodologies allowed us to reveal distinct areas of intra-molecular communication.

Chapter 7

Conclusion

This Thesis constitutes the application of atomistic graph analysis of proteins for the purpose of exploring alternative drug-targeting approaches. We studied three protein systems important to the disease systems coronavirus disease 2019 (COVID-19) and breast cancer (BC). Our work provides targeted approaches for their inhibition in an efficient manner. Broadly, we demonstrated how information on biological concepts like allosteric modulation and dimer interactions, can be gained from graph-based computational studies rooted in physicochemical descriptions of a system. We provide detailed insights into molecular mechanisms, and we hope that our results provide valuable guidance for rational drug design.

Alternative drug targeting mechanisms are of considerable importance in closely related protein families and systems that develop resistance against common drugs, as often seen in recurrent tumours. To overcome the limitations of active site targeting[17], we introduced two alternative targeting mechanisms in Chapter 1: interruption of the dimerisation process in functionally obligate dimers and allosteric inhibition. By extending the scope of computer-aided drug design (CADD) onto dimer interfaces and allosteric sites, the chemical search space is widened, and the scope for drug design in traditionally "undruggable" targets is expanded. The increase of data availability and constant advancements in techniques motivate the importance of CADD, and Chapter 2 describes the significance of computational methodologies in exploring alternative drug design concepts.

This work applies an all-atom yet efficient computational framework to study three dimeric protein systems and describes their potential for alternative drug targeting. We build our atomistic graph analysis of dimeric proteins on work by B. Amor[168] and F. Vianello[163] and apply two complementary methodologies: Markov transient (MT)[159] and bond-to-bond propensity (BBP)[149] analyses. These methods provide a measure for fast and strong coupling within a protein and find residues and sites that are crucial for modulating protein activity. We here extended the application onto dimeric proteins that are more complex with multi-layered signalling processes that contribute to their functionality. In doing so, we can validate known molecular mechanisms, provide insights into activation signals, and find hotspots for allosteric modulation or dimer disruption.

The following Sections summarise the main findings in the three study systems and highlight how each protein can be targeted with alternative drug targeting for disease. We also explore open questions and future directions that arise from our work.

7.1 Summary of biological results

Estrogen receptor α

Firstly, we studied estrogen receptor α (ERα), which modulates the cellular response to estrogens and is a key factor in the context of BC[3]. In Chapter 4, we applied atomistic graph analysis to study the homodimeric ERα ligand-binding domain (LBD), where natural agonists and antagonistic chemotherapeutics bind, and that is essential for the transcription activation function 2 (AF-2). We studied the ERα LBD with MT and BBP analyses and found that the former did not show signalling patterns beyond a uniform diffusion process from the source sites. This is in line with previous work in our group that detected that Markov Transients are primarily applicable in enzymatic proteins[159,161] and motivated us to focus on bond-to-bond propensities in the non-catalytic ERα.

We built on our previous work[202] that validated the molecular mechanism of ERα under the regulation of agonists and antagonists and confirmed the importance of helix 12 (H12) posi-

tioning and the dimerisation. We further explored the dimer interface that is formed in the homodimeric ERα ligand-binding domain (LBD) and showed how bond-to-bond propensities can highlight critical residues in the interface. We propose that these residues might be target points for inhibiting the functionally essential dimerisation process, as has been proposed before[98].

Notably, we found the dimer interface to be further implicated in resistance mechanisms that have been described for the BC L536R mutation of the ERα LBD. Based on experimental data provided by Fui Lai and Simak Ali, we investigated the differential inhibitory effects that two recently developed selective estrogen receptor degraders (SERDs) have on the L536R mutant. Our results indicate that the experimentally observed differential resistance might be conferred over dimer interface stability and that targeting the dimerisation process might help to overcome resistance in recurrent tumours.

SARS-CoV-2 main protease

The global pandemic of COVID-19 that started in early 2020 led to an unprecedented focus of scientific efforts to target proteins of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to combat the disease. Chapter 5 explores the homodimeric main protease (M^{pro}) of SARS-CoV-2 to identify targeting strategies that inhibit viral replication by interfering with the proteolytic machinery. We found that bond-to-bond propensities were valuable in studying the underlying molecular mechanism that includes the dimerisation process. Similar to what we found for the ERα LBD interface, we detected high scoring residues that are involved in indirect activation of M^{pro}. These studies were guided by previous results in the former severe acute respiratory syndrome coronavirus (SARS-CoV) that caused an outbreak of SARS in 2002/2003[285]. We were able to confirm for both viruses that residues that are important for dimerisation communicate with the N-finger residues which connect to the substrate-binding site.

We further demonstrated the application of MT and BBP analyses in a more traditional approach[54,149,159,161,162] to predict four putative allosteric hotspots that can be targeted for alter-

native inhibition approaches. Notably, one of these hotspots is located in the dimer interface and targeting this area might have a dual advantage as it could also be used to interrupt the obligatory dimerisation of M^{pro}. We went on to confirm the targetability of our hotspots by integrating data from a recent crystallographic fragment screen[292] and established which fragments exhibit a direct connectivity towards the active site. These results provide relevant insights into allosteric regulation in the SARS-CoV-2 M^{pro}, and we are confident that they are valuable starting points for rational drug design.

CDK4/6 and D-type cyclins

The cell cycle is under tight regulation from the interplay between kinases and cyclins[6]. They form heterodimeric complexes essential for functionality and are implicated in a wide range of cancer types[335]. In Chapter 6, we explored the G1-phase cyclin-dependent kinases (CDKs) 4 and 6, which form complexes with D-type cyclins and are less studied than other CDKs, like CDK2. Our atomistic graph analysis revealed the interplay of multi-factorial input signals required to achieve activation. Further, we revealed that Markov Transients predict extended hot areas on the protein surface that we propose to be PPI sites for co-factor assembly.

This Chapter also demonstrated how MT and BBP analyses complement each other in predicting hotspots on the protein while highlighting different high scoring residues. In terms of the interface between CDK4 and cyclin D1, these results revealed two areas that might confer different communication signals within the complex and could exhibit differential effects when targeted. Given the limitations placed on our approach by the availability of structural data, the results found for CDK4 in complex with D-type cyclins await confirmation for the CDK6 - cyclin D complex.

However, with available structures of CDK6, we were able to present insights into the differential mechanisms of chemotherapeutic inhibition in cell cycle kinases. Again, our results need to be considered in light of structure availability as we had to model large loop regions in the structures. Nonetheless, we were able to confirm experimental inhibition patterns[354] and are confident that Markov Transients provided valuable insights into the inhibition mechanism.

Furthermore, we did not pick up the same weakened communication with the Cα helix in CDK2 and postulate that a different inhibition mechanism is at play in CDK2.

In conclusion, we have shown that atomistic graph analysis is a valuable tool for studying the intra-structural communication in dimeric protein complexes. This Thesis confirms the molecular mechanisms in disease-relevant proteins and explores how they can be modulated by targeting allosteric sites and dimer interfaces. The prediction of allosteric sites in proteins has been demonstrated extensively by our group for Markov Transients[159,161] and bond-to-bond propensities[149,162]. We further confirmed the validity of this application in large benchmarking sets[54] and published our approach for public use as a web server[51]. The computational efficiency of the approach allows us to react quickly to new threats like COVID-19, and we show that we can provide valuable insights into resistance mechanisms of cancer mutations. We hope this Thesis will guide targeted drug design, and we describe below which experimental studies could benefit from our results.

7.2 Open questions and future work

7.2.1 Suggested future experiments

We have shown numerous times that our atomistic graph analysis is an efficient tool to study communication within proteins and predict allosteric signalling and sites. Over the years, our group performed a range of benchmarking studies on state-of-the-art data sets available at the time[54,149] and in-depth explorations of biological systems[149,159,162,163] with predictions that were experimentally verified *in vitro* and *in vivo*[161].

Furthermore, the results we presented in this Thesis are in agreement with experimental observations in multiple instances. Especially for Chapter 4, we detail how our results overlap with experimental data and with what has been described as the molecular mechanism of ERα in literature[3]. Over the course of the last year, we have seen an increase in studies of the SARS-CoV-2 M^{pro} that describe allosteric effects and sites which overlap with our results as

presented in Chapter 5. Thus we are confident in the predictive power of our methodologies and hope that they give rise to rationally guided experiments.

Possible experiments that would confirm our results can be guided by the hotspots and high scoring residues that have been presented in this work. Generally, these experiments would require the design of small molecules that can bind in proximity to the allosteric hotspots or high scoring residues in the dimerisation sites. For the latter, these molecules could be extended towards inhibitory peptide design, leading to dimerisation disruption.

For ERα, it would be of high interest to see whether the observation that differences in drug resistance in the cancer mutant L536R are conferred over dimer stability, can be confirmed. To this end it might be fruitful to study dimer dissociation rates of ERα constructs while bound to different SERDs. This approach has been under discussion with our collaborators, the group of Simak Ali in the Department of Surgery & Cancer at Imperial College London.

For the allosteric hotspots that we predicted in the SARS-CoV-2 M^{pro}, it would be exciting to see whether a small molecule can be designed to bind at the high scoring residues. Good starting points for chemical modifications could be the fragments that are binding in proximity to our hotspots and for which structural data was made available through a crystallographic screen[292]. Once such molecules are designed, it would be intriguing to see whether a binding event leads to up or down-regulation of M^{pro}. Studies exploring the activity of fragments or drugs binding distal from the active site have so far found that they exhibit inhibitory allostery on M^{pro} [291,294]. This gives us confidence that it would be viable to study our allosteric hotspots with experimental means.

For CDK4 in complex with cyclin D we identified two differing residue clusters at the dimer interface with MT and BBP analyses. It would be most interesting to see whether an inhibition targeted at these high scoring residues would lead to a differential effect. We propose that the residues identified with MT analysis might lead to a disruption of the catalytic machinery, while the high scoring BBP residues might impact the dimer stability.

7.2.2 Impact of different inhibitors

We have shown in Chapters 4 and 6 that our methodologies can fruitfully be applied to investigate how binding of chemotherapeutics affects the connectivity within protein structures. Our explorations in that regard are limited by available structures of proteins in complex with inhibitors. To overcome this limitation, it would be of high interest to use docking approaches to create structures of proteins bound to a range of different modulators. As described shortly in Section 2.1, docking can be used in virtual screenings to explore which molecules would bind to a given target[60]. However, it can also be used to model the structure of a specific ligand bound to a target protein and determine the binding pose. Such a knowledge-guided docking experiment benefits from a known binding area as well as already bound template ligands to guide the simulation. For the LBD of ERα as well as the monomeric forms of the different CDKs, this information is available and could be used to set up docking experiments. Specific questions that come to mind based on the results that we presented in these systems are the following:

- Which impact do other SERDs have on the connectivity in mutated structures of ERα? Structures were only available for the ERα LBD bound to AZD9496 and AZD9833, and thus, no information can be provided for other SERDs. However, our collaborators obtained experimental data on their inhibitory values, and it would be interesting to see whether we detect the same effect over the dimer interface connectivities.

- Do we see differences between CDK4/6 versus CDK2 when inhibited with palbociclib, abemaciclib or ribociclib? These three cancer inhibitors have been approved for therapy based on their specific inhibition of CDK4 and 6. Unfortunately, there is no structural data available for these inhibitors bound to CDK2, for which they show weaker or no inhibitory behaviour[354]. It would be interesting to see whether we can elucidate the different mechanisms in a direct comparison of CDKs bound to palbociclib, abemaciclib and ribociclib.

7.2.3 Elucidation of different allosteric mechanisms in proteins

This work saw the application of two complementary approaches on atomistic graphs. MT analysis allows us to study which areas of the protein are reached the fastest by an input signal at a source site[159]. BBP analysis finds the areas of the protein that are coupled the strongest to the source site[149]. In Chapters 5 and 6, we found that the methodologies find similar signalling clusters in the protein but highlight different patterns on the single residue level. We further stated that Markov Transients are more applicable in proteins of catalytic activity, as shown by previous studies in our group[159,161].

Nonetheless, it would be interesting to explore these observations in larger datasets. As done for bond-to-bond propensity in work by Wu et al.[54], Markov transient analysis could be benchmarked in a large dataset of known allosteric proteins. This would provide us with quantitative measures for how well Markov transient analysis is performing overall and whether trends in the predictive pattern are skewed towards enzymatic proteins.

Ultimately, these large-scale studies do not only serve as an evaluation of the predictive power of our methodologies, but they can also teach us about allosteric mechanisms that are at play. By studying for which proteins allosteric sites can and cannot be predicted, we obtain a dataset of protein features that can be used for classification approaches. This will teach us about the applicability and limitations of our methodology but can also aid in defining different classes of allosteric mechanisms as described in Section 1.3.

7.2.4 *In silico* mutational analysis

As shortly mentioned in Section 2.2 in the context of PPI interactions, mutational studies are a widely applied tool to detect which residues are crucially involved in signalling patterns. In the context of our work, we could set up *in silico* alanine mutations of every position in the protein and run our atomistic graph analysis on the so created structures. These scans would result in high-dimensional output data that can be compared with statistical means to detect

outliers in two aspects. Firstly, it would be possible to detect positions that change the global connectivity within the whole protein or towards specific structural features. This approach has been used in ribosomal protein S6 kinase 4 (RSK4) to determine which residues impact the Markov transient signalling propagation between allosteric and orthosteric sites the most, which led to the detection of allosteric paths[161]. In the case of ERα, the data could be analysed with regard to H12 residues, which would allow identifying positions that have a particular high communication towards the helix. These positions might stabilise AF-2 over connecting to H12 and could foreshadow future resistance mechanisms.

Secondly, a full alanisation scan and subsequent atomistic graph analysis would identify positions that are impacted the most by alterations in the protein. The residues that are the most sensitive to mutations might constitute topologically relevant positions in the protein. It would further be possible to evaluate whether these positions are chemically important by mutating them into residues other than alanine and evaluating the physicochemical perturbation that would be introduced into the atomistic graph.

ProteinLens 2.0

In Section 3.2.2, we introduced ProteinLens, an interactive web server that makes our atomistic graph analysis accessible to the community. The web server was published in a first version to include the current graph construction process as coded in BagPype[160], and MT and BBP analyses. It further includes a scoring functionality that allows the user to investigate sites of interest against the backdrop of the whole protein[51].

For future versions of the web server, we would envision providing additional functionality to the user. For example, we think it would be interesting to include a feature that allows the user to explore the impact of mutations on signalling within the protein. In recent years it has been repeatedly proposed that the impact of single residue mutations might be conferred over allosteric effects in proteins[358,359]. Our methodologies would be ideal for exploring these effects by incorporating mutated residues into the graph. Our methodologies would then allow the user to investigate how a chosen mutated position influences the allostericity within the

protein and highlight differentially perturbed residues or paths.

Appendix A

Methodological Details

A.1 Structure details and pre-processing

The general workflow for structure curation followed the steps listed below:

1. **Download files from the PDB in `.pdb` format**[*].
2. **Clean A/B confirmations of side chains.** Sometimes side chain atoms are indicated with an alternate location record. We used a python script to choose the A conformation coordinates for all residue with double records.
3. **Adjust biological assemblies.** Some PDB files contained the wrong number of protein chains in the asymmetric unit cell for the dimeric assembly they were meant to represent. If that was the case, we either deleted surplus chains or modelled the missing dimer halves based on the symmetry information contained in the file header (`REMARK 350`). We used `MakeMultimer.py`, a python script that is freely available online[†], to replicate asymmetric unit cells.
4. **Curating water molecules that are functionally important.** Generally, all solvent water molecules are stripped from the structures before graph construction as `Reduce`, the

[*]Details of the current `.pdb` file format v.3.30 can be found at: wwpdb.org/documentation/file-format-content/format33/v3.3.html

[†]Available at: watcut.uwaterloo.ca/tools/makemultimer/

command line tool that we use for protonation, can not add hydrogens to single oxygen atoms. However, when described in the publication of a structure that a water molecule is involved in the function or binding mode of a protein, we curated these molecules to be included. We manually added hydrogen atoms with the `AddH` functionality in `Chimera`[356] to all water molecules and kept the relevant ones in the structure as indicated for each protein below.

5. **Modelling missing loops and residues.** Missing residues and residues with missing atoms are indicated in `REMARK 465` and `REMARK 470`, respectively. We modelled all non-terminal missing loops in `Chimera`[356] using the `Refine/Model Loops` option that integrates the computational tool `Modeller`[360]. In each run, we obtained five models and we diverged from the default settings by setting the number of adjacent residues that were allowed to moved to 0. The model with the best scores[360] and highest overlap with a close full-length structure was chosen. For missing side chain atoms we used PyRosetta to model their conformation[263]. The regions that we modelled are indicated for each protein below.

For Chapters 5 and 6 of this work, steps 2) and 3) were integrated in the BagPype graph construction workflow[160]. In the Sections below we provide details on the structures used for each study system.

A.1.1 Estrogen receptor alpha

For a basic comparison between agonist and antagonist-bound conformations of the ERα ligand binding domain (LBD) we use the PDB entries 1G50[227] bound to estradiol (2.9 Å resolution) and 3ERT[235] (1.9 Å resolution) bound to 4-hydroxytamoxifen, a commonly used chemotherapeutic. For the agonist-bound structure, we deleted chain A as it was not part of the homodimer and renumbered the residue identifiers from 1304 - 1549 to 304 - 549 in chain B and from 2304 - 2547 to 304 - 547 in chain C. The estradiol molecules were recorded at position 600 in both chains (`EST600`).

For the antagonist-bound structure, we modelled the second monomer based on the symmetry matrix using `MakeMultimer.py`. We further modelled side chain atoms for 11 residues in each monomer and added hydrogens to one water molecule in each binding site with the residue identifier 2. The 4-hydroxytamoxifen ligands are recorded at position 600 in chain A and B (`OHT600`).

Structural features

A structural feature that was investigated in both the agonist and antagonist-bound conformations was helix 12 (H12). In the agonist-bound conformation, H12 stretches from residue identifiers 539 - 547[219]. For the antagonist-bound conformation, a shift in H12 is recorded to include residues 536 - 544[235]. For the investigation of the dimer interface, we looked at the whole interface as well as the different structural features that make up the interface. Table A.1 lists the structural features in agonist and antagonist-bound conformation as defined in the `HELIX` section of their respective `.pdb` files.

Table A.1: Structural features in ERα LBD dimer interface. Residues in each structural element are shown for agonist and antagonist-bound conformation.

Structural element	Agonist-bound	Antagonist-bound
Helix 5/6	371-394	372-395
Helix 7	413-418	412-417
Helix 8	421-437	424-438
Helix 9	441-456	442-455
Loop 10	457-464	456-465
Helix 10	465-492	466-492
Helix 11	496-531	497-528
Helix 12	539-547	536-544

A.1.2 SARS-CoV-2 M^{pro}

We used two main protease structures, one in SARS-CoV-2 and one in SARS-CoV. The SARS-CoV-2 M^{pro} structure was solved to 1.75Å and deposited in the PDB as entry 6Y2E in March 2020[5]. One water molecule at position 582 was kept as it is in close proximity to HIS^{41}. The second monomer was modelled with `MakeMultimer.py`.

For the SARS-CoV M^{pro} we chose PDB entry 2DUC which contained no gaps or missing residues and was solved to 1.7 Å[277]. We kept the two water molecules at positions 342 in chain A and 312 in chain B.

Structural features

For the purpose of scoring the connectivity towards the active site, Nan Wu defined an extended region as follows. We used all 23 structures from a recent crystallographic fragment screen[291] with fragments non-covalently bound to the binding site of M^{pro}. We then used PyMol[191] to find all residues with atoms within 4 Å of any of the 23 fragments. Ultimately, the active site region we scored contained the following residues: THR^{25}, THR^{26}, HIS^{41}, CYS^{44}, THR^{45}, SER^{46}, MET^{49}, TYR^{54}, PHE^{140}, LEU^{141}, ASN^{142}, SER^{144}, CYS^{145}, MET^{162}, HIS^{163}, HIS^{164}, MET^{165}, GLU^{166}, LEU^{167}, PRO^{168}, ASP^{187}, ARG^{188}, GLN^{189} and THR^{190}.

A.1.3 Cyclin-dependent kinase 4 and 6

As described in Chapter 6, there is only a limited number of structures available in the PDB[48]. Figure A.1 provides an overview of available structures according to the proposed activation steps. The only structures available of the dimeric complex were CDK4 with cyclin D1 and D3 as described below.

For Chapter 6.2, we compared monomeric structures. For CDK2 we chose PDB entry 1HCL at 1.8 Å resolution[346]. Residues 37-40 were modelled as described above. Monomeric structures for CDKs 4 and 6 were not available in the PDB and were instead obtained from AlphaFold[49]. Figure A.2 shows the monomeric structures downloaded from AlphaFold. The terminal regions that were modelled with low and very low confidence were deleted from the structures.

For CDK4 in complex with cyclin D1 we chose PDB entry 2W9Z with a resolution of 2.45 Å[328] as it contained only one small gap in the kinase chain. This gap (residues 241-244) was closed with `MODELLER` in `Chimera`. The structure is mutated at the phosphorylation site 172 from

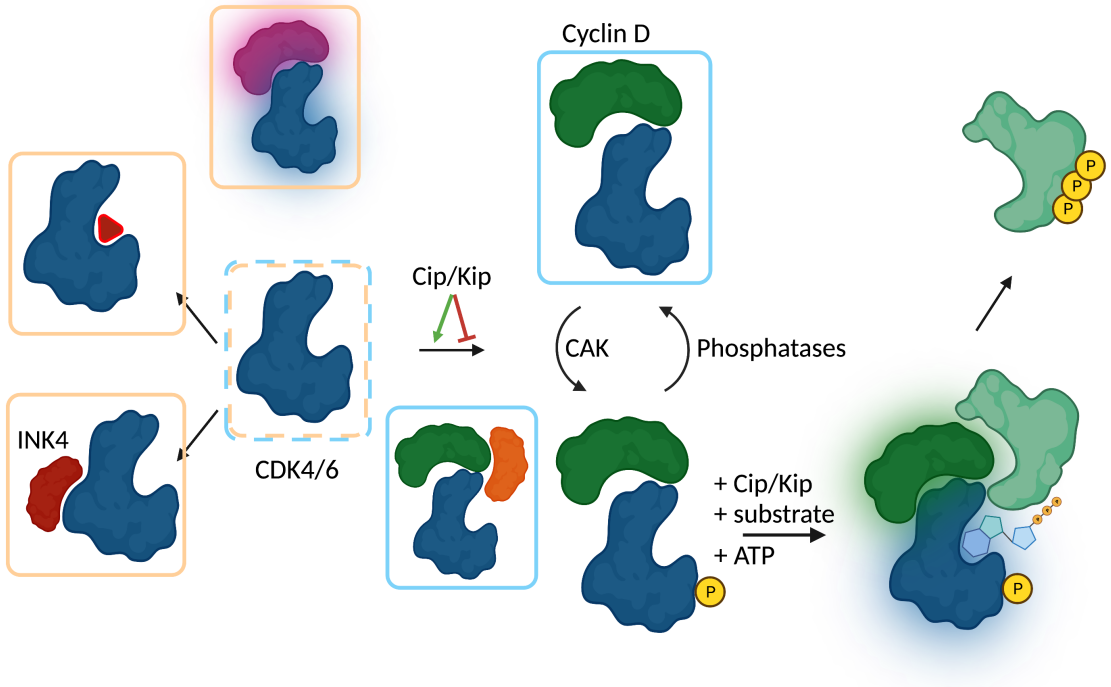


Figure A.1: Available structures of CDK4/6. The CDK4/6 activation pathway is depicted with CDK structures in dark blue and cyclin structures in dark green. Blue and yellow boxes depict structures that are available for CDK4 and CDK6, respectively. The dashed line indicates modelled structures which were obtained from AlphaFold[49]. Orange - Cip/Kip proteins; red - inhibitors (small molecules and INK4 proteins); light green - substrate; yellow circle - phosphate; purple - viral cyclin; CAK - CDK-activating kinase; ATP shown as simplified chemical structure.*

THR to ALA.

For the second complex used in this work, we used PDB entry 3G33[329] which contained CDK4 bound to cyclin D3 at 3 Å resolution. The file contained two dimers and we deleted chain A and D to obtain one dimer. We further renamed chain B to A and chain C to B and renumbered the kinase residues to match with the numbering in structure 2W9Z for ease of comparison. The 3G33 structure contained one gap in the cyclin monomer from 217-219, which was modelled with the approach described above. Missing atoms were added with PyRosetta[263] for nine residues.

Chapter 6.4 investigates monomeric structures of CDK6 and CDK2 bound to different chemotherapeutics. Table A.2 lists the three CDK6 structures obtained from the work by Guiley et al.[327]. The gaps were closed as described above and missing atoms were added with PyRosetta[263].

*Created with biorender.com

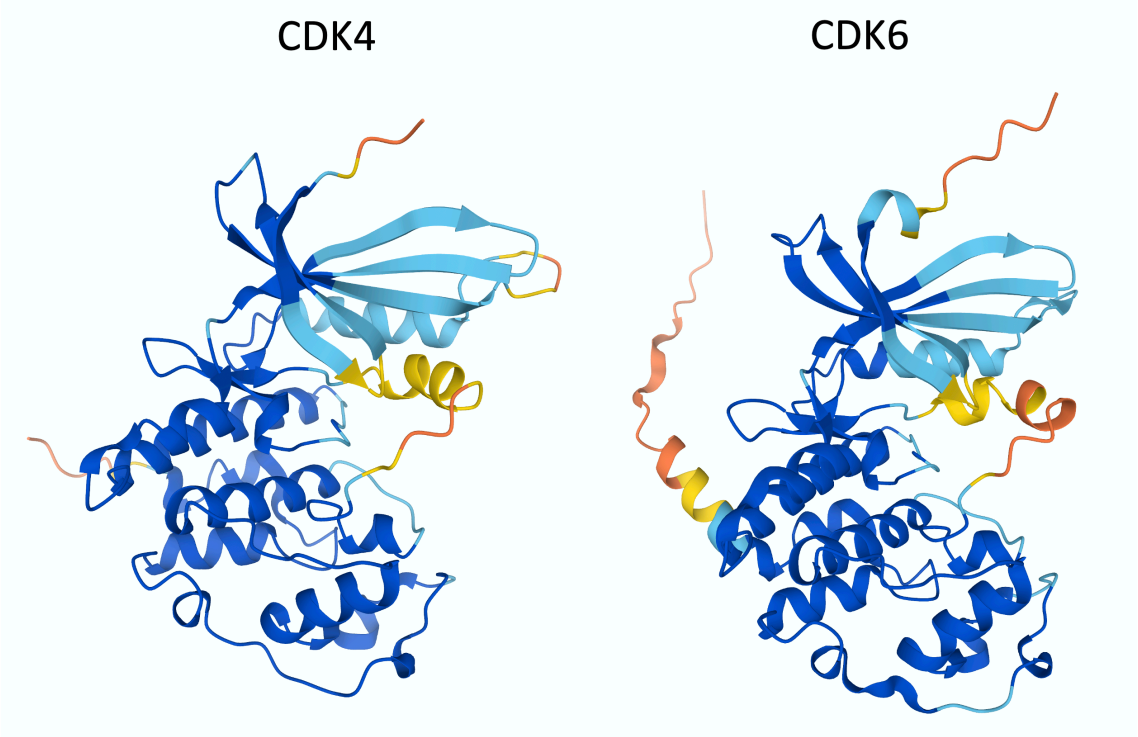


Figure A.2: AlphaFold prediction of monomeric CDK4/6 structures. Downloaded from alphafold.ebi.ac.uk/entry/P11802 for CDK4 and alphafold.ebi.ac.uk/entry/Q00534 for CDK6. Structures are coloured by modelling confidence score as defined by AlphaFold[49]: dark blue - very high (score > 90); light blue - confident (90 > score > 70); yellow - low(70 > score > 50); orange - very low (score < 50).

Table A.2: Characteristics of monomeric CDK6 structures with inhibitors.

PDB id	Inhibitor	Resolution [Å]	Gaps	Missing atoms
5L2I	Palbociclib	2.75	47-54, 85-92, 167-180, 255-256	1 residue
5L2S	Abemaciclib	2.27	48-54, 85-92, 168-180	6 residues
5L2T	Ribociclib	2.37	48-56, 85-92, 168-180	5 residues

For entry 5L2S we kept one water molecule at position 1011 as it contributed to ligand binding.

For an inhibited structure of CDK2 we chose the PDB entry 2B54 at a resolution of 1.85 Å[357]. The structure contained no gaps or missing atoms and the inhibitor DIN-232305 was recorded at position 300.

Structural features

As there is no structure available for CDK4/6 bound to the natural ligand ATP, we defined binding site residues based on an alignment with the binding site of CDK2. Echalier et al.[361] performed a detailed alignment of binding site residues and their conservation across multiple

members of the CDK family. Based on their work, we focused on the first shell binding site residues that form highly conserved motifs. We decided to apply a narrow definition of the binding site here, as we have shown in a recent study that the algorithm performs better with a small but relevant residue set[54]. A full list of binding site residues that were used in this work is shown in Table C.7.

A.2 ERα mutations and chemotherapeutics

> The experimental details in the following Sections were kindly provided by Simak Ali and Fui Lai. Experiments were designed by Simak Ali and Fui Lai and performed by Fui Lai.

Tissue culture and growth assays

MCF7-Luc cells (hereafter referred to as MCF7; Cambridge Bioscience, Cambridge, UK) and derived mutant ESR1 clones were authenticated by short tandem repeat profiling using the AmpFlSTR Identifiler Plus kit (Applied Biosystems, Warrington, UK), as described[362]. Mycoplasma negativity was maintained by regular testing using the MycoAlert Mycoplasma Detection Kit (Lonza, UK). Cell lines were routinely cultured in Dulbecco's Modified Eagle's medium (DMEM) containing 10% fetal calf serum (FCS) and penicillin-streptomycin-L-glutamine (PSG). For estrogen depletion, the cells were transferred to DMEM lacking phenol red and containing 5% dextran-coated charcoal-stripped FCS (DSS) for 72 h. Stock solutions of 17β-estradiol (EST) and anti-estrogens, prepared in DMSO, were added to the culture medium at a dilution of 1 in 1000. An equal volume of DMSO was added to the vehicle controls. Cell growth was measured using the sulphorhodamine B (SRB) assay, as described previously[363], or imaging using the IncuCyte ZOOM (Essen Bioscience, Welwyn Garden City, UK). For the latter, three images per well were acquired every 12 hours for a period of 6-9 days, and confluency (%) was calculated using the IncuCyte ZOOM software package (Essen Bioscience). For determination of the half-maximal effective concentration (IC50) values, cells were seeded in

96-well culture plates and treated with increasing concentrations of anti-estrogens for 6 days. Cell growth was determined using the SRB assay. IC50 values were calculated from non-linear regression curve fitting using GraphPad Prism v9. Doubling times were calculated in Prism, using the exponential growth equation.

Generation of ER-mutant MCF7 cell lines using CRISPR-Cas9 genome editing

The MCF7-Y537S A4 clone (here referred to as Y537S CL3) has been previously described[363]. The other ESR1 mutant lines were generated using the same approach, following site-directed mutagenesis of an 1,803 bp fragment of the ESR1 gene flanking the exon 8 coding region, except that MCF7 cells were transfected with the hCas9 and donor template plasmids, together with the CRISPR sgRNA CRISPR4834192 or CRISPR4834193. These CRISPRs, targeting intron 7 of ESR1, were designed using a web-based software*. Single colony cloning and screening of genomic DNA using mutant-specific PCR was undertaken as previously detailed[363]. PCR of genomic DNA followed by Sanger sequencing was used to confirm integration of the appropriate mutation in the ESR1 gene locus.

*Available at: zlab.bio/guide-design-resources

Appendix B

Supplementary Figures

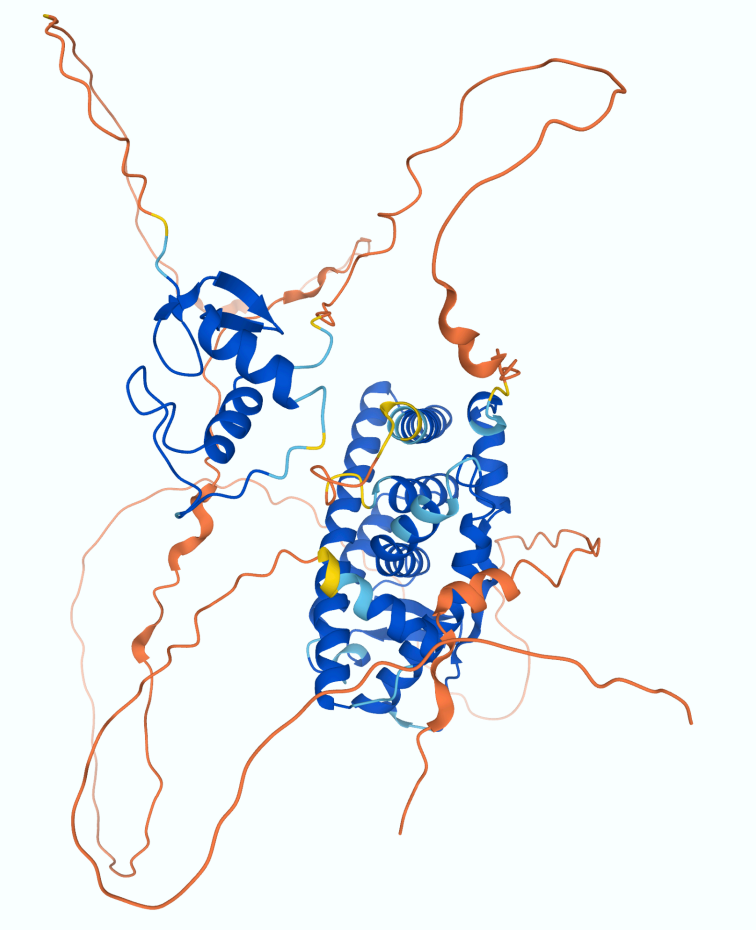


Figure B.1: AlphaFold prediction of monomeric ERα structure. Downloaded from alphafold.ebi.ac.uk/entry/P03372. Structures are coloured by modelling confidence score as defined by AlphaFold[49]: dark blue - very high (score > 90); light blue - confident (90 > score > 70); yellow - low(70 > score > 50); orange - very low (score < 50).

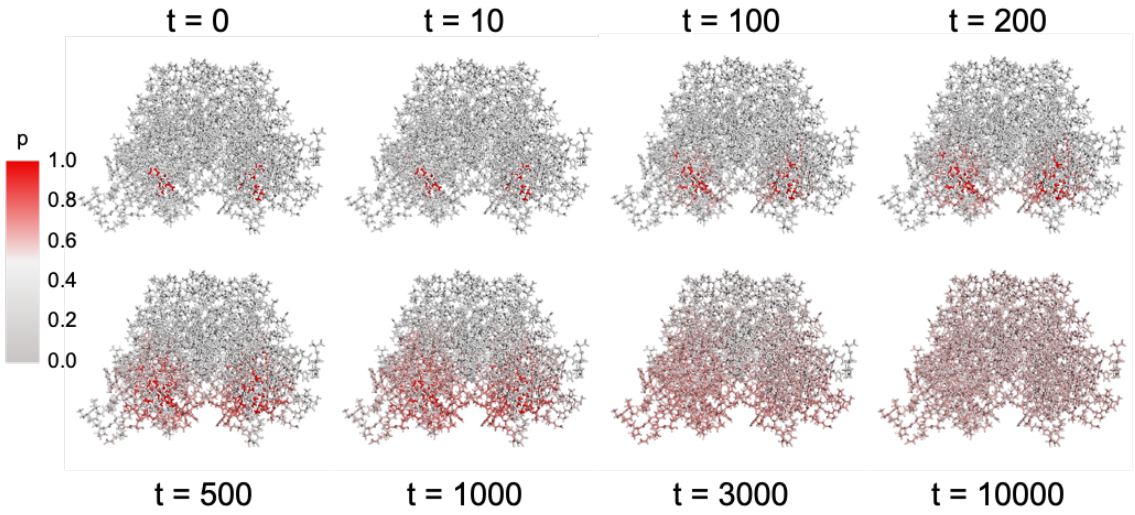


Figure B.2: Markov transient time steps in the antagonist-bound ERα ligand binding domain when sourced from the OHT molecules. The ERα LBD (PDB id: 3ERT[235]) is shown in an all atom stick representation. Atoms are coloured by probability (0 - grey to 1 - red) of the random walker being at this node at a given Markov time step t. Shown are different time steps in the Markov transient analysis as indicated.

A

	IC50 (nM)	MCF-7 Luc	L536R CL1	L536R CL2	L536R CL3	Y537C CL1	Y537C CL2	Y537N CL1	Y537N CL2	Y537N CL3	Y537S CL1	Y537S CL2	Y537S CL3	D538G CL1	D538G CL2	D538G CL3	D538G CL4
SERM	Tamoxifen	0.34	4.61	2.21	3.72	0.39	0.42	1.07	2.07	0.93	28.71	9.03	18.51	1.55	7.52	3.53	8.11
	Raloxifene	0.24	7.03	3.79	3.76	0.33	0.34	1.27	3.76	0.54	44.38	11.98	4.76	1.93	6.32	1.82	6.24
	Bazedoxifene	0.15	2.05	1.51	1.39	0.14	0.17	0.59	2.67	0.35	15.89	13.31	4.14	1.00	4.36	1.25	3.37
	Lasofoxifene	0.08	1.36	0.71	1.13	0.08	0.12	0.35	0.79	0.29	5.80	1.69	2.24	0.61	0.96	0.54	1.87
SERD	Faslodex	0.18	0.91	0.91	1.00	0.18	0.26	0.86	1.90	0.31	7.10	10.61	1.99	1.10	6.05	0.74	4.76
	AZD9496	0.13	0.65	0.53	0.52	0.12	0.23	0.62	2.15	0.27	19.95	15.34	2.31	1.44	2.20	0.93	3.02
	AZD9833	0.11	7.72	2.96	3.69	0.13	0.15	0.71	1.28	0.23	5.52	8.76	1.29	1.03	3.29	0.69	4.38
	GDC-0810	0.04	0.59	0.55	0.52	0.04	0.06	0.24	0.42	0.07	2.45	2.52	0.63	0.35	1.31	0.23	1.30
	RAD1901	0.46	17.65	8.17	10.21	0.56	0.68	2.43	6.68	0.97	26.09	22.69	10.94	4.54	14.41	4.38	17.52

B

	Fold Mutant clones vs WT	MCF-7 Luc	L536R CL1	L536R CL2	L536R CL3	Y537C CL1	Y537C CL2	Y537N CL1	Y537N CL2	Y537N CL3	Y537S CL1	Y537S CL2	Y537S CL3	D538G CL1	D538G CL2	D538G CL3	D538G CL4
SERM	Tamoxifen	1.0	13.5	6.5	10.9	1.1	1.2	3.1	6.1	2.7	84.4	26.5	54.4	4.5	22.1	10.4	23.8
	Raloxifene	1.0	28.8	15.6	15.4	1.4	1.4	5.2	15.4	2.2	182.0	49.1	19.5	7.9	25.9	7.5	25.6
	Bazedoxifene	1.0	13.7	10.1	9.3	0.9	1.1	3.9	17.8	2.4	105.9	88.7	27.6	6.7	29.1	8.3	22.4
	Lasofoxifene	1.0	17.7	9.3	14.7	1.0	1.6	4.5	10.4	3.8	75.7	22.0	29.2	7.9	12.5	7.1	24.4
SERD	Faslodex	1.0	5.1	5.1	5.6	1.0	1.5	4.8	10.6	1.7	39.8	59.4	11.1	6.2	33.9	4.1	26.6
	AZD9496	1.0	4.8	4.0	3.9	0.9	1.7	4.6	16.0	2.0	148.0	113.8	17.2	10.7	16.3	6.9	22.4
	AZD9833	1.0	67.9	26.1	32.4	1.2	1.3	6.3	11.2	2.1	48.5	77.1	11.3	9.1	28.9	6.0	38.5
	GDC-0810	1.0	14.6	13.6	12.8	0.9	1.4	5.9	10.2	1.8	60.3	62.0	15.5	8.7	32.2	5.7	32.1
	RAD1901	1.0	38.1	17.7	22.1	1.2	1.5	5.2	14.4	2.1	56.4	49.0	23.6	9.8	31.1	9.5	37.9

Figure B.3: Effect of chemotherapeutics on ERα cancer mutants. A) and B) IC50 values (nM) generated for a variety of mutant clones, together with the fold difference in IC50 between WT (MCF7) cells and mutant clones. Values highlighted in orange show >10-fold difference from the IC50 value determined for MCF7 cells, with the cells in red identifying >20-fold difference in sensitivity to drug. Experimental data and figure provided by Fui Lai and Simak Ali.

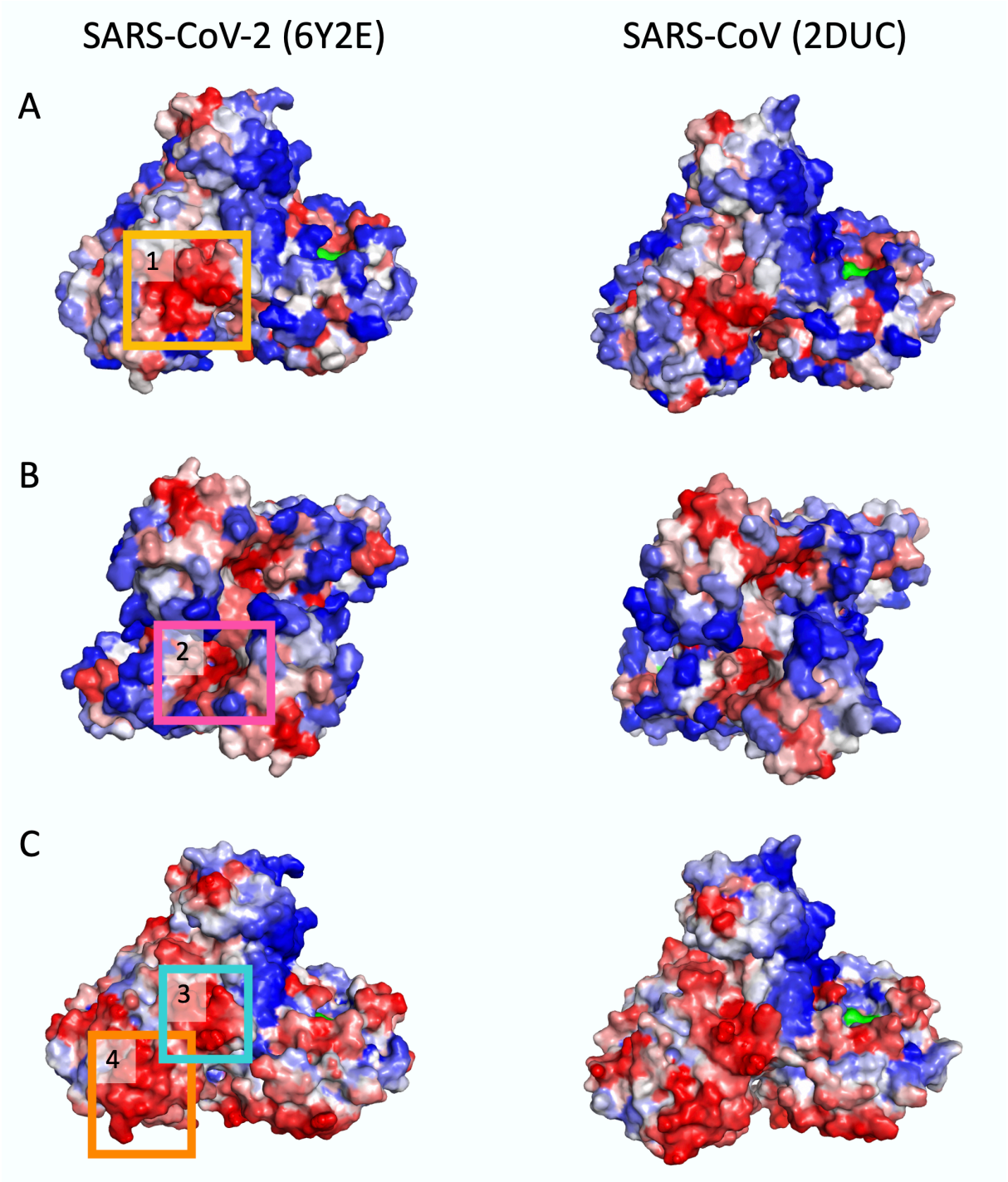


Figure B.4: Consistency of allosteric hotspots between SARS-CoV-2 and SARS-CoV. Surface representations of the SARS-CoV-2 (PDB id: 6Y2E[5]) and SARS-CoV (PDB id: 2DUC[277]). Coloured by QS from BBP analysis (for **A, B**) and MT analysis (for **C**) sourced from active site residues. **A)** Hotspot 1 as described in Figure 5.5A. **B)** Hotspot 2 as described in Figure 5.5B. **C)** Hotspots 3 and 4 as described in Figure 5.6C. Adapted from Strömich et al.[55].

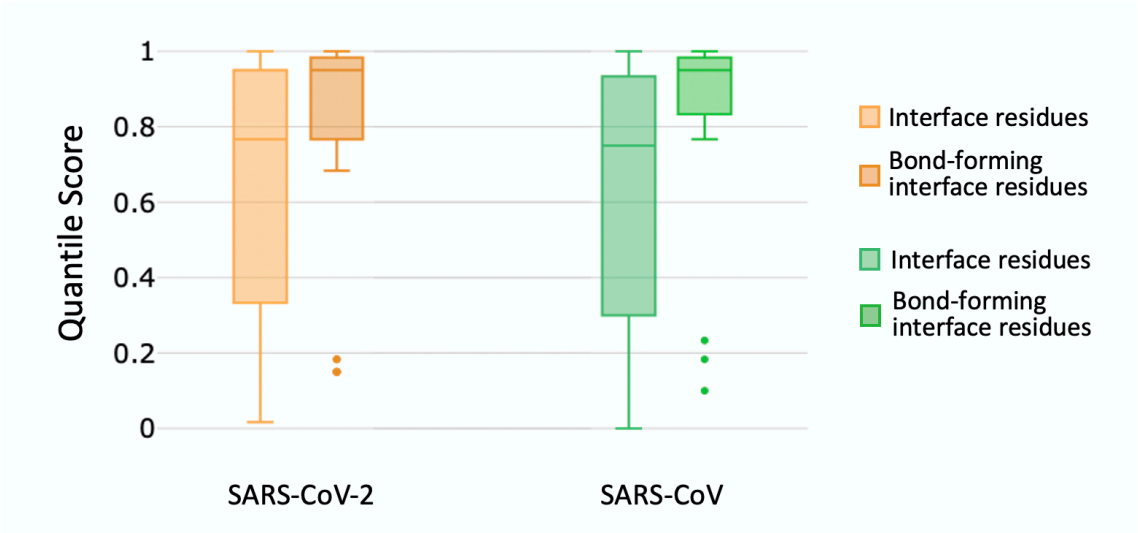


Figure B.5: Scoring of whole dimer interface in SARS-CoV-2 and SARS-CoV. Shown are QS distributions for the interface residues in SARS-CoV-2 and SARS-CoV M^{pro}. Bond forming interface residues are a subclass of interface residues, which from hydrogen bonds and salt bridges.

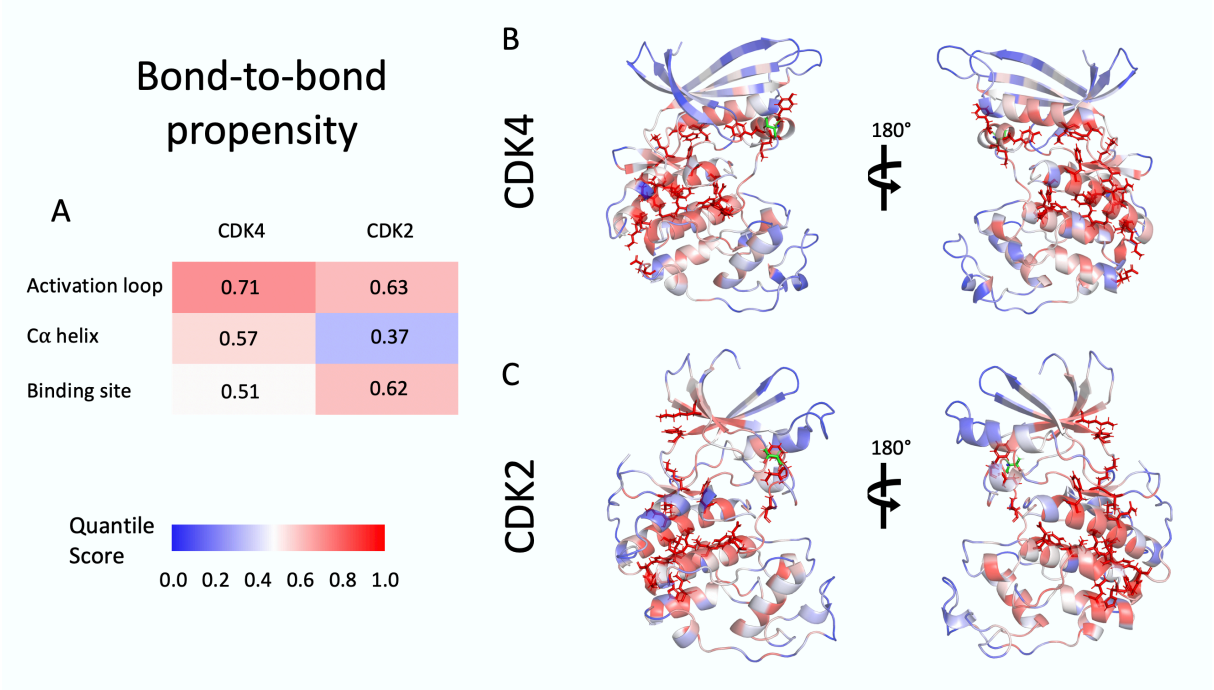


Figure B.6: Bond-to-bond propensity analysis of monomeric CDK4 and 2 when sourced from the phosphorylation site. A) Average quantile score (QS) results for bond-to-bond propensities are shown for each structural element in CDK4 and CDK2. **B) and C)** The structures of CDK4 (AlphaFold model[49]) and CDK2 (PDB id: 1HCL[346]) are shown in two orientations with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks.

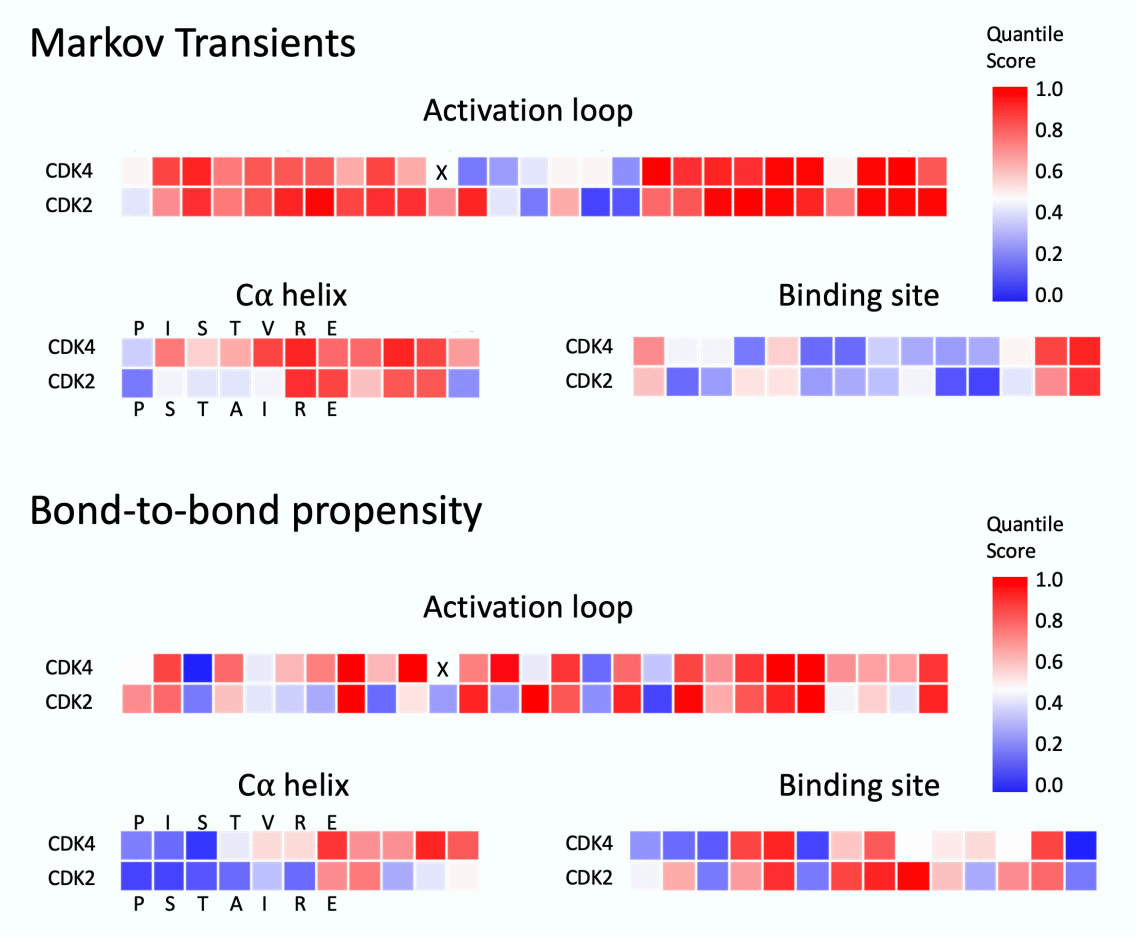


Figure B.7: Residue-wise MT and BBP results in CDK2 and 4 coloured by quantile score when sourced from the phosphorylation site. Shown are structural features in the kinases with one box per residue (for a full list of residues in these structural features see Tbl. C.7). PISTVRE/PSTAIRE helix residues are indicated. X - residue not present.

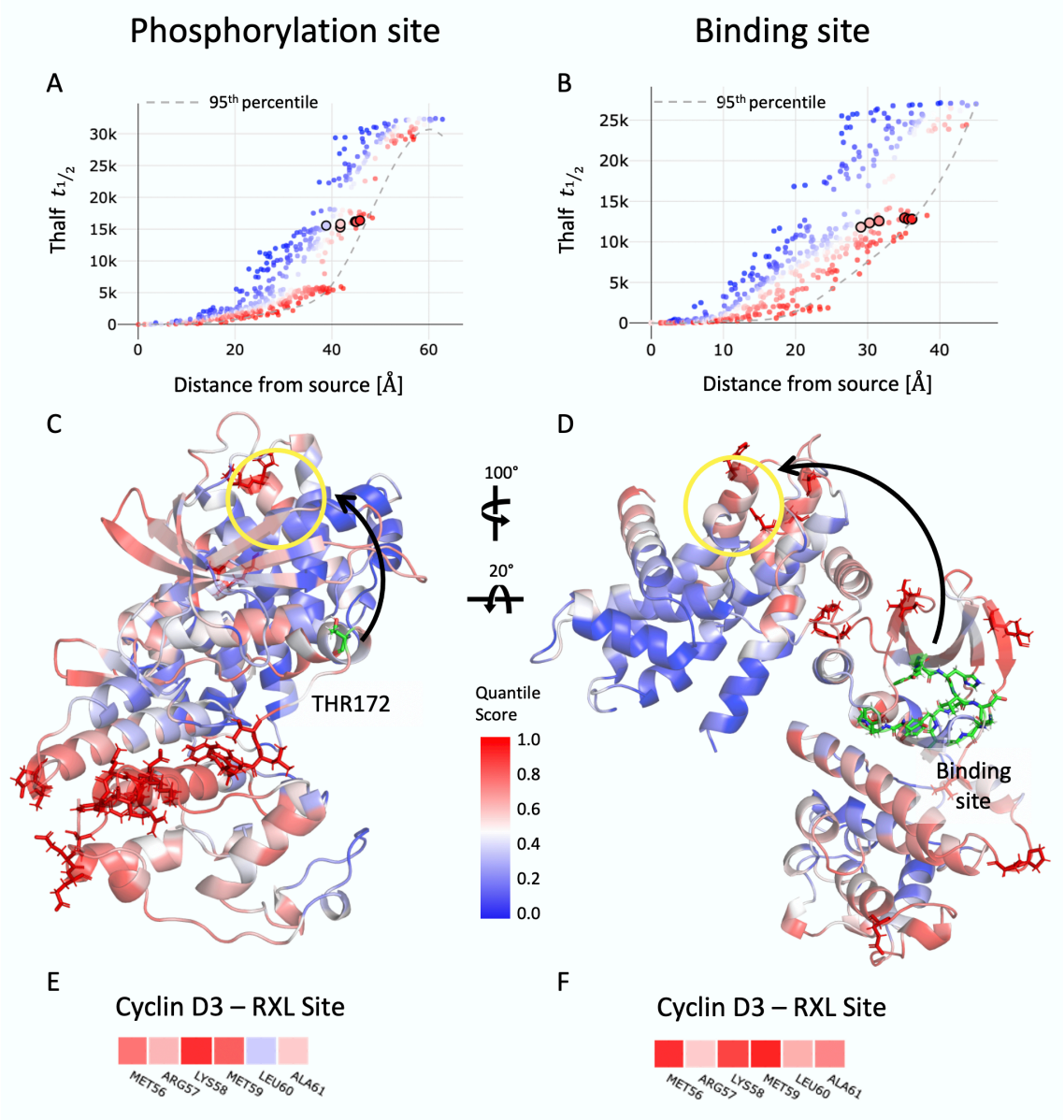


Figure B.8: Markov transient analysis of CDK4 - cyclin D3. Shown on the right are the results of the analysis sourced from the phosphorylation site THR^{172} and on the left when sourced from the binding site residues. Colours are according to QS from 0 - blue to 1 - red. **A) and B)** Data distribution of all residues with $t_{1/2}$ values over the distance from the source. RXL site residues are highlighted as larger dots with a black outline. **C) and D)** The complex (PDB id: 3G33[329]) is shown in two orientations with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a yellow circle are the RXL sites on cyclin D3. **E) and F)** Detailed sequence for the RXL site residues coloured by QS.

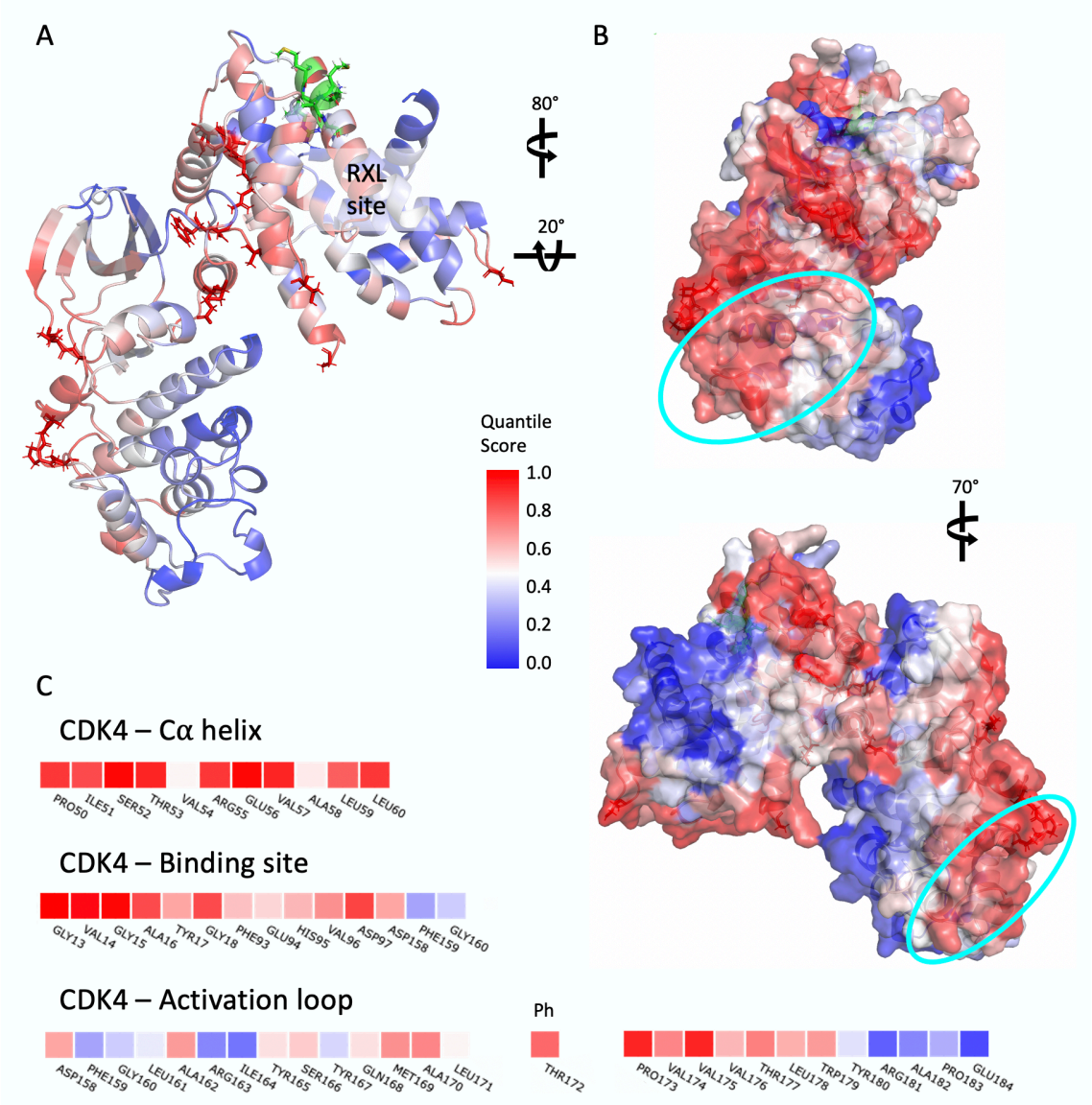


Figure B.9: The RXL site as a source in Markov transient analysis of CDK4 - cyclin D3. A) The complex (PDB id: 3G33[329]) is shown in the front orientation with residues coloured by QS (0 - blue to 1 - red). Residues with a QS > 0.95 are shown as sticks, and source residues are shown as green sticks. **B)** Two surface visualisations of the complex analogous to the results presented in Fig. 6.8. Highlighted in cyan is an extended hotspot region that we propose to be a PPI site. **C)** Detailed sequences for functionally important features on the kinase coloured by QS. Ph - phosphorylation site.

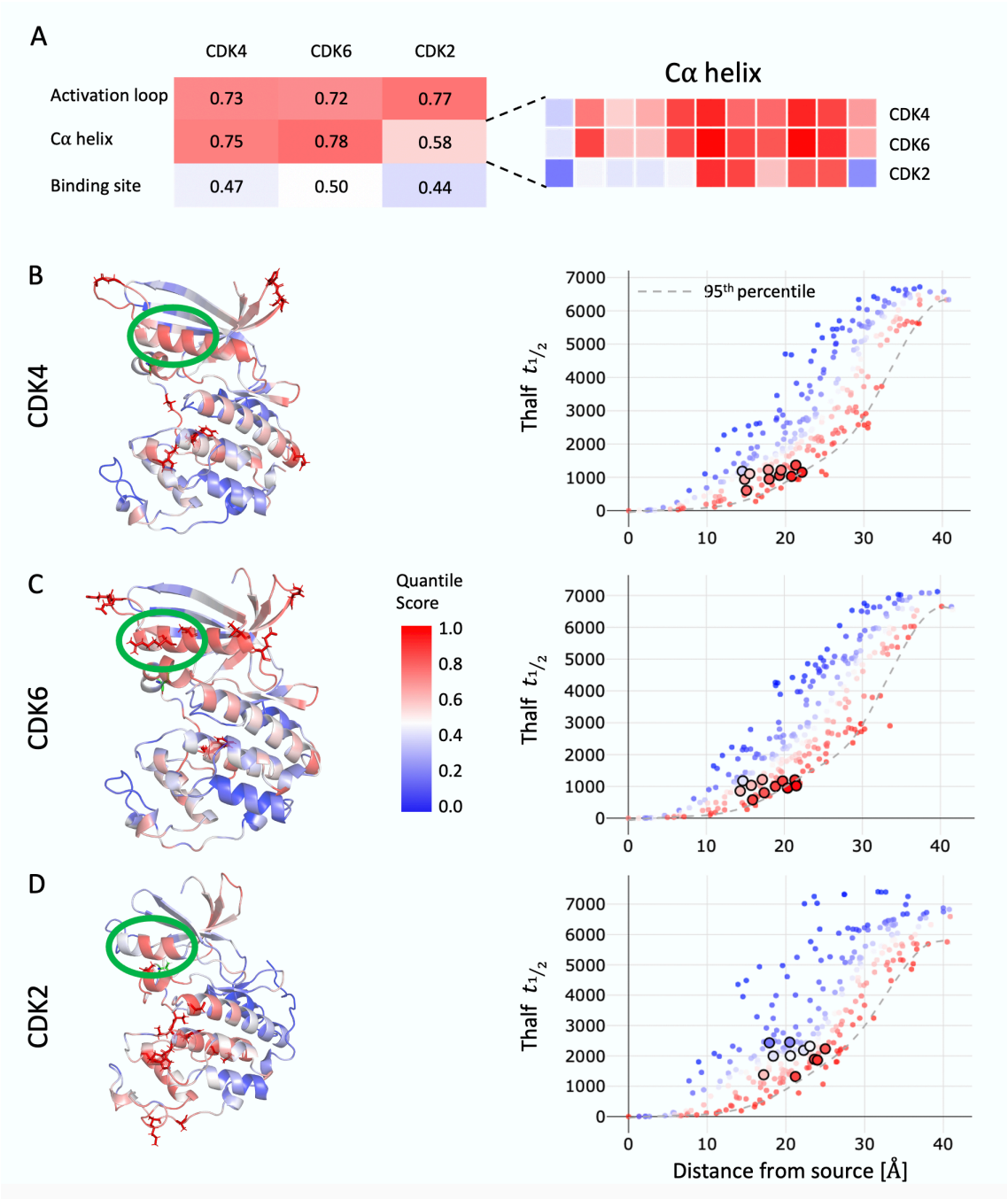


Figure B.10: MT analysis of monomeric CDK structures when sourced from the phosphorylation site. A) Average quantile score (QS) results for Markov Transients are shown for each structural element in CDKs 4, 6 and 2. A zoom in into the sequence of the Cα helix is provided to the right. **B) and C) and D)** The structures of CDK4, CDK6 (both AlphaFold models[49]) and CDK2 (PDB id: 1HCL[346]) are shown in back orientation with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a green circle are the Cα helices. The scatterplots show $t_{1/2}$ values over the distance from the source for each residue in the protein. Cα residues are highlighted as larger dots with a black outline.

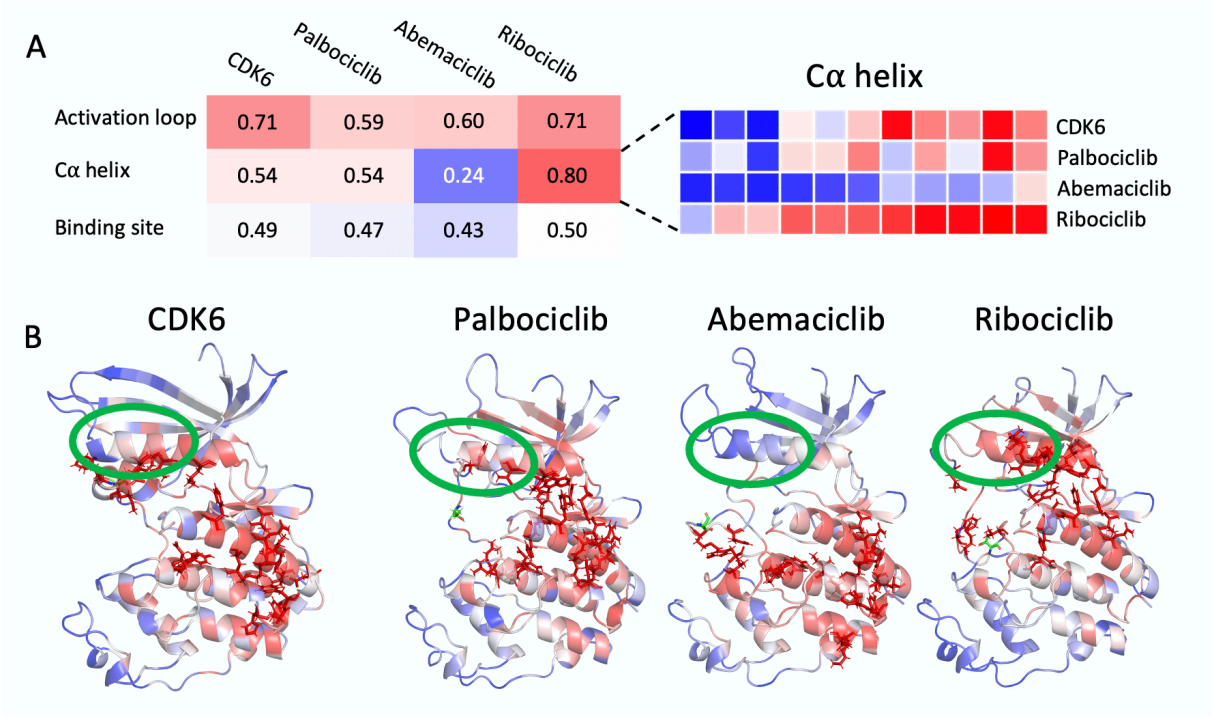


Figure B.11: BBP analysis of monomeric CDK6 in apo and inhibited form when sourced from the phosphorylation site. A) Average quantile score (QS) results for bond-to-bond propensities are shown for each structural element in CDK6 in apo and inhibited forms. A zoom-in into the sequence of the Cα helix is provided to the right. **B)** The structures of monomeric CDK6 (AlphaFold model[49]) and inhibited forms with palbociclib, abemaciclib and ribociclib (PDB ids: 5L2I, 5L2S, 5L2T[354]) are shown from the back view with residues coloured by QS. Residues with a QS > 0.95 are shown as sticks. The source residues are shown as green sticks. Highlighted with a green circle are the Cα helices.

Appendix C

Supplementary Tables

Table C.1: Dimer interface residues in the agonist-bound ERα LBD. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

Chain B	Bond	Chain C	Bond
CYS381		CYS381	
GLU385		GLU385	
MET427		GLU423	
ALA430		MET427	
THR431		ALA430	
SER433		THR431	
ARG434	H	SER433	H
MET437		ARG434	
ILE451		MET437	
ASN455	H	ILE451	
SER456		ASN455	H
GLY457		SER456	
TYR459		GLY457	
THR460		TYR459	H
LEU469		THR460	

Table C.1 continued from previous page

Chain B	Bond	Chain C	Bond
LYS472	H	LEU469	
ASP473		LYS472	
HIS476		ASP473	
LEU479		HIS476	
ASP480	H	LEU479	
LYS481		ASP480	H
THR483		LYS481	
ASP484	H	THR483	
ILE487		ASP484	H
LEU497		ILE487	
GLN498	H	LEU497	
GLN500		GLN498	
HIS501		GLN500	
GLN502	H	HIS501	
LEU504		GLN502	H
ALA505		LEU504	
GLN506	H	ALA505	
LEU508		GLN506	H
LEU509	H	LEU508	
ILE510		LEU509	H
LEU511		ILE510	
SER512	H	LEU511	
HIS513		SER512	H
ARG515	H	HIS513	
HIS516		ARG515	H
ASN519		HIS516	

Table C.1 continued from previous page

Chain B	Bond	Chain C	Bond
LYS520		ASN519	
MET522		LYS520	
GLU523		MET522	
HIS547		GLU523	
ARG548		HIS547	
LEU549			

Table C.2: Dimer interface residues in the antagonist-bound ERα LBD. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

Chain A	Bond	Chain B	Bond
GLU385		GLU385	
MET427		MET427	
ALA430		ALA430	
THR431		THR431	
ARG434		ARG434	
MET437		MET437	
ILE451		ILE451	
ASN455	H	ASN455	H
SER456		SER456	
TYR459		TYR459	
THR460		THR460	
LEU469		LEU469	
LYS472		LYS472	
HIS476		HIS476	
LEU479		LEU479	
ASP480	H	ASP480	H

Table C.2 continued from previous page

Chain A	Bond	Chain B	Bond
LYS481		LYS481	
THR483		THR483	
ASP484	H	ASP484	H
ILE487		ILE487	
LEU497		LEU497	
GLN498	H	GLN498	H
HIS501		HIS501	
GLN502	H	GLN502	H
LEU504		LEU504	
ALA505		ALA505	
GLN506	H	GLN506	H
LEU508		LEU508	
LEU509	H	LEU509	H
ILE510		ILE510	
LEU511		LEU511	
SER512		SER512	
HIS513		HIS513	
ARG515		ARG515	
HIS516	H	HIS516	H
ASN519	H	ASN519	H
LYS520		LYS520	
GLU523		GLU523	

Table C.3: Dimer interface residues in the SARS-CoV-2 M^{pro}. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

Chain A	Bond	Chain B	Bond
SER1	HS	SER1	HS
GLY2		GLY2	
PHE3		PHE3	
ARG4	HS	ARG4	HS
LYS5		LYS5	
MET6		MET6	
ALA7	H	ALA7	H
PHE8		PHE8	
PRO9		PRO9	
SER10	H	SER10	H
GLY11	H	GLY11	H
LYS12		LYS12	
GLU14	H	GLU14	H
MET17		MET17	
GLY71		GLY71	
LEU115		LEU115	
ALA116		ALA116	
TYR118	H	TYR118	H
ASN119		ASN119	
GLY120		GLY120	
SER121	H	SER121	H
PRO122	H	PRO122	H
SER123		SER123	
GLY124		GLY124	
VAL125	H	VAL125	H

Table C.3 continued from previous page

Chain A	Bond	Chain B	Bond
TYR126		TYR126	
GLN127	H	GLN127	H
CYS128		CYS128	
ALA129		ALA129	
LYS137	H	LYS137	H
GLY138		GLY138	
SER139	H	SER139	H
PHE140	H	PHE140	H
LEU141		LEU141	
GLU166	HS	GLU166	HS
GLY170		GLY170	
HIS172		HIS172	
THR280		THR280	
GLY283		GLY283	
SER284		SER284	
ALA285		ALA285	
LEU286		LEU286	
GLU290	HS	GLU290	HS
ARG298		ARG298	
GLN299	H	GLN299	H
CYS300		CYS300	
SER301		SER301	
GLY302		GLY302	
VAL303		VAL303	
THR304	H	THR304	H
PHE305	H	PHE305	H

Table C.3 continued from previous page

Chain A	Bond	Chain B	Bond
GLN306	H	GLN306	H

Table C.4: Dimer interface residues in the SARS-CoV M^{pro}. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

Chain A	Bond	Chain B	Bond
SER1		SER1	HS
GLY2		GLY2	
PHE3		PHE3	
ARG4	H	ARG4	HS
LYS5		LYS5	
MET6		MET6	
ALA7	H	ALA7	H
PHE8		PHE8	
PRO9		PRO9	
SER10	H	SER10	H
GLY11	H	GLY11	H
LYS12		LYS12	
GLU14	H	GLU14	H
LEU115		ASN72	
ALA116		LEU115	
SER121		ALA116	
PRO122		TYR118	
SER123	H	SER121	
GLY124		PRO122	H
VAL125	H	SER123	
TYR126		GLY124	

Table C.4 continued from previous page

Chain A	Bond	Chain B	Bond
GLN127		VAL125	H
CYS128		TYR126	
LYS137	H	GLN127	H
GLY138		CYS128	
SER139		ILE136	
PHE140	H	GLY138	
LEU141		SER139	
GLU166	HS	PHE140	
PRO168		LEU141	
GLY170		ASN142	
HIS172		ASP155	
THR285		GLU166	
ILE286		THR285	
GLU290	HS	ILE286	
GLN299		GLU290	
GLY302		ARG298	H
VAL303		GLN299	
THR304		CYS300	
PHE305	H	SER301	
GLN306		PHE305	

Table C.5: Allosteric hotspots in the SARS-CoV M^{pro} as determined with BBP analysis. QSs are given for each residue and solvent-accessible surface area (SASA) was determined in PyMol[191].

Hotspot	Residue	QS	SASA [Å^2]
Hotspot 1	LEU30	0.95	22.89
	LEU32	1.00	1.35
	ASP33	0.89	69.14
	ASN95	0.95	6.90
	THR98	0.98	154.30
	LYS100	0.97	156.31
	TYR101	0.98	69.11
	PHE103	0.98	60.19
	PHE159	1.00	0.00
Hotspot 2	ARG4	0.54	89.49
	ARG131	0.99	7.96
	ASP197	0.91	36.91
	THR199	0.76	32.12
	ASP289	0.98	15.28
	GLU290	0.83	12.88

Table C.6: Allosteric hotspots in the SARS-CoV M^{pro} as determined with MT analysis. QSs are given for each residue and solvent-accessible surface area (SASA) was determined in PyMol[191]. Highlighted in blue is a cysteine residue that can be targeted for covalent binding.

Hotspot	Residue	QS	SASA [Å^2]
Hotspot 3	LYS100	0.91	156.31
	LYS102	0.78	136.75
	ASN151	0.98	30.43
	ILE152	0.93	7.47
	ASP153	0.99	62.00
	TYR154	0.83	192.09
	ASP155	0.85	54.97
	CYS156	0.98	17.69
	VAL157	0.71	0.00
	SER158	0.89	18.84
Hotspot 4	ASP33	0.94	69.14
	ASP34	0.96	53.76
	VAL35	0.92	18.11
	TYR37	0.91	8.60
	ARG76	0.84	156.72
	ILE78	0.87	69.36
	LYS90	0.64	110.37
	VAL91	0.63	1.34
	ASP92	0.95	72.23
	THR93	0.89	66.96
	ALA94	0.91	67.31

Table C.7: Alignment of structural features in CDKs. Residues are listed with residue name and number for each structural feature. Orange - DFG motif (in binding site and activation loop); green - PSTAIRE motif; blue - phosphorylation site; X - residue not present

Structural feature	CDK2	CDK4	CDK6
Binding site	GLY 11	GLY 13	GLY 20
	GLU 12	VAL 14	GLU 21
	GLY 13	GLY 15	GLY 22
	THR 14	ALA 16	ALA 23
	TYR 15	TYR 17	TYR 24
	GLY 16	GLY 18	GLY 25
	PHE 80	PHE 93	PHE 98
	GLU 81	GLU 94	GLU 99
	PHE 82	HIS 95	HIS 100
	LEU 83	VAL 96	VAL 101
	HIS 84	ASP 97	ASP 102
	ASP 145	ASP 158	ASP 163
	PHE 146	PHE 159	PHE 164
	GLY 147	GLY 160	GLY 165
Cα helix	PRO 45	PRO 50	PRO 55
	SER 46	ILE 51	LEU 56
	THR 47	SER 52	SER 57
	ALA 48	THR 53	THR 58
	ILE 49	VAL 54	ILE 59
	ARG 50	ARG 55	ARG 60
	GLU 51	GLU 56	GLU 61
	ILE 52	VAL 57	VAL 62
	SER 53	ALA 58	ALA 63
	LEU 54	LEU 59	VAL 64

Table C.7 continued from previous page

Structural feature	CDK2	CDK4	CDK6
	LEU 55	LEU 59	LEU 65
Activation loop	ASP 145	ASP 158	ASP 163
	PHE 146	PHE 159	PHE 164
	GLY 147	GLY 160	GLY 165
	LEU 148	LEU 161	LEU 166
	ALA 149	ALA 162	ALA 167
	ARG 150	ARG 163	ARG 168
	ALA 151	ILE 164	ILE 169
	PHE 152	TYR 165	TYR 170
	GLY 153	SER 166	SER 171
	VAl 154	TYR 167	PHE 172
	PRO 155	X	X
	VAL 156	GLN 168	GLN 173
	ARG 157	MET 169	MET 174
	THR 158	ALA 170	ALA 175
	TYR 159	LEU 171	LEU 176
	THR 160	THR 172	THR 177
	HIS 161	PRO 173	SER 178
	GLU 162	VAL 174	VAL 179
	VAL 163	VAL 175	VAL 180
	VAL 164	VAL 176	VAL 181
	THR 165	THR 177	THR 182
	LEU 166	LEU 178	LEU 183
	TRP 167	TRP 179	TRP 184
	TYR 168	TYR 180	TYR 185
	ARG 169	ARG 181	ARG 186

Table C.7 continued from previous page

Structural feature	CDK2	CDK4	CDK6
	ALA 170	ALA 182	ALA 187
	PRO 171	PRO 183	PRO 188
	GLU 172	GLU 184	GLU 189

Table C.8: Dimer interface residues between CDK4 and cyclin D1. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

CDK4	Bond	Cyclin D1	Bond
ARG5		LEU23	
ASN41		ARG26	
GLY42		VAL27	
GLU43		ALA30	
GLU44	H	MET31	
GLY48		LYS33	HS
LEU49	H	ALA34	
ILE51		THR37	
THR53		PHE108	
VAL54		VAL109	
ARG55	H	LYS112	H
VAL57		MET113	
ALA58		LYS114	
LEU59		GLU115	
ARG61	H	THR116	
ARG62	H	PRO118	
GLU64		LEU119	
ALA65	H	THR120	
PHE66		ALA121	H

Table C.8 continued from previous page

CDK4	Bond	Cyclin D1	Bond
GLU67	HS	PRO134	
MET75		GLU135	
ASP76		LEU138	
VAL77		GLN139	
CYS78		GLU141	H
ALA79	H	LEU142	
SER81		VAL145	
ARG82		ASN146	H
ILE87		LYS149	H
VAL89		TRP150	
PHE130		ASN151	
ALA133		LEU152	
		ALA153	H
		ALA154	
		MET155	
		ASP159	

Table C.9: Dimer interface residues between CDK4 and cyclin D3. Residues determined by PDBePisa[78]. H - hydrogen bond, HS - salt bridge

CDK4	Bond	Cyclin D3	Bond
VAL39		ARG26	
ASN41		SER30	
GLY42		LEU31	
GLY43		ARG33	
GLY45		LEU34	
GLY46		ARG37	HS

Table C.9 continued from previous page

CDK4	Bond	Cyclin D3	Bond
GLY47		TYR38	
GLY48		LEU109	
LEU49		LYS112	H
ILE51		LEU113	
THR53		ARG114	
VAL54		GLU115	H
ARG55	H	THR116	
VAL57		PRO118	
ALA58		THR120	
LEU59		ILE121	
ARG61	H	GLU122	
ARG62		PRO134	
GLU64	HS	ARG138	
ALA65		ASP139	
PHE66		GLU141	
MET75		VAL142	
ASP76		LEU145	
VAL77		GLY146	
CYS78		LYS149	H
ALA79		TRP150	
THR80		ASP151	
SER81		LEU152	
ILE87		ALA153	
VAL89		ALA154	
PHE287		VAL155	
		ASP159	

Table C.10: Average QS of CDK4 - cyclin D3 interface when sourced from different sites. Phosphorylation site is THR^{172} and a full list of binding site residues is given in Table C.7. RXL site residues are listed in Table 6.1.

	Source site		
Methodology	Phosphorylation site	Binding site	RXL site
Markov Transients	0.40	0.53	0.55
Random Site Score [95% CI]	0.46, [0.46,0.47]	0.47, [0.47,0.48]	0.50, [0.50,0.51]
Bond-to-bond propensity	0.44	0.52	0.54
Random Site Score [95% CI]	0.53, [0.52,0.53]	0.49, [0.48,0.49]	0.52, [0.52,0.53]

Appendix D

Publication Permissions of Third Parties

ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules

Author: Mersmann, Sophia F; Strömich, Léonie

Publication: Nucleic Acids Research

Publisher: Oxford University Press

Date: 2021-05-12

Spectrum and Degree of CDK Drug Interactions Predicts Clinical Performance

Author:
Ping Chen,Nathan V. Lee,Wenyue Hu,Meirong Xu,Rose Ann Ferre,Hieu Lam,Simon Bergqvist,James Solowiej,Wade Diehl,You-Ai He,Xiu Yu,Asako Nagata,Todd VanArsdale,Brion W. Murray

Publication: Molecular Cancer Therapeutics

Publisher: American Association for Cancer Research

Date: 2016-10-01

Order Completed

Thank you for your order.

This Agreement between Léonie Strömich ("You") and American Association for Cancer Research ("American Association for Cancer Research") consists of your license details and the terms and conditions provided by American Association for Cancer Research and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	5238840596662
License date	Jan 30, 2022

Printable Details

Licensed Content

Licensed Content Publisher	American Association for Cancer Research
Licensed Content Publication	Molecular Cancer Therapeutics
Licensed Content Title	Spectrum and Degree of CDK Drug Interactions Predicts Clinical Performance
Licensed Content Author	Ping Chen,Nathan V. Lee,Wenyue Hu,Meirong Xu,Rose Ann Ferre,Hieu Lam,Simon Bergqvist,James Solowiej,Wade Diehl,You-Ai He,Xiu Yu,Asako Nagata,Todd VanArsdale,Brion W. Murray
Licensed Content Date	Oct 1, 2016
Licensed Content Volume	15
Licensed Content Issue	10

Order Details

Type of Use	Thesis/Dissertation
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no
Circulation	20
Territory of distribution	Worldwide

About Your Work

Title	Atomistic graph analysis of protein dimers in disease
Institution name	Imperial College London
Expected presentation date	Apr 2022

Additional Data

Portions	Figure 4

Synthesis and Biological Evaluation of 1-Aryl-4,5-dihydro-1H-pyrazolo[3,4-d]pyrimidin-4-one Inhibitors of Cyclin-Dependent Kinases

Author: Jay A. Markwalder, Marc R. Arnone, Pamela A. Benfield, et al

Publication: Journal of Medicinal Chemistry

Publisher: American Chemical Society

Date: Nov 1, 2004

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

BACK CLOSE WINDOW

Bibliography

[1] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. W.W. Norton & Company, 6th edition, 2015.

[2] Yaliraki, S.N. and Barahona, M. Chemistry across scales: From molecules to cells. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1861):2921–2934, 2007.

[3] Yasar, P., Ayaz, G., User, S.D., Güpür, G., and Muyan, M. Molecular mechanism of estrogen-estrogen receptor signaling. *Reproductive Medicine and Biology*, 16(1):4–20, 2017.

[4] Punekar, N.S. *ENZYMES: Catalysis, Kinetics and Mechanisms*. Springer Singapore, Singapore, 1st edition, 2018.

[5] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[6] Morgan, D.O. *The Cell Cycle, Principles of Control*. New Science Press Ltd., London, 1st edition, 2007.

[7] Patel, M.S., Nemeria, N.S., Furey, W., and Jordan, F. The pyruvate dehydrogenase complexes: Structure-based function and regulation. *Journal of Biological Chemistry*, 289(24):16615–16623, 2014.

[8] Weis, W.I. and Kobilka, B.K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annual Review of Biochemistry*, 87:897–919, 2018.

[9] Turcios, N.L. Cystic fibrosis: An overview. *Journal of Clinical Gastroenterology*, 39(4): 307–317, 2005.

[10] Ryan, D.P. and Matthews, J.M. Protein-protein interactions in human disease. *Current Opinion in Structural Biology*, 15(4):441–446, 2005.

[11] Araujo, R.P., Liotta, L.A., and Petricoin, E.F. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. *Nature Reviews Drug Discovery*, 6:878–880, 2007.

[12] Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montaño, B., Blundell, T.L., and Ascher, D.B. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in Biophysics and Molecular Biology*, 128: 3–13, 2017.

[13] Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, G., Karlsson, A., Al-lazikani, B., Hersey, A., Oprea, T.I., and Overington, J.P. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017.

[14] Grover, A.K. Use of Allosteric Targets in the Discovery of Safer Drugs. *Medical Principles and Practice*, 22(5):418–426, 2013.

[15] Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., and Zarrinkar, P.P. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011.

[16] Bridges, T.M. and Lindsley, C.W. G-protein-coupled receptors: From classical modes of modulation to allosteric mechanisms. *ACS Chemical Biology*, 3(9):530–541, 2008.

[17] Wenthur, C.J., Gentry, P.R., Mathews, T.P., and Lindsley, C.W. Drugs for Allosteric Sites on Receptors. *Annual Review of Pharmacology and Toxicology*, 54(1):165–184, 2014.

[18] Aranda, A. and Pascual, A. Nuclear Hormone Receptors and Gene Expression. *Physiological Reviews*, 81(3):1269–1304, 2001.

[19] Wood, D.J. and Endicott, J.A. Structural insights into the functional diversity of the CDK–cyclin family. *Open Biology*, 8(9):180112, 2018.

[20] Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. Dissecting Subunit Interfaces in Homodimeric Proteins. *Proteins: Structure, Function and Genetics*, 53(3):708–719, 2003.

[21] Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., and Shi, J. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy*, 5(1), 2020.

[22] Fry, D.C. Protein-Protein Interactions as Targets for Small Moelcule Drug Discovery. *Biopolymers*, 84:535–552, 2006.

[23] Wells, J.A. and McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009, 2007.

[24] Rosell, M. and Fernández-Recio, J. Hot-spot analysis for drug discovery targeting protein-protein interactions. *Expert Opinion on Drug Discovery*, 13(4):327–338, 2018.

[25] Helmer, D. and Schmitz, K. Peptides and Peptide Analogs to Inhibit Protein-Protein Interactions. In Böldicke, T., editor, *Protein Targeting Compounds: Advances in Experimental Medicine and Biology.*, pages 147–183. Springer, Cham, 2016.

[26] Wang, X., Ni, D., Liu, Y., and Lu, S. Rational Design of Peptide-Based Inhibitors Disrupting Protein-Protein Interactions. *Frontiers in Chemistry*, 9(May):1–15, 2021.

[27] Monod, J. and Jacob, F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:389–401, 1961.

[28] Monod, J., Wyman, J., and Changeux, J.P. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12(1):88–118, 1965.

[29] Koshland, D.E., Némethy, G., and Filmer, D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry*, 5(1):365–385, 1966.

[30] Cooper, A. and Dryden, D.T. Allostery without conformational change. A plausible model. *European Biophysics Journal*, 11:103–109, 1984.

[31] Kumar, S., Ma, B., Tsai, C.J., Wolfson, H., and Nussinov, R. Folding funnels and conformational transitions via hinge-bending motions. *Cell Biochemistry and Biophysics*, 31(2):141–164, 1999.

[32] Ma, B., Kumar, S., Tsai, C.J., and Nussinov, R. Folding funnels and binding mechanisms. *Protein engineering*, 12(9):713–720, 1999.

[33] Tsai, C.J., Kumar, S., Ma, B., and Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190, 1999.

[34] Lockless, S.W. and Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[35] del Sol, A., Tsai, C.J., Ma, B., and Nussinov, R. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, 17(8):1042–1050, 2009.

[36] Hilser, V.J., Wrabl, J.O., and Motlagh, H.N. Structural and Energetic Basis of Allostery. *Annual Review of Biophysics*, 41:585–609, 2012.

[37] Motlagh, H.N., Wrabl, J.O., Li, J., and Hilser, V.J. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

[38] Tsai, C.J. and Nussinov, R. A Unified View of "How Allostery Works". *PLoS Computational Biology*, 10(2), 2014.

[39] Gunasekaran, K., Ma, B., and Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function and Genetics*, 57(3):433–443, 2004.

[40] Ahmed, M., Ghatge, M., and Safo, M. Hemoglobin: Structure, Function and Allostery. In Hoeger, U. and Harris, J., editors, *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins.*, pages 345–382. Springer, Cham, subcellula edition, 2020.

[41] Kantrowitz, E.R. Allostery and cooperativity in Escherichia coli aspartate transcarbamoylase. *Archives of Biochemistry and Biophysics*, 519(2):81–90, 2012.

[42] Hardy, J.A. and Wells, J.A. Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology*, 14(6):706–715, 2004.

[43] Ni, D., Lu, S., and Zhang, J. Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug discovery. *Medicinal Research Reviews*, 39:2314–2342, 2019.

[44] Mannes, M., Martin, C., Menet, C., and Ballet, S. Wandering beyond small molecules: peptides as allosteric protein modulators. *Trends in Pharmacological Sciences*, In Press., 2021.

[45] Hughes, J.P., Rees, S.S., Kalindjian, S.B., and Philpott, K.L. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011.

[46] Agamah, F.E., Mazandu, G.K., Hassan, R., Bope, C.D., Thomford, N.E., Ghansah, A., and Chimusa, E.R. Computational/in silico methods in drug target and lead prediction. *Briefings in Bioinformatics*, 21(5):1663–1675, 2020.

[47] Nwanochie, E. and Uversky, V.N. Structure determination by single-particle cryo-electron microscopy: Only the sky (and intrinsic disorder) is the limit. *International Journal of Molecular Sciences*, 20(17):4186, 2019.

[48] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[49] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[50] Frye, L., Bhat, S., Akinsanya, K., and Abel, R. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discovery Today: Technologies*, 39:111–117, 2021.

[51] Mersmann, S., Strömich, L., Song, F.J., Wu, N., Vianello, F., Barahona, M., and Yaliraki, S. ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules. *Nucleic Acids Research*, 2021.

[52] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71: 209–249, 2021.

[53] Gao, X., Leone, G.W., and Wang, H. Cyclin D-CDK4/6 functions in cancer. In *Advances in Cancer Research*, volume 148, pages 147–169. Elsevier Inc., 2020.

[54] Wu, N., Strömich, L., and Yaliraki, S.N. Prediction of allosteric sites and signaling: Insights from benchmarking datasets. *Patterns*, 3(1):100408, 2021.

[55] Strömich, L., Wu, N., Barahona, M., and Yaliraki, S.N. Allosteric Hotspots in the Main Protease of SARS-CoV-2. *bioRxiv*, page 2020.11.06.369439, 2020.

[56] Torjesen, I. Drug development: the journey of a medicine from lab to shelf. The Pharmaceutical Journal, 2015. Available at: https://pharmaceutical-journal.com/article/feature/drug-development-the-journey-of-a-medicine-from-lab-to-shelf.

[57] Macalino, S.J.Y., Gosu, V., Hong, S., and Choi, S. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, 38(9):1686–1701, 2015.

[58] Yu, W. and MacKerell Jr., A. Computer-Aided Drug Design Methods. *Methods in Molecular Biology*, 1520:85–106, 2017.

[59] Perkins, R., Fang, H., Tong, W., and Welsh, W.J. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry*, 22(8):1666–1679, 2003.

[60] Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004.

[61] Chen, Y.C. Beware of docking! *Trends in Pharmacological Sciences*, 36(2):78–95, 2015.

[62] Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Ojeda-Montes, M.J., Gimeno, A., Cereto-Massagué, A., Garcia-Vallvé, S., and Pujadas, G. Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Medicinal Research Reviews*, (In Press), 2021.

[63] De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, 2016.

[64] Curry, S. Structural biology: A century-long journey into an unseen world. *Interdisciplinary Science Reviews*, 40(3):308–328, 2015.

[65] Bai, X., McMullan, G., and Scheres, S.H. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1):49–57, 2015.

[66] AlQuraishi, M. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, 65:1–8, 2021.

[67] Gurung, A.B., Ali, M.A., Lee, J., Farah, M.A., and Al-Anazi, K.M. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. *BioMed Research International*, page 8853056, 2021.

[68] Lin, A., Giuliano, C.J., Palladino, A., John, K.M., Abramowicz, C., Yuan, M.L., Sausville, E.L., Lukow, D.A., Liu, L., Chait, A.R., Galluzzo, Z.C., Tucker, C., and Sheltzer, J.M. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, 11(509), 2019.

[69] Keskin, O., Tuncbag, N., and Gursoy, A. Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, 116(8):4884–4909, 2016.

[70] Morilla, I., Lees, J.G., Reid, A.J., Orengo, C., and Ranea, J.A. Assessment of protein domain fusions in human protein interaction networks prediction: Application to the human kinetochore model. *New Biotechnology*, 27(6):755–765, 2010.

[71] Jansen, R., Greenbaum, D., and Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, 2002.

[72] Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America*, 104(11):4337–4341, 2007.

[73] Ofran, Y. and Rost, B. ISIS: Interaction sites identified from sequence. *Bioinformatics*, 23(2):13–16, 2007.

[74] Romero-Molina, S., Ruiz-Blanco, Y.B., Harms, M., Münch, J., and Sanchez-Garcia, E. PPI-Detect: A support vector machine model for sequence-based prediction of protein-protein interactions. *Journal of Computational Chemistry*, 40(11):1233–1242, 2019.

[75] Lu, H.C., Fornili, A., and Fraternali, F. Protein-Protein interaction networks studies and importance of 3D structure knowledge. *Expert Review of Proteomics*, 10(6):511–520, 2013.

[76] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., and Von Mering, C. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[77] Laddach, A., Chung, S.S., and Fraternali, F. Prediction of Protein-Protein Interactions: Looking Through the Kaleidoscope. In Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C.B.T.E.o.B., and Biology, C., editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 834–848. Academic Press, Oxford, 2019.

[78] Krissinel, E. and Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, 372(3):774–797, 2007.

[79] Vakser, I.A. Protein-protein docking: From interaction to interactome. *Biophysical Journal*, 107(8):1785–1793, 2014.

[80] Das, S. and Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific Reports*, 11(1):1–12, 2021.

[81] Periole, X., Zeppelin, T., and Schiøtt, B. Dimer Interface of the Human Serotonin Transporter and Effect of the Membrane Composition. *Scientific Reports*, 8(1):1–15, 2018.

[82] Moreira, I., Fernandes, P., and Ramos, M. Hot spots - A review of the protein-protein interface determinant amino-acid residues. *Proteins: Structure, Function and Bioinformatics*, 68(4):803–812, 2007.

[83] Clackson, T. and Wells, J.A. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

[84] Morrow, J.K. and Zhang, S. Computational Prediction of Protein Hot Spot Residues. *Current Drug Metabolism*, 18(9):1255–1265, 2012.

[85] Kortemme, T. and Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14116–14121, 2002.

[86] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Research*, 33(suppl2):W382–W388, 2005.

[87] Ofran, Y. and Rost, B. Protein-protein interaction hotspots carved into sequences. *PLoS Computational Biology*, 3(7):1169–1176, 2007.

[88] Tuncbag, N., Gursoy, A., and Keskin, O. Identification of computational hot spots in protein interfaces: Combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520, 2009.

[89] Zhu, X. and Mitchell, J.C. KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins: Structure, Function and Bioinformatics*, 79(9):2671–2683, 2011.

[90] Huo, S., Massova, I., and Kollman, P.A. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *Journal of Computational Chemistry*, 23(1): 15–27, 2002.

[91] Rajamani, D., Thiel, S., Vajda, S., and Camacho, C.J. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11287–11292, 2004.

[92] London, N., Raveh, B., Movshovitz-Attias, D., and Schueler-Furman, O. Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? *Proteins: Structure, Function and Bioinformatics*, 78(15):3140–3149, 2010.

[93] London, N., Raveh, B., and Schueler-Furman, O. Druggable protein-protein interactions - from hot spots to hot segments. *Current Opinion in Chemical Biology*, 17(6):952–959, 2013.

[94] Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., and Kmiecik, S. Protein–peptide docking: opportunities and challenges. *Drug Discovery Today*, 23(8):1530–1537, 2018.

[95] Pelay-Gimeno, M., Glas, A., Koch, O., and Grossmann, T.N. Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes. *Angewandte Chemie - International Edition*, 54(31):8896–8927, 2015.

[96] Sedan, Y., Marcu, O., Lyskov, S., and Schueler-Furman, O. Peptiderive server: derive peptide inhibitors from protein-protein interactions. *Nucleic Acids Research*, 44(W1): W536–W541, 2016.

[97] Chakraborty, S., Cole, S., Rader, N., King, C., Rajnarayanan, R., and Biswas, P.K. In silico design of peptidic inhibitors targeting estrogen receptor alpha dimer interface. *Molecular Diversity*, 16(3):441–451, 2012.

[98] Chakraborty, S., Asare, B.K., Biswas, P.K., and Rajnarayanan, R.V. Designer interface peptide grafts target estrogen receptor alpha dimerization. *Biochemical and Biophysical Research Communications*, 478(1):116–122, 2016.

[99] Collier, G. and Ortiz, V. Emerging computational approaches for the study of protein allostery. *Archives of Biochemistry and Biophysics*, 538(1):6–15, 2013.

[100] Schueler-Furman, O. and Wodak, S.J. Computational approaches to investigating allostery. *Current Opinion in Structural Biology*, 41:159–171, 2016.

[101] Greener, J.G. and Sternberg, M.J. Structure-based prediction of protein allostery. *Current Opinion in Structural Biology*, 50:1–8, 2018.

[102] Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69, 2003.

[103] Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[104] Smock, R.G., Rivoire, O., Russ, W.P., Swain, J.F., Leibler, S., Ranganathan, R., and Gierasch, L.M. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Molecular Systems Biology*, 6(414):1–10, 2010.

[105] Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. Hotspots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575, 2011.

[106] Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., Zhao, Y., Liu, Y., Chen, Z., Hou, T., and Zhang, J. Allosite: A method for predicting allosteric sites. *Bioinformatics*, 29(18):2357–2359, 2013.

[107] Chen, A.S.Y., Westwood, N.J., Brear, P., Rogers, G.W., Mavridis, L., and Mitchell, J.B.O. A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Molecular Informatics*, 35(3-4):125–135, 2016.

[108] Akbar, R. and Helms, V. ALLO: A tool to discriminate and prioritize allosteric pockets. *Chemical Biology and Drug Design*, 91(4):845–853, 2018.

[109] Shukla, D., Meng, Y., Roux, B., and Pande, V.S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature Communications*, 5:3397, 2014.

[110] Panjkovich, A. and Daura, X. Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics*, 13(1):273, 2012.

[111] Greener, J.G. and Sternberg, M.J.E. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics*, 16(1):335, 2015.

[112] Qi, Y., Wang, Q., Tang, B., and Lai, L. Identifying allosteric binding sites in proteins with a two-state gô model for novel allosteric effector discovery. *Journal of Chemical Theory and Computation*, 8(8):2962–2971, 2012.

[113] Greener, J.G., Filippis, I., and Sternberg, M.J. Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure*, 25(3):546–558, 2017.

[114] Amamuddy, O.S., Veldman, W., Manyumwa, C., Khairallah, A., Agajanian, S., Oluyemi, O., Verkhivker, G.M., and Bishop, Ö.T. Integrated computational approaches and tools for allosteric drug discovery. *International Journal of Molecular Sciences*, 21(3):847, 2020.

[115] Song, K., Liu, X., Huang, W., Lu, S., Shen, Q., Zhang, L., and Zhang, J. Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling*, 57(9):2358–2363, 2017.

[116] Weinkam, P., Pons, J., and Sali, A. Structure-based model of allostery predicts coupling between distant sites. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13):4875–4880, 2012.

[117] Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. GSATools: Analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics*, 29(16):2053–2055, 2013.

[118] Tekpinar, M., Neron, B., and Delarue, M. Extracting Dynamical Correlations and Identifying Key Residues for Allosteric Communication in Proteins by correlationplus. *Journal of Chemical Information and Modeling*, 61(10):4832–4838, 2021.

[119] Xie, J., Wang, S., Xu, Y., Deng, M., and Lai, L. Uncovering the Dominant Motion Modes of Allosteric Regulation Improves Allosteric Site Prediction. *Journal of Chemical Information and Modeling*, 62:187–195, 2022.

[120] Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., Wang, Q., Shi, T., Zhao, Y., Wang, Y., Li, W., Li, Y., Chen, H., ZhangChen, G., and Zhang, J. ASD: A comprehensive database of allosteric proteins and modulators. *Nucleic Acids Research*, 39(suppl1):D663–D669, 2011.

[121] Huang, Z., Mou, L., Shen, Q., Lu, S., Li, C., Liu, X., Wang, G., Li, S., Geng, L., Liu, Y., Wu, J., Chen, G., and Zhang, J. ASD v2.0: Updated content and novel features focusing on allosteric regulation. *Nucleic Acids Research*, 42(D1):D510–D516, 2014.

[122] Shen, Q., Wang, G., Li, S., Liu, X., Lu, S., Chen, Z., Song, K., Yan, J., Geng, L., Huang, Z., Huang, W., Chen, G., and Zhang, J. ASD v3.0: Unraveling Allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Research*, 44(D1): D527–D535, 2016.

[123] Liu, X., Lu, S., Song, K., Shen, Q., Ni, D., Li, Q., He, X., Zhang, H., Wang, Q., Chen, Y., Li, X., Wu, J., Sheng, C., Chen, G., Liu, Y., Lu, X., and Zhang, J. Unraveling allosteric landscapes of allosterome with ASD. *Nucleic Acids Research*, 48(D1):D394–D401, 2020.

[124] Huang, W., Wang, G., Shen, Q., Liu, X., Lu, S., Geng, L., Huang, Z., and Zhang, J. ASBench: Benchmarking sets for allosteric discovery. *Bioinformatics*, 31(15):2598–2600, 2015.

[125] Zlobin, A., Suplatov, D., Kopylov, K., and Švedas, V. CASBench: A benchmarking set of proteins with annotated catalytic and allosteric sites in their structures. *Acta Naturae*, 11(1):74–80, 2019.

[126] Stumpf, M.P., Thorne, T., De Silva, E., Stewart, R., Hyeong, J.A., Lappe, M., and Wiuf, C. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6959–6964, 2008.

[127] Chowdhury, S., Hepper, S., Lodi, M.K., Saier, M.H., and Uetz, P. The protein interactome of glycolysis in escherichia coli. *Proteomes*, 9(2):1–16, 2021.

[128] Barabási, A.L., Gulbahce, N., and Loscalzo, J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[129] Athanasios, A., Charalampos, V., Vasileios, T., and Ashraf, G.M. Protein-Protein Interaction (PPI) Network: Recent Advances in Drug Discovery. *Current drug metabolism*, 18(1):5–10, 2017.

[130] Wuchty, S. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18(9):1694–1702, 2001.

[131] Levy, E.D., Pereira-Leal, J.B., Chothia, C., and Teichmann, S.A. 3D complex: A structural classification of protein complexes. *PLoS Computational Biology*, 2(11):1395–1406, 2006.

[132] Koch, I., Kaden, F., and Selbig, J. Analysis of protein sheet topologies by graph theoretical methods. *Proteins: Structure, Function, and Bioinformatics*, 12(4):314–323, 1992.

[133] May, P., Kreuchwig, A., Steinke, T., and Koch, I. PTGL: A database for secondary structure-based protein topologies. *Nucleic Acids Research*, 38(suppl1):D326–D330, 2009.

[134] Wolf, J.N., Keßler, M., Ackermann, J., and Koch, I. PTGL: Extension to graph-based topologies of cryo-EM data for large protein structures. *Bioinformatics*, 37(7):1032–1034, 2021.

[135] Koch, I. and Schäfer, T. Protein super-secondary structure and quaternary structure topology: theoretical description and application. *Current Opinion in Structural Biology*, 50:134–143, 2018.

[136] Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., and Giuliani, A. Protein contact networks: An emerging paradigm in chemistry. *Chemical Reviews*, 113(3):1598–1613, 2013.

[137] Chennubhotla, C. and Bahar, I. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Computational Biology*, 3(9):1716–1726, 2007.

[138] Fokas, A.S., Cole, D.J., Ahnert, S.E., and Chin, A.W. Residue Geometry Networks: A Rigidity-Based Approach to the Amino Acid Network and Evolutionary Rate Analysis. *Scientific Reports*, 6:33213, 2016.

[139] Ribeiro, A.A. and Ortiz, V. Determination of signaling pathways in proteins through network theory: Importance of the topology. *Journal of Chemical Theory and Computation*, 10(4):1762–1769, 2014.

[140] Liang, Z., Verkhivker, G.M., and Hu, G. Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: Theory, tools and applications. *Briefings in Bioinformatics*, 21(3):815–835, 2020.

[141] Pacini, L., Dorantes-Gilardi, R., Vuillon, L., and Lesieur, C. Mapping Function from Dynamics: Future Challenges for Network-Based Models of Protein Structures. *Frontiers in Molecular Biosciences*, 8:744646, 2021.

[142] Brinda, K.V., Kannan, N., and Vishveshwara, S. Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Engineering*, 15(4):265–277, 2002.

[143] Del Sol, A. and O'Meara, P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins: Structure, Function and Genetics*, 58(3):672–682, 2005.

[144] Feher, V.A., Durrant, J.D., Van Wart, A.T., and Amaro, R.E. Computational approaches to mapping allosteric pathways. *Current Opinion in Structural Biology*, 25:98–103, 2014.

[145] Kaya, C., Armutlulu, A., Ekesan, S., and Haliloglu, T. MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Research*, 41(W1):W249–W255, 2013.

[146] Li, H., Chang, Y.Y., Lee, J.Y., Bahar, I., and Yang, L.W. DynOmics: Dynamics of structural proteome and beyond. *Nucleic Acids Research*, 45(W1):W374–W380, 2017.

[147] Guarnera, E. and Berezovsky, I.N. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS Computational Biology*, 12(3):e1004678, 2016.

[148] Sen, T.Z., Feng, Y., Garcia, J.V., Kloczkowski, A., and Jernigan, R.L. The Extent of Cooperativity of Protein Motions Observed with Elastic Network Models is Similar for Atomic and Coarser-Grained Models. *Journal of Chemical Theory and Computation*, 2 (3):696–704, 2006.

[149] Amor, B.R.C., Schaub, M.T., Yaliraki, S.N., and Barahona, M. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature Communications*, 7(1):12477, 2016.

[150] Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. Protein flexibility predictions using graph theory. 44(2):150–165, 2001.

[151] Thorpe, M.F., Lei, M., Rader, A.J., Jacobs, D.J., and Kuhn, L.A. Protein flexibility and dynamics using constraint theory. *Journal of Molecular Graphics and Modelling*, 19(1): 60–69, 2001.

[152] Hespenheide, B.M., Rader, A.J., Thorpe, M.F., and Kuhn, L.A. Identifying protein folding cores from the evolution of flexible regions during unfolding. *Journal of Molecular Graphics and Modelling*, 21(3):195–207, 2002.

[153] Radestock, S. and Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Engineering in Life Sciences*, 8(5):507–522, 2008.

[154] Nutschel, C., Coscolín, C., David, B., Mulnaes, D., Ferrer, M., Jaeger, K.E., and Gohlke, H. Promiscuous Esterases Counterintuitively Are Less Flexible than Specific Ones. *Journal of Chemical Information and Modeling*, 61(5):2383–2395, 2021.

[155] Pfleger, C., Minges, A., Boehm, M., Mcclendon, C.L., Torella, R., and Gohlke, H. Ensemble- and Rigidity Theory-Based Perturbation Approach To Analyze Dynamic Allostery. *Journal of Chemical Theory and Computation*, 13(12):6343–6357, 2017.

[156] Pfleger, C., Kusch, J., Kondapuram, M., Schwabe, T., Sattler, C., Benndorf, K., and Gohlke, H. Allosteric signaling in C-linker and cyclic nucleotide-binding domain of HCN2 channels. *Biophysical Journal*, 120(5):950–963, 2021.

[157] Veloso, C.J., Silveira, C.H., Melo, R.C., Ribeiro, C., Lopes, J.C., Santoro, M.M., and Meira, W. On the characterization of energy networks of proteins. *Genetics and Molecular Research*, 6(4):799–820, 2007.

[158] Delmotte, A., Tate, E.W., Yaliraki, S.N., and Barahona, M. Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. *Physical Biology*, 8(5):055010, 2011.

[159] Amor, B., Yaliraki, S.N., Woscholski, R., and Barahona, M. Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Molecular BioSystems*, 10(8):2247–2258, 2014.

[160] Song, F., Barahona, M., and Yaliraki, S.N. BagPype: A Python package for the construction of atomistic, energy-weighted graphs from biomolecular structures. figshare, 2021. Available at: https://doi.org/10.6084/m9.figshare.14039723.v1.

[161] Chrysostomou, S., Roy, R., Prischi, F., Thamlikitkul, L., Chapman, K.L., Mufti, U., Peach, R., Ding, L., Hancock, D., Moore, C., Molina-Arcas, M., Mauri, F., Pinato, D.J.,

Abrahams, J.M., Ottaviani, S., Castellano, L., Giamas, G., Pascoe, J., Moonamale, D., Pirrie, S., Gaunt, C., Billingham, L., Steven, N.M., Cullen, M., Hrouda, D., Winkler, M., Post, J., Cohen, P., Salpeter, S.J., Bar, V., Zundelevich, A., Golan, S., Leibovici, D., Lara, R., Klug, D.R., Yaliraki, S.N., Barahona, M., Wang, Y., Downward, J., Skehel, J.M., Ali, M.M.U., Seckl, M.J., and Pardo, O.E. Repurposed floxacins targeting RSK4 prevent chemoresistance and metastasis in lung and bladder cancer. *Science Translational Medicine*, 13(602):eaba4627, 2021.

[162] Hodges, M., Barahona, M., and Yaliraki, S.N. Allostery and cooperativity in multimeric proteins: bond-to-bond propensities in ATCase. *Scientific Reports*, 8(1):11079, 2018.

[163] Vianello, F. Computational characterisation of protein interaction sites: from small ligand pockets to large domain interfaces. PhD thesis, Imperial College London, 2020.

[164] Sydow, D., Burggraaff, L., Szengel, A., Van Vlijmen, H.W., Ijzerman, A.P., Van Westen, G.J., and Volkamer, A. Advances and Challenges in Computational Target Prediction. *Journal of Chemical Information and Modeling*, 59(5):1728–1742, 2019.

[165] Lu, S., He, X., Ni, D., and Zhang, J. Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *Journal of Medicinal Chemistry*, 62(14):6405–6421, 2019.

[166] Meliga, S. Graph Clustering of Atomic Networks of Protein Dynamics. PhD thesis, Imperial College London, 2009.

[167] Delmotte, A. All-scale structural analysis of biomolecules through dynamical graph partitioning. PhD thesis, Imperial College London, 2014.

[168] Amor, B.R.C. Exploring allostery in proteins with graph theory. PhD thesis, Imperial College London, 2016.

[169] Song, F. Modelling biomolecules through atomistic graphs: theory, implementation and applications. Phd, Imperial College London, 2022.

[170] Word, J., Lovell, S.C., Richardson, J.S., and Richardson, D.C. Asparagine and glutamine:

using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, 1999.

[171] Rader, A.J., Hespenheide, B.M., Kuh, L.A., and Thorpe, M.F. Protein unfolding: Rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3540–3545, 2002.

[172] Python Software Foundation. The Python Programming Language., 2001-2021. Available at: www.python.org.

[173] Lin, M.S., Fawzi, N.L., and Head-Gordon, T. Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure*, 15(6):727–740, 2007.

[174] Beguerisse-Díaz, M., Vangelov, B., and Barahona, M. Finding role communities in directed networks using Role-Based Similarity, Markov Stability and the Relaxed Minimum Spanning Tree. *IEEE Global Conference on Signal and Information Processing*, pages 937–940, 2013.

[175] Huheey, J.E., Keiter, E.A., and Keiter, R.L. *Inorganic chemistry: principles of structure and reactivity*. HarperCollins College Publishers, New York, NY, 1993.

[176] Mayo, S.L., Olafson, B.D., and Goddard, W.A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry*, 94(26):8897–8909, 1990.

[177] Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. Automated design of the surface positions of protein helices. *Protein Science*, 6(6):1333–1337, 1997.

[178] Gilson, M.K. and Honig, B.H. The dielectric constant of a folded protein. *Biopolymers*, 25(11):2097–2119, 1986.

[179] Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[180] Hunter, C.A. and Sanders, J.K. The Nature of π-π Interactions. *Journal of the American Chemical Society*, 112(14):5525–5534, 1990.

[181] Warshel, A., Sharma, P.K., Kato, M., and Parson, W.W. Modeling electrostatic effects in proteins. *Biochimica et biophysica acta*, 1764(11):1647–1676, 2006.

[182] Schaub, M.T., Lehmann, J., Yaliraki, S.N., and Barahona, M. Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution. *Network Science*, 2(1):66–89, 2014.

[183] Chennubhotla, C. and Bahar, I. Markov propagation of allosteric effects in biomolecular systems: Application to GroEL-GroES. *Molecular Systems Biology*, 2:36, 2006.

[184] Delvenne, J.C., Yaliraki, S.N., and Barahon, M. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12755–12760, 2010.

[185] Lambiotte, R., Delvenne, J.C., and Barahona, M. Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, 2014.

[186] Zhang, H., Salazar, J.D., and Yaliraki, S.N. Proteins across scales through graph partitioning: application to the major peanut allergen Ara h 1. *Journal of Complex Networks*, 6(5):679–692, 2018.

[187] Hodges, M., Yaliraki, S.N., and Barahona, M. Edge-based formulation of elastic network models. *Physical Review Research*, 1(3):1–9, 2019.

[188] Peach, R.L., Klug, D.R., Saman, D., Yaliraki, S.N., and Willison, K.R. Unsupervised Graph - Based Learning Predicts. *bioRxiv*, 2019.

[189] Koenker, R. *Quantile regression*. Cambridge University Press, 2005.

[190] Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman and Hall/CRC., 1st edition, 1994.

[191] Delano, W.L. The PyMOL Molecular Graphics System., 2004.

[192] Krissinel, E. and Henrick, K. Detection of Protein Assemblies in Crystals. In R. Berthold, M., Glen, R.C., Diederichs, K., Kohlbacher, O., and Fischer, I., editors, *International Symposium on Computational Life Science*, pages 163–174, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[193] Panjkovich, A. and Daura, X. PARS: A web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics*, 30(9):1314–1315, 2014.

[194] Goncearenco, A., Mitternacht, S., Yong, T., Eisenhaber, B., Eisenhaber, F., and Berezovsky, I.N. SPACER: Server for predicting allosteric communication and effects of regulation. *Nucleic Acids Research*, 41(Web Server issue):266–272, 2013.

[195] Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R.W., Chen, J., and Gerstein, M. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure*, 24(5):826–837, 2016.

[196] Wang, J., Jain, A., McDonald, L.R., Gambogi, C., Lee, A.L., and Dokholyan, N.V. Mapping allosteric communications within individual proteins. *Nature Communications*, 11:3862, 2020.

[197] Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., Shen, Y., Chen, F., Lai, L., and Pei, J. CavityPlus: A web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Research*, 46(W1):W374–W379, 2018.

[198] Tian, H., Jiang, X., and Tao, P. PASSer: Prediction of allosteric sites server. *Machine Learning: Science and Technology*, 2(3):035015, 2021.

[199] Tan, Z.W., Guarnera, E., Tee, W.V., and Berezovsky, I.N. AlloSigMA 2: paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. *Nucleic Acids Research*, 48(W1):W116–W124, 2020.

[200] Django Software Foundation. Django (Version 3.1.), 2020.

[201] Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlic, A., and Rose, P.W. NGL viewer: Web-based molecular graphics for large complexes. *Bioinformatics*, 34(21):3755–3758, 2018.

[202] Strömich, L. Molecular mechanisms and allostery in estrogen receptor alpha using bond-to-bond propensities. Master's thesis, Imperial College London, 2018.

[203] Weake, V.M. and Workman, J.L. Inducible gene expression: Diverse regulatory mechanisms. *Nature Reviews Genetics*, 11(6):426–437, 2010.

[204] Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. The Human Transcription Factors. *Cell*, 172(4):650–665, 2018.

[205] Ribeiro, R.C., Kushner, P.J., and Baxter, J.D. The nuclear hormone receptor gene superfamily. *Annual Review of Medicine*, 46:443–453, 1995.

[206] Katzenellenbogen, B.S., Choi, I., Delage-Mourroux, R., Ediger, T.R., Martini, P.G., Montano, M., Sun, J., Weis, K., and Katzenellenbogen, J.A. Molecular mechanisms of estrogen action: Selective ligands and receptor pharmacology. *Journal of Steroid Biochemistry and Molecular Biology*, 74(5):279–285, 2000.

[207] Toft, D. and Gorski, J. A receptor molecule for estrogens: isolation from the rat uterus and preliminary characterization. *Biochemistry*, 55(6):1574–1581, 1966.

[208] Toft, D., Shyamala, G., and Gorski, J. A receptor molecule for estrogens: studies using a cell-free system. *Biochemistry*, 57(6):1740–1743, 1967.

[209] Kuiper, G.G., Enmark, E., Pelto-Huikko, M., Nilsson, S., and Gustafsson, J.Å. Cloning of a novel estrogen receptor expressed in rat prostate and ovary. *Proceedings of the National Academy of Sciences of the United States of America*, 93(12):5925–5930, 1996.

[210] Jia, M., Dahlman-Wright, K., and Gustafsson, J.Å. Estrogen receptor alpha and beta in health and disease. *Best Practice and Research: Clinical Endocrinology and Metabolism*, 29(4):557–568, 2015.

[211] Levin, E.R. Plasma membrane estrogen receptors. *Trends in Endocrinology and Metabolism*, 20(10):477–482, 2009.

[212] Klinge, C.M. Estrogenic control of mitochondrial function and biogenesis. *Journal of Cellular Biochemistry*, 105(6):1342–1351, 2008.

[213] Gruber, C.J., Gruber, D.M., Gruber, I.M., Wieser, F., and Huber, J.C. Anatomy of the estrogen response element. *Trends in Endocrinology and Metabolism*, 15(2):73–78, 2004.

[214] Zwart, W., Theodorou, V., Kok, M., Canisius, S., Linn, S., and Carroll, J.S. Oestrogen receptor-co-factor chromatin specificity in the transcriptional regulation of breast cancer. *EMBO Journal*, 30(23):4764–4776, 2011.

[215] Yi, P., Wang, Z., Feng, Q., Pintilie, G.D., Foulds, C.E., Lanz, R.B., Ludtke, S.J., Schmid, M.F., Chiu, W., and O'Malley, B.W. Structure of a Biologically Active Estrogen Receptor-Coactivator Complex on DNA. *Molecular Cell*, 57(6):1047–1058, 2015.

[216] Kushner, P., Agard, D., Greene, G., Scanlan, T., Shiau, A., Uht, R., and Webb, P. Estrogen receptor pathways to AP-1. *The Journal of Steroid Biochemistry and Molecular Biology*, 74(5):311–317, 2000.

[217] Safe, S. Transcriptional activation of genes by 17 beta-estradiol through estrogen receptor-Sp1 interactions. *Vitamins and Hormones*, 62:231–252, 2001.

[218] Kumar, V. and Chambon, P. The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell*, 55(1):145–156, 1988.

[219] Brzozowski, A.M., Pike, A.C.W., Dauter, Z., Hubbard, R.E., Bonn, T., Engström, O., Öhman, L., Greene, G.L., Gustafsson, J.Å., and Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, 389(6652):753–758, 1997.

[220] Tanenbaum, D.M., Wang, Y., Williams, S.P., and Sigler, P.B. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11):5998–6003, 1998.

[221] Lees, J.A., Fawell, S.E., White, R., and Parker, M.G. A 22-amino-acid peptide restores DNA-binding activity to dimerization-defective mutants of the estrogen receptor. *Molecular and Cellular Biology*, 10(10):5529–5531, 1990.

[222] Kumar, R. and Thompson, E.B. Transactivation functions of the N-terminal domains of nuclear hormone receptors: Protein folding and coactivator interactions. *Molecular Endocrinology*, 17(1):1–10, 2003.

[223] Kraus, W.L., Mcinerney, E.M., and Katzenellenbogen, B.S. Ligand-dependent, transcriptionally productive association of the amino- and carboxyl-terminal regions of a steroid hormone nuclear receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 92(26):12314–12318, 1995.

[224] Webb, P., Nguyen, P., Shinsako, J., Anderson, C., Feng, W., Nguyen, M.P., Chen, D., Huang, S.M., Subramanian, S., McKinerney, E., Katzenellenbogen, B.S., Stallcup, M.R., and Kushner, P.J. Estrogen receptor activation function 1 works by binding p160 coactivator proteins. *Molecular Endocrinology*, 12(10):1605–1618, 1998.

[225] Pike, A.C., Brzozowski, A.M., and Hubbard, R.E. A structural biologist's view of the oestrogen receptor. *The Journal of Steroid Biochemistry and Molecular Biology*, 74(5): 261–268, 2000.

[226] Tamrazi, A., Carlson, K.E., Daniels, J.R., Hurth, K.M., and Katzenellenbogen, J.A. Estrogen Receptor Dimerization: Ligand Binding Regulates Dimer Affinity and DimerDissociation Rate. *Molecular Endocrinology*, 16(12):2706–2719, 2002.

[227] Eiler, S., Gangloff, M., Duclaud, S., Moras, D., and Ruff, M. Overexpression, Purification, and Crystal Structure of Native ERα LBD. *Protein Expression and Purification*, 22(2): 165–173, 2001.

[228] Rastinejad, F., Ollendorff, V., and Polikarpov, I. Nuclear receptor full-length architectures: confronting myth and illusion with high resolution. *Trends in biochemical sciences*, 40(1):16–24, 2015.

[229] Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., and Rastinejad, F. Structure of the intact PPAR-γ-RXR-α nuclear receptor complex on DNA. *Nature*, 456(7220):350–356, 2008.

[230] Chandra, V., Huang, P., Potluri, N., Wu, D., Kim, Y., and Rastinejad, F. Multidomain integration in the structure of the HNF-4α nuclear receptor complex. *Nature*, 495(7441): 394–398, 2013.

[231] Lou, X., Toresson, G., Benod, C., Suh, J.H., Philips, K.J., Webb, P., and Gustafsson, J.A. Structure of the retinoid X receptor α-liver X receptor β (RXRα-LXRβ) heterodimer on DNA. *Nature Structural and Molecular Biology*, 21(3):277–281, 2014.

[232] Chandra, V., Wu, D., Li, S., Potluri, N., Kim, Y., and Rastinejad, F. The quaternary architecture of RARβ-RXRα heterodimer facilitates domain-domain signal transmission. *Nature Communications*, 8(1):1–9, 2017.

[233] Gruber, C.J., Tschugguel, W., Schneeberger, C., and Huber, J.C. Production and Actions of Estrogens. *New England Journal of Medicine*, 346(5):340–352, 2002.

[234] Mak, H.Y., Hoare, S., Henttu, P.M., and Parker, M.G. Molecular determinants of the estrogen receptor-coactivator interface. *Molecular and Cellular Biology*, 19(5):3895–3903, 1999.

[235] Shiau, A.K., Barstad, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A., and Greene, G.L. The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell*, 95(7):927–937, 1998.

[236] Anderson, E. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Research*, 4(5):197–201, 2002.

[237] Weigelt, B. and Reis-Filho, J.S. Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nature Reviews Clinical Oncology*, 6(12):718–730, 2009.

[238] Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, Ø., Pergamenschikov, A., Williams,

C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.L., Brown, P.O., and Botstein, D. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[239] Pearce, S.T. and Jordan, V.C. The biological role of estrogen receptors alpha and beta in cancer. *Critical Reviews in Oncology/Hematology*, 50(1):3–22, 2004.

[240] Yip, C.H. and Rhodes, A. Estrogen and progesterone receptors in breast cancer. *Future Oncology*, 10(14):2293–2301, 2014.

[241] Jensen, E.V. and Jordan, V.C. The estrogen receptor: a model for molecular medicine. *Clinical Cancer Research*, 9(6):1980–1989, 2003.

[242] Jordan, V.C. Tamoxifen: A most unlikely pioneering medicine. *Nature Reviews Drug Discovery*, 2(3):205–213, 2003.

[243] Puyang, X., Furman, C., Zheng, G.Z., Wu, Z.J., Banka, D., Aithal B, K., Agoulnik, S., Bolduc, D.M., Buonamici, S., Caleb, B., Das, S., Eckley, S., Fekkes, P., Hao, M.H., Hart, A., Houtman, R., Irwin, S., Joshi, J.J., Karr, C., Kim, A., Kumar, N., Kumar, P., Kuznetsov, G., Lai, W.G., Larsen, N., MacKenzie, C., Martin, L.A., Melchers, D., Moriarty, A., Nguyen, T.V., Norris, J., O'Shea, M., Pancholi, S., Prajapati, S., Rajagopalan, S., Reynolds, D.J., Rimkunas, V., Rioux, N., Ribas, R., Siu, A., Sivakumar, S., Subramanian, V., Thomas, M., Vaillancourt, F.H., Wang, J., Wardell, S., Wick, M.J., Yao, S., Yu, L., Warmuth, M., Smith, P.G., Zhu, P., and Korpal, M. Discovery of Selective Estrogen Receptor Covalent Antagonists (SERCAs) for the treatment of ERa(WT) and ERa(MUT) breast cancer. *Cancer Discovery*, 8(9):1176–1193, 2018.

[244] Cyrus, K., Wehenkel, M., Choi, E.Y., Lee, H., Swanson, H., and Kim, K.B. Jostling for position: Optimizing linker location in the design of estrogen receptor-targeting PROTACs. *ChemMedChem*, 5(7):979–985, 2010.

[245] Ali, S., Buluwela, L., and Coombes, R.C. Antiestrogens and Their Therapeutic Applications in Breast Cancer and Other Diseases. *Annual Review of Medicine*, 62(1):217–232, 2011.

[246] McDonnell, D.P. The molecular pharmacology of SERMs. *Trends in Endocrinology and Metabolism*, 10(8):301–311, 1999.

[247] Vajdos, F.F., Hoth, L.R., Geoghegan, K.F., Simons, S.P., LeMotte, P.K., Danley, D.E., Ammirati, M.J., and Pandit, J. The 2.0 Å crystal structure of the ERα ligand-binding domain complexed with lasofoxifene. *Protein Science*, 16(5):897–905, 2007.

[248] Fanning, S.W., Jeselsohn, R., Dharmarajan, V., Mayne, C.G., Karimi, M., Buchwalter, G., Houtman, R., Toy, W., Fowler, C.E., Han, R., Lainé, M., Carlson, K.E., Martin, T.A., Nowak, J., Nwachukwu, J.C., Hosfield, D.J., Chandarlapaty, S., Tajkhorshid, E., Nettles, K.W., Griffin, P.R., Shen, Y., Katzenellenbogen, J.A., Brown, M., and Greene, G.L. The SERM/SERD bazedoxifene disrupts ESR1 helix 12 to overcome acquired hormone resistance in breast cancer cells. *eLife*, 7:e37161, 2018.

[249] Nathan, M.R. and Schmid, P. A Review of Fulvestrant in Breast Cancer. *Oncology and Therapy*, 5(1):17–29, 2017.

[250] Osborne, C.K., Wakeling, A., and Nicholson, R.I. Fulvestrant: An oestrogen receptor antagonist with a novel mechanism of action. *British Journal of Cancer*, 90:S2–S6, 2004.

[251] De Savi, C., Bradbury, R.H., Rabow, A.A., Norman, R.A., De Almeida, C., Andrews, D.M., Ballard, P., Buttar, D., Callis, R.J., Currie, G.S., Curwen, J.O., Davies, C.D., Donald, C.S., Feron, L.J., Gingell, H., Glossop, S.C., Hayter, B.R., Hussain, S., Karoutchi, G., Lamont, S.G., MacFaul, P., Moss, T.A., Pearson, S.E., Tonge, M., Walker, G.E., Weir, H.M., and Wilson, Z. Optimization of a Novel Binding Motif to acrylic Acid (AZD9496), a Potent and Orally Bioavailable Selective Estrogen Receptor Downregulator and Antagonist. *Journal of Medicinal Chemistry*, 58(20):8128–8140, 2015.

[252] Lai, A., Kahraman, M., Govek, S., Nagasawa, J., Bonnefous, C., Julien, J., Douglas, K., Sensintaffar, J., Lu, N., Lee, K.J., Aparicio, A., Kaufman, J., Qian, J., Shao, G., Prudente, R., Moon, M.J., Joseph, J.D., Darimont, B., Brigham, D., Grillot, K., Heyman, R., Rix, P.J., Hager, J.H., and Smith, N.D. Identification of GDC-0810 (ARN-810), an Orally Bioavailable Selective Estrogen Receptor Degrader (SERD) that Demonstrates

Robust Activity in Tamoxifen-Resistant Breast Cancer Xenografts. *Journal of Medicinal Chemistry*, 58(12):4888–4904, 2015.

[253] Bardia, A., Aftimos, P., Bihani, T., Anderson-Villaluz, A.T., Jung, J., Conlan, M.G., and Kaklamani, V.G. EMERALD: Phase III trial of elacestrant (RAD1901) vs endocrine therapy for previously treated ER+ advanced breast cancer. *Future Oncology*, 15(28): 3209–3218, 2019.

[254] Scott, J.S., Moss, T.A., Balazs, A., Barlaam, B., Breed, J., Carbajo, R.J., Chiarparin, E., Davey, P.R., Delpuech, O., Fawell, S., Fisher, D.I., Gagrica, S., Gangl, E.T., Grebe, T., Greenwood, R.D., Hande, S., Hatoum-Mokdad, H., Herlihy, K., Hughes, S., Hunt, T.A., Huynh, H., Janbon, S.L., Johnson, T., Kavanagh, S., Klinowska, T., Lawson, M., Lister, A.S., Marden, S., McGinnity, D.F., Morrow, C.J., Nissink, J.W.M., O'Donovan, D.H., Peng, B., Polanski, R., Stead, D.S., Stokes, S., Thakur, K., Throner, S.R., Tucker, M.J., Varnes, J., Wang, H., Wilson, D.M., Wu, D., Wu, Y., Yang, B., and Yang, W. Discovery of AZD9833, a Potent and Orally Bioavailable Selective Estrogen Receptor Degrader and Antagonist. *Journal of Medicinal Chemistry*, 63(23):14530–14559, 2020.

[255] Weir, H.M., Bradbury, R.H., Rabow, A.A., Buttar, D., Callis, R.J., Curwen, J.O., De Almeida, C., Ballard, P., Hulse, M., Donald, C.S., Feron, L.J., Karoutchi, G., MacFaul, P., Moss, T., Norman, R.A., Pearson, S.E., Tonge, M., Davies, G., Walker, G.E., Wilson, Z., Rowlinson, R., Powell, S., Sadler, C., Richmond, G., Ladd, B., Pazolli, E., Mazzola, A.M., D'Cruz, C., and De Savi, C. AZD9496: An oral estrogen receptor inhibitor that blocks the growth of ER-positive and ESR1-mutant breast tumors in preclinical models. *Cancer Research*, 76(11):3307–3318, 2016.

[256] Chang, M. Tamoxifen resistance in breast cancer. *Biomolecules & Therapeutics*, 20(3): 256–267, 2012.

[257] Dustin, D., Gu, G., and Fuqua, S.A.W. ESR1 mutations in breast cancer. *Cancer*, 125 (21):3714–3728, 2019.

[258] Fanning, S.W., Mayne, C.G., Dharmarajan, V., Carlson, K.E., Martin, T.A., Novick,

S.J., Toy, W., Green, B., Panchamukhi, S., Katzenellenbogen, B.S., Tajkhorshid, E., Griffin, P.R., Shen, Y., Chandarlapaty, S., Katzenellenbogen, J.A., and Greene, G.L. Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *eLife*, 5: e12792, 2016.

[259] Bafna, D., Ban, F., Rennie, P.S., Singh, K., and Cherkasov, A. Computer-aided ligand discovery for estrogen receptor alpha. *International Journal of Molecular Sciences*, 21 (12):1–49, 2020.

[260] Raj, G.V., Sareddy, G.R., Ma, S., Lee, T.K., Viswanadhapalli, S., Li, R., Liu, X., Murakami, S., Chen, C.C., Lee, W.R., Mann, M., Krishnan, S.R., Manandhar, B., Gonugunta, V.K., Strand, D., Tekmal, R.R., Ahn, J.M., and Vadlamudi, R.K. Estrogen receptor coregulator binding modulators (ERXs) effectively target estrogen receptor positive human breast cancers. *eLife*, 6:e26857, 2017.

[261] Fratev, F. Activation helix orientation of the estrogen receptor is mediated by receptor dimerization: evidence from molecular dynamics simulations. *Physical Chemistry Chemical Physics*, 17(20):13403–13420, 2015.

[262] Yudt, M.R. and Koide, S. Preventing estrogen receptor action with dimer-interface peptides. *Steroids*, 66(7):549–558, 2001.

[263] Chaudhury, S., Lyskov, S., and Gray, J.J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, 2010.

[264] Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., Penzar, D., Perlman, S., Poon, L.L.M., Samborskiy, D.V., Sidorov, I.A., Sola, I., Ziebuhr, J., and of Viruses, C.S.G.o.t.I.C.o.T. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, 5(4):536–544, 2020.

[265] Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., and Zhang, Y.Z. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.

[266] Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., and Shi, Z.L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, 2020.

[267] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., and Tan, W. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733, 2020.

[268] Graham, R.L., Donaldson, E.F., and Baric, R.S. A decade after SARS: strategies for controlling emerging coronaviruses. *Nature Reviews Microbiology*, 11(12):836–848, 2013.

[269] Peiris, J.S.M., Guan, Y., and Yuen, K.Y. Severe acute respiratory syndrome. *Nature Medicine*, 10:S88–S97, 2004.

[270] Dyall, J., Gross, R., Kindrachuk, J., Johnson, R.F., Olinger Jr, G.G., Hensley, L.E., Frieman, M.B., and Jahrling, P.B. Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome: Current Therapeutic Options and Potential Targets for Novel Therapies. *Drugs*, 77(18):1935–1966, 2017.

[271] Yin, Y. and Wunderink, R.G. MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology*, 23(2):130–137, 2018.

[272] Fehr, A.R. and Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. In *Coronaviruses: Methods and Protocols*, volume 1282, pages 1–23. Springer New York, 2015.

[273] Shi, J., Wei, Z., and Song, J. Dissection study on the severe acute respiratory syndrome 3C-like protease reveals the critical role of the extra domain in dimerization of the enzyme: defining the extra domain as a new target for design of highly specific protease inhibitors. *The Journal of Biological Chemistry*, 279(23):24765–24773, 2004.

[274] Chen, Y.W., Yiu, C.P.B., and Wong, K.Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CLpro) structure: Virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, 9(129), 2020.

[275] Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z., Sun, L., Mo, L., Ye, S., Pang, H., Gao, G.F., Anand, K., Bartlam, M., Hilgenfeld, R., and Rao, Z. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (23):13190–13195, 2003.

[276] Anand, K., Palm, G.J., Mesters, J.R., Siddell, S.G., Ziebuhr, J., and Hilgenfeld, R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *The EMBO journal*, 21(13):3213–3224, 2002.

[277] Muramatsu, T., Takemoto, C., Kim, Y.T., Wang, H., Nishii, W., Terada, T., Shirouzu, M., and Yokoyama, S. SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, 113(46):12997–13002, 2016.

[278] Chen, S., Chen, L., Tan, J., Chen, J., Du, L., Sun, T., Shen, J., Chen, K., Jiang, H., and Shen, X. Severe acute respiratory syndrome coronavirus 3C-like proteinase N terminus is indispensable for proteolytic activity but not for enzyme dimerization: Biochemical and thermodynamic investigation in conjunction with molecular dynamics simulations. *Journal of Biological Chemistry*, 280(1):164–173, 2005.

[279] Hsu, W.C., Chang, H.C., Chou, C.Y., Tsai, P.J., Lin, P.I., and Chang, G.G. Critical assessment of important regions in the subunit association and catalytic action of the severe

acute respiratory syndrome coronavirus main protease. *Journal of Biological Chemistry*, 280(24):22741–22748, 2005.

[280] Chou, C.Y., Chang, H.C., Hsu, W.C., Lin, T.Z., Lin, C.H., and Chang, G.G. Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. *Biochemistry*, 43(47):14958–14970, 2004.

[281] Shi, J. and Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS Journal*, 273(5):1035–1045, 2006.

[282] Shi, J., Han, N., Lim, L., Lua, S., Sivaraman, J., Wang, L., Mu, Y., and Song, J. Dynamically-Driven Inactivation of the Catalytic Machinery of the SARS 3C-Like Protease by the N214A Mutation on the Extra Domain. *PLOS Computational Biology*, 7(2): e1001084, 2011.

[283] Lim, L., Shi, J., Mu, Y., and Song, J. Dynamically-driven enhancement of the catalytic machinery of the SARS 3C-like protease by the S284-T285-I286/A mutations on the extra domain. *PLoS ONE*, 9(7), 2014.

[284] Owen, D.R., Allerton, C.M.N., Anderson, A.S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R.D., Carlo, A., Coffman, K.J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K.S., Gibson, S.A., Greasley, S.E., Hurst, B.L., Kadar, E.P., Kalgutkar, A.S., Lee, J.C., Lee, J., Liu, W., Mason, S.W., Noell, S., Novak, J.J., Obach, R.S., Ogilvie, K., Patel, N.C., Pettersson, M., Rai, D.K., Reese, M.R., Sammons, M.F., Sathish, J.G., Singh, R.S.P., Steppan, C.M., Stewart, A.E., Tuttle, J.B., Updyke, L., Verhoest, P.R., Wei, L., Yang, Q., and Zhu, Y. An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science*, 374(6575):1586–1593, 2021.

[285] Hilgenfeld, R. From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. *FEBS Journal*, 281(18):4085–4096, 2014.

[286] Yang, H. and Yang, J. A review of the latest research on Mprotargeting SARS-COV inhibitors. *RSC Medicinal Chemistry*, 12(7):1026–1036, 2021.

[287] Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Pujadas, G., and Garcia-Vallvé, S. A review of the current landscape of SARS-CoV-2 main protease inhibitors: have we hit the bull's-eye yet? *International Journal of Molecular Sciences*, 23(1):259, 2022.

[288] Kidera, A., Moritsugu, K., Ekimoto, T., and Ikeguchi, M. Allosteric Regulation of 3CL Protease of SARS-CoV-2 and SARS-CoV Observed in the Crystal Structure Ensemble. *Journal of Molecular Biology*, 433(24):167324, 2021.

[289] Barrila, J., Bacha, U., and Freire, E. Long-range cooperative interactions modulate dimerization in SARS 3CL pro. *Biochemistry*, 45(50):14908–14916, 2006.

[290] Barrila, J., Gabelli, S.B., Bacha, U., Amzel, L.M., and Freire, E. Mutation of Asn28 disrupts the dimerization and enzymatic activity of SARS 3CLpro. *Biochemistry*, 49(20): 4308–4317, 2010.

[291] El-baba, T.J., Lutomski, C.A., Kantsadi, A.L., Malla, T.R., John, T., Mikhailov, V., Bolla, J.R., Schofield, C.J., Zitzmann, N., Vakonakis, I., and Robinson, C.V. Allosteric inhibition of the SARS-CoV-2 main protease - insights from mass spectrometry-based assays. *Angewandte Chemie International Edition*, 59(52):23544–23548, 2020.

[292] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C.D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandaõ-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T.D., Dunnett, L., Fairhead, M., Firth, J.D., Jones, S.P., Keely, A., Keserü, G.M., Klein, H.F., Martin, M.P., Noble, M.E.M., O'Brien, P., Powell, A., Reddi, R., Skyner, R., Snee, M., Waring, M.J., Wild, C., London, N., von Delft, F., and Walsh, M.A. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature Communications*, 11:5047, 2020.

[293] Cantrelle, F.X., Boll, E., Brier, L., Moschidi, D., Belouzard, S., Landry, V., Leroux, F., Dewitte, F., Landrieu, I., Dubuisson, J., Deprez, B., Charton, J., and Hanoulle, X. NMR Spectroscopy of the Main Protease of SARS-CoV-2 and Fragment-Based Screen-

ing Identify Three Protein Hotspots and an Antiviral Fragment. *Angewandte Chemie - International Edition*, 60(48):25428–25435, 2021.

[294] Günther, S., Reinke, P.Y., Fernández-Garciá, Y., Lieske, J., Lane, T.J., Ginn, H.M., Koua, F.H., Ehrt, C., Ewert, W., Oberthuer, D., Yefanov, O., Meier, S., Lorenzen, K., Krichel, B., Kopicki, J.D., Gelisio, L., Brehm, W., Dunkel, I., Seychell, B., Gieseler, H., Norton-Baker, B., Escudero-Pérez, B., Domaracky, M., Saouane, S., Tolstikova, A., White, T.A., Hänle, A., Groessler, M., Fleckenstein, H., Trost, F., Galchenkova, M., Gevorkov, Y., Li, C., Awel, S., Peck, A., Barthelmess, M., Schlünzen, F., Xavier, P.L., Werner, N., Andaleeb, H., Ullah, N., Falke, S., Srinivasan, V., Franca, B.A., Schwinzer, M., Brognaro, H., Rogers, C., Melo, D., Zaitseva-Doyle, J.J., Knoska, J., Penã-Murillo, G.E., Mashhour, A.R., Hennicke, V., Fischer, P., Hakanpaä¨, J., Meyer, J., Gribbon, P., Ellinger, B., Kuzikov, M., Wolf, M., Beccari, A.R., Bourenkov, G., Stetten, D.V., Pompidor, G., Bento, I., Panneerselvam, S., Karpics, I., Schneider, T.R., Garcia-Alai, M.M., Niebling, S., Günther, C., Schmidt, C., Schubert, R., Han, H., Boger, J., Monteiro, D.C., Zhang, L., Sun, X., Pletzer-Zelgert, J., Wollenhaupt, J., Feiler, C.G., Weiss, M.S., Schulz, E.C., Mehrabi, P., Karnĭcar, K., Usenik, A., Loboda, J., Tidow, H., Chari, A., Hilgenfeld, R., Uetrech, C., Cox, R., Zaliani, A., Beck, T., Rarey, M., Günther, S., Turk, D., Hinrichs, W., Chapman, H.N., Pearson, A.R., Betzel, C., and Meents, A. X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science*, 372(6542):642–646, 2021.

[295] Du, R., Cooper, L., Chen, Z., Lee, H., Rong, L., and Cui, Q. Discovery of chebulagic acid and punicalagin as novel allosteric inhibitors of SARS-CoV-2 3CLpro. *Antiviral Research*, 190:105075, 2021.

[296] Komatsu, T.S., Okimoto, N., Koyama, Y.M., Hirano, Y., Morimoto, G., Ohno, Y., and Taiji, M. Drug Binding Dynamics of the Dimeric SARS-CoV-2 Main Protease, determined by Molecular Dynamics Simulation. *Scientific Reports*, 10:16986, 2020.

[297] El Ahdab, D., Lagardère, L., Inizan, T.J., Célerse, F., Liu, C., Adjoua, O., Jolly, L.H., Gresh, N., Hobaika, Z., Ren, P., Maroun, R.G., and Piquemal, J.P. Interfacial Water

Many-Body Effects Drive Structural Dynamics and Allosteric Interactions in SARS-CoV-2 Main Protease Dimerization Interface. *Journal of Physical Chemistry Letters*, 12(26): 6218–6226, 2021.

[298] Carli, M., Sormani, G., Rodriguez, A., and Laio, A. Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations. *Journal of Physical Chemistry Letters*, 12(1):65–72, 2021.

[299] Sztain, T., Amaro, R., and McCammon, J.A. Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease. *Journal of Chemical Information and Modeling*, 61(7):3495–3501, 2021.

[300] Dubanevics, I. and McLeish, T.C. Computational analysis of dynamic allostery and control in the SARS-CoV-2 main protease. *Journal of the Royal Society Interface*, 18 (174):20200591, 2021.

[301] Yuce, M., Cicek, E., Inan, T., Dag, A.B., Kurkcuoglu, O., and Sungur, F.A. Repurposing of FDA-approved drugs against active site and potential allosteric drug-binding sites of COVID-19 main protease. *Proteins: Structure, Function and Bioinformatics*, 89(11): 1425–1441, 2021.

[302] Sheik Amamuddy, O., Afriyie Boateng, R., Barozi, V., Wavinya Nyamai, D., and Tastan Bishop, Ö. Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. *Computational and Structural Biotechnology Journal*, 19:6431–6455, 2021.

[303] Amaral, J.L., Oliveira, J.T., Lopes, F.E., Freitas, C.D., Freire, V.N., Abreu, L.V., and Souza, P.F. Quantum biochemistry, molecular docking, and dynamics simulation revealed synthetic peptides induced conformational changes affecting the topology of the catalytic site of SARS-CoV-2 main protease. *Journal of Biomolecular Structure and Dynamics*, 5: 1–13, 2021.

[304] Bhat, Z.A., Chitara, D., Iqbal, J., Sanjeev, B.S., and Madhumalar, A. Targeting allosteric

pockets of SARS-CoV-2 main protease Mpro. *Journal of Biomolecular Structure and Dynamics*, 2021.

[305] Liang, J., Pitsillou, E., Ververis, K., Guallar, V., Hung, A., and Karagiannis, T.C. Small molecule interactions with the SARS-CoV-2 main protease: In silico all-atom microsecond MD simulations, PELE Monte Carlo simulations, and determination of in vitro activity inhibition. *Journal of Molecular Graphics and Modelling*, 110:108050, 2022.

[306] Menéndez, C.A., Byléhn, F., Perez-Lemus, G.R., Alvarado, W., and de Pablo, J.J. Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease. *Science Advances*, 6(37):1–7, 2020.

[307] Sencanski, M., Perovic, V., Pajovic, S.B., Adzic, M., Paessler, S., and Glisic, S. Drug Repurposing for Candidate SARS-CoV-2 Main Protease Inhibitors by a Novel in Silico Method. *Molecules*, 25(17):3830, 2020.

[308] Novak, J., Rimac, H., Kandagalla, S., Pathak, P., Naumovich, V., Grishina, M., and Potemkin, V. Proposition of a new allosteric binding site for potential SARS-CoV-2 3CL protease inhibitors by utilizing molecular dynamics simulations and ensemble docking. *Journal of Biomolecular Structure and Dynamics*, 2021.

[309] Verma, S. and Pandey, A.K. Factual insights of the allosteric inhibition mechanism of SARS-CoV-2 main protease by quercetin: an in silico analysis. *Biotech*, 11(67), 2021.

[310] ElSawy, K.M., Alminderej, F.M., and Caves, L.S. Disruption of 3CLpro protease self-association by short peptides as a potential route to broad spectrum coronavirus. *Journal of Biomolecular Structure and Dynamics*, 2021.

[311] Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[312] Lahiry, P., Torkamani, A., Schork, N.J., and Hegele, R.A. Kinase mutations in human disease: Interpreting genotype-phenotype relationships. *Nature Reviews Genetics*, 11(1): 60–74, 2010.

[313] Hopkins, A.L. and Groom, C.R. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, 2002.

[314] Cohen, P., Cross, D., and Jänne, P.A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery*, 20(7):551–569, 2021.

[315] Malumbres, M. Cyclin-dependent kinases. *Genome Biology*, 15(6):1–10, 2014.

[316] Tatum, N.J. and Endicott, J.A. Chatterboxes: the structural and functional diversity of cyclins. *Seminars in Cell and Developmental Biology*, 107:4–20, 2020.

[317] Malumbres, M., Sotillo, R., Santamaría, D., Galán, J., Cerezo, A., Ortega, S., Dubus, P., and Barbacid, M. Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell*, 118(4):493–504, 2004.

[318] Grossel, M.J. and Hinds, P.W. Beyond the cell cycle: A new role for cdk6 in differentiation. *Journal of Cellular Biochemistry*, 97(3):485–493, 2006.

[319] Pavletich, N.P. Mechanisms of cyclin-dependent kinase regulation: Structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. *Journal of Molecular Biology*, 287 (5):821–828, 1999.

[320] Merrick, K.A., Larochelle, S., Zhang, C., Allen, J.J., Shokat, K.M., and Fisher, R.P. Distinct Activation Pathways Confer Cyclin-Binding Specificity on Cdk1 and Cdk2 in Human Cells. *Molecular Cell*, 32(5):662–672, 2008.

[321] Merrick, K.A. and Fisher, R.P. Putting one step before the other: Distinct activation pathways for Cdk1 and Cdk2 bring order to the mammalian cell cycle. *Cell Cycle*, 9(4): 706–714, 2010.

[322] Majumdar, A., Burban, D.J., Muretta, J.M., Thompson, A.R., Engel, T.A., Rasmussen, D.M., Subrahmanian, M.V., Veglia, G., Thomas, D.D., and Levinson, N.M. Allostery governs Cdk2 activation and differential recognition of CDK inhibitors. *Nat Chem Biol*, 17(4):456–464, 2021.

[323] Sherr, C.J. and Roberts, J.M. CDK inhibitors: Positive and negative regulators of G1-phase progression. *Genes and Development*, 13(12):1501–1512, 1999.

[324] Besson, A., Dowdy, S.F., and Roberts, J.M. CDK Inhibitors: Cell Cycle Regulators and Beyond. *Developmental Cell*, 14(2):159–169, 2008.

[325] LaBaer, J., Garrett, M.D., Stevenson, L.F., Slingerland, J.M., Sandhu, C., Chou, H.S., Fattaey, A., and Harlow, E. New functional activities for the p21 family of CDK inhibitors. *Genes and Development*, 11(7):847–862, 1997.

[326] Ray, A., James, M.K., Larochelle, S., Fisher, R.P., and Blain, S.W. p27 Kip1 Inhibits Cyclin D-Cyclin-Dependent Kinase 4 by Two Independent Modes. *Molecular and Cellular Biology*, 29(4):986–999, 2009.

[327] Guiley, K.Z., Stevenson, J.W., Lou, K., Barkovich, K.J., Kumarasamy, V., Wijeratne, T.U., Bunch, K.L., Tripathi, S., Knudsen, E.S., Witkiewicz, A.K., Shokat, K.M., and Rubin, S.M. p27 allosterically activates cyclin-dependent kinase 4 and antagonizes palbociclib inhibition. *Science*, 366(6471):eaaw2106, 2019.

[328] Day, P.J., Cleasby, A., Tickle, I.J., O'Reilly, M., Coyle, J.E., Holding, F.P., McMenamin, R.L., Yon, J., Chopra, R., Lengauer, C., and Jhoti, H. Crystal structure of human CDK4 in complex with a D-type cyclin. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4166–4170, 2009.

[329] Takaki, T., Echalier, A., Brown, N.R., Hunt, T., Endicott, J.A., and Noble, M.E. The structure of CDK4/cyclin D3 has implications for models of CDK activation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4171–4176, 2009.

[330] Schachter, M.M., Merrick, K.A., Larochelle, S., Hirschi, A., Zhang, C., Shokat, K.M., Rubin, S.M., and Fisher, R.P. A Cdk7-Cdk4 T-Loop Phosphorylation Cascade Promotes G1 Progression. *Molecular Cell*, 50(2):250–260, 2013.

[331] Anders, L., Ke, N., Hydbring, P., Choi, Y.J., Widlund, H.R., Chick, J.M., Zhai, H., Vidal, M., Gygi, S.P., Braun, P., and Sicinski, P. A Systematic Screen for CDK4/6 Substrates

Links FOXM1 Phosphorylation to Senescence Suppression in Cancer Cells. *Cancer Cell*, 20(5):620–634, 2011.

[332] Dick, F.A. and Rubin, S.M. Molecular mechanisms underlying RB protein function. *Nature Reviews Molecular Cell Biology*, 14(5):297–306, 2013.

[333] Sherr, C.J. and McCormick, F. The RB and p53 pathways in cancer. *Cancer Cell*, 2(2): 103–112, 2002.

[334] Hanahan, D. and Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell*, 144 (5):646–674, 2011.

[335] Malumbres, M. and Barbacid, M. Cell cycle kinases in cancer. *Current Opinion in Genetics and Development*, 17(1):60–65, 2007.

[336] Stone, A., Sutherland, R.L., and Musgrove, E.A. Inhibitors of cell cycle kinases: Recent advances and future prospects as cancer therapeutics. *Critical Reviews in Oncogenesis*, 17(2):175–198, 2012.

[337] VanArsdale, T., Boshoff, C., Arndt, K.T., and Abraham, R.T. Molecular pathways: Targeting the cyclin D-CDK4/6 axis for cancer treatment. *Clinical Cancer Research*, 21 (13):2905–2910, 2015.

[338] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[339] Piezzo, M., Cocco, S., Caputo, R., Cianniello, D., Di Gioia, G., Di Lauro, V., Fusco, G., Martinelli, C., Nuzzo, F., Pensabene, M., and De Laurentiis, M. Targeting cell cycle in breast cancer: CDK4/6 inhibitors. *International Journal of Molecular Sciences*, 21(18): 1–23, 2020.

[340] Lamb, R., Lehn, S., Rogerson, L., Clarke, R.B., and Landberg, G. Cell cycle regulators cyclin D1 and CDK4/6 have estrogen receptor-dependent divergent functions in breast cancer migration and stem cell-like activity. *Cell Cycle*, 12(15):2384–2394, 2013.

[341] Zhong, Z., Yeow, W.S., Zou, C., Wassell, R., Wang, C., Pestell, R.G., Quong, J.N., and Quong, A.A. Cyclin D1/cyclin-dependent kinase 4 interacts with filamin A and affects the migration and invasion potential of breast cancer cells. *Cancer Research*, 70(5):2105–2114, 2010.

[342] Braal, C.L., Jongbloed, E.M., Wilting, S.M., Mathijssen, R.H., Koolen, S.L., and Jager, A. Inhibiting CDK4/6 in Breast Cancer with Palbociclib, Ribociclib, and Abemaciclib: Similarities and Differences. *Drugs*, 81(3):317–331, 2021.

[343] Susanti, N.M.P. and Tjahjono, D.H. Cyclin-dependent kinase 4 and 6 inhibitors in cell cycle dysregulation for breast cancer treatment. *Molecules*, 26(15):4462, 2021.

[344] Scheidemann, E.R. and Shajahan-Haq, A.N. Resistance to CDK4 6 Inhibitors in Estrogen Receptor-Positive Breast Cancer. *International Journal of Molecular Sciences*, 22(22): 12292, 2021.

[345] Modi, V. and Dunbrack, R.L. Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 116(14):6818–6827, 2019.

[346] Schulze-Gahmen, U., De Bondt, H.L., and Kim, S.H. High-Resolution Crystal Structures of Human Cyclin-Dependent Kinase 2 with and without ATP: Bound Waters and Natural Ligand as Guides for Inhibitor Design. *J Med Chem*, 39(23):4540–4546, 1996.

[347] Jeffrey, P.D., Russo, A.A., Polyak, K., Gibbs, E., Hurwitz, J., Massagué, J., and Pavletich, N.P. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*, 376(6538):313–320, 1995.

[348] Schulze-Gahmen, U. and Kim, S.H. Structural basis for CDK6 activation by a virus-encoded cyclin. *Nature Structural Biology*, 9(3):177–181, 2002.

[349] Gibson, T.J., Thompson, J.D., Blocker, A., and Kouzarides, T. Evidence for a protein domain superfamily shared by the cyclins, TFIIB and RB/p107. *Nucleic Acids Research*, 22(6):946–952, 1994.

[350] Russo, A.A., Jeffrey, P.D., Patten, A.K., Massagué, J., and Pavletich, N.P. Crystal structure of the p27(Kip1) cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature*, 382(6589):325–331, 1996.

[351] Wang, B. and Song, J. Structural basis for the ORC1-Cyclin A association. *Protein Science*, 28(9):1727–1733, 2019.

[352] Errico, A., Deshmukh, K., Tanaka, Y., Pozniakovsky, A., and Hunt, T. Identification of substrates for cyclin dependent kinases. *Advances in Enzyme Regulation*, 50(1):375–399, 2010.

[353] Hallett, S.T., Pastok, M.W., Morgan, R.M.L., Wittner, A., Blundell, K.L., Felletar, I., Wedge, S.R., Prodromou, C., Noble, M.E., Pearl, L.H., and Endicott, J.A. Differential Regulation of G1 CDK Complexes by the Hsp90-Cdc37 Chaperone System. *Cell Reports*, 21(5):1386–1398, 2017.

[354] Chen, P., Lee, N.V., Hu, W., Xu, M., Ferre, R.A., Lam, H., Bergqvist, S., Solowiej, J., Diehl, W., He, Y.A., Yu, X., Nagata, A., Vanarsdale, T., and Murray, B.W. Spectrum and degree of CDK drug interactions predicts clinical performance. *Molecular Cancer Therapeutics*, 15(10):2273–2281, 2016.

[355] Van Beusekom, B., Joosten, K., Hekkelman, M.L., Joosten, R.P., and Perrakis, A. Homology-based loop modeling yields more complete crystallographic protein structures. *IUCrJ*, 5:585–594, 2018.

[356] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[357] Markwalder, J.A., Arnone, M.R., Benfield, P.A., Boisclair, M., Burton, C.R., Chang, C.h., Cox, S.S., Czerniak, P.M., Dean, C.L., Doleniak, D., Grafstrom, R., Harrison, B.A., Kaltenbach, R.F., Nugiel, D.A., Rossi, K.A., Sherk, S.R., Sisk, L.M., Stouten, P., Trainor, G.L., Worland, P., and Seitz, S.P. Synthesis and Biological Evaluation of 1-Aryl-4 , 5-

dihydro-1 H -pyrazolo [3 , 4- d] pyrimidin-4-one Inhibitors of Cyclin-Dependent Kinases. *Journal of Medicinal Chemistry*, 47(24):5894–5911, 2004.

[358] Tee, W.V., Guarnera, E., and Berezovsky, I.N. On the Allosteric Effect of nsSNPs and the Emerging Importance of Allosteric Polymorphism. *Journal of Molecular Biology*, 431 (19):3933–3942, 2019.

[359] Guarnera, E. and Berezovsky, I.N. Allosteric drugs and mutations: chances, challenges, and necessity. *Current Opinion in Structural Biology*, 62:149–157, 2020.

[360] Webb, B. and Sali, A. Comparative Protein Structure Modeling using MODELLER. *Current Protocols in Bioinformatics*, 54:5.6.1–5.6.37, 2016.

[361] Echalier, A., Hole, A.J., Lolli, G., Endicott, J.A., and Noble, M.E. An inhibitor's-eye view of the atp-binding site of CDKs in different regulatory states. *ACS Chemical Biology*, 9(6):1251–1256, 2014.

[362] Sava, G.P., Fan, H., Fisher, R.A., Lusvarghi, S., Pancholi, S., Ambudkar, S.V., Martin, L.A., Charles Coombes, R., Buluwela, L., and Ali, S. ABC-transporter upregulation mediates resistance to the CDK7 inhibitors THZ1 and ICEC0942. *Oncogene*, 39(3):651–663, 2020.

[363] Harrod, A., Fulton, J., Nguyen, V.T., Periyasamy, M., Ramos-Garcia, L., Lai, C.F., Metodieva, G., De Giorgio, A., Williams, R.L., Santos, D.B., Gomez, P.J., Lin, M.L., Metodiev, M.V., Stebbing, J., Castellano, L., Magnani, L., Coombes, R.C., Buluwela, L., and Ali, S. Genomic modelling of the ESR1 Y537S mutation for evaluating function and new therapeutic approaches for metastatic breast cancer. *Oncogene*, 36(16):2286–2296, 2017.