Specificity determinants within families of protein – protein interactions

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

2017

Stefan M Ivanov

School of Chemistry

List of contents

Word count: 37 165

List of figures

List of tables

List of equations

Glossary of abbreviations

Abbreviation	**Meaning**
bZIP	Basic leucine zipper domain-containing protein
E1	Ubiquitin-activating enzyme
E2	Ubiquitin-conjugating enzyme
E3	Ubiquitin-ligating enzyme, ubiquitin ligase
NetQ	Electrostatic energy of interaction
ΔSASAnp	Change in nonpolar solvent accessible surface area upon binding
ΔSASApol	Change in polar solvent accessible surface area upon binding
MD	Molecular dynamics
MC	Monte Carlo
QSAR	Quantitative structure-activity relationship
MM-GB(PB)SA	Molecular mechanics – Generalized Born (Poisson-Boltzmann) surface area
FEP	Free energy perturbation
TI	Thermodynamic integration
BAR	Bennett acceptance ratio
MBAR	Multistate Bennett acceptance ratio
RISM	Reference interaction site model
GIST	Grid inhomogeneous solvation theory

Abstract

University: The University of Manchester

Candidate's Name: Stefan M Ivanov

Degree Title: Doctor of Philosophy

Thesis Title: Specificity determinants within families of protein – protein interactions

Protein – protein interactions govern every aspect of the cellular life cycle. Despite the pivotal role of interprotein association, many of its aspects remain poorly understood. This pertains particularly to the specificity determinants in interactions between large families of proteins and in intrafamily interactions. To elucidate the origins of affinity and specificity in paralogous inter- and intrafamily interactions, a series of *in silico* techniques of increasing theoretical sophistication and computational cost were employed on several datasets from key physiological pathways, under the initial assumption that interactions are mediated through a common interface on a conserved steric scaffold.

A large-scale bioinformatics study on all combinations of potential interactors within the examined systems was carried out first, performing side chain replacement on X-ray- and NMR-derived templates to produce up to thousands of models of the various binary interactions within the examined systems. Simultaneously, polar and nonpolar areas, buried upon complexation, and the energy of electrostatic interaction between the binding partners were computed. Comparison of surfaces and energies between interacting and non-interacting pairs, identified from literature, reveals that all three parameters are significantly different between interactors and non-interactors, with electrostatics being most discriminatory of the three interfacial descriptors.

Despite the statistical significance of the separation between binders and non-binders, considerable overlap remains, making any predictions solely based on buried surface and charge interactions unreliable. To probe deeper into the binding process, extensive molecular mechanics – Poisson-Boltzmann surface area calculations were then performed on a medium-sized set of 60 protein – peptide complexes from the Bcl-2-family of proteins – key regulators of the intrinsic apoptotic pathway. Per-residue decomposition of the enthalpy of interaction between the different protein – peptide pairs provides much finer detail on the binding process than the large-scale surface and charge calculations previously performed. This allowed pinpointing where affinity and specificity within the system originate, identification of key interactions, determination of how affinity is dependent on peptide properties, and provided a quantitative estimate of the energetics of binding. Crucially, this work demonstrates that the proteins' per-residue energies can be viewed as an energy fingerprint.

Finally, this point was further developed by performing free energy calculations at a higher level of theory – thermodynamic integration – on eight large, drug and drug-like compounds bound to the Bcl-xL and Bcl-2 proteins. Comparison of the information content provided by energetic fingerprinting with a traditional two-dimensional quantitative structure-activity relationship study demonstrates the added value of free energy calculations. Crucially, this method affords a more comprehensive description of the binding process and every individual protein – ligand/peptide/protein complex, and extends the framework of four-dimensional molecular dynamics - quantitative structure-activity relationships (4D-MD/QSAR). Finally, directions for future work aiming to derive and validate hyperpredictive 4D-MD/QSAR models incorporating ligand- and receptor-based descriptors are set out.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

Verily, at a juncture such as this, one cannot help but look back at the events leading up to the current circumstances and acknowledge the people who moulded those very circumstances into their so very favorable present form. I will now attempt to condense all my gratitude toward those very people in a less-than-astronomical number of words.

I thank Jim Warwicker for granting me with the opportunity of a lifetime; I thank Pete Bond for granting me with my second opportunity of a lifetime; Roland Huber for voluntarily taking upon himself the burden of mentoring me, in- and outside of Science; Chandra Verma for his subtle, yet immensely positive influence; Jan Marzinek, Daniel Holdbrook, Nils Berglund, and Vas Kargas for all their help, work-related and otherwise, and for creating an amazing atmosphere in- and outside of work, an atmosphere I will fondly remember for the rest of my life; Stephen “Foxy” Fox for all his help and encouragement; and Robin Yee and James Baker for their kind assistance with technical problems. On the mancunian side, I thank Joe Flood, Max Hebditch, and Spyros Charonis for being such great friends and for the wonderful memories; Max Hebditch, Pen Richardson, and Paul Mulherin for invaluable, indispensable technical assistance; Jim Warwicker, Jessica Bowler, and Christoph Ballestrem for their confidence and support when I needed confidence and support the most; Sam Hay for his constant positivity; and Nick Fowler for all the pleasant and insightful discussions on physics. Were it not for their kindness, I would not be where I am today.

I also thank Igor Berezovsky and Enrico Guarnera for the valuable, yet painful, lessons I learned, and, finally, Zejun Zheng for his valiant but hopeless efforts in trying to make those lessons less painful. Verily, valor in the face of inevitable doom merits profound distinction.

September 8th 2017, Manchester, UK

About the author

Stefan completed his secondary education in 2005 in Sofia, Bulgaria, retaining and enhancing a strong interest in biology throughout. He obtained a Master's degree in Pharmacy from the Medical University of Sofia in February 2011. Shortly thereafter, he became committed to pursuing a carrier in science. He became enrolled in the Research Attachment Programme at the Agency for Science, Technology, and Research in February 2013, and began active PhD studies in September of the same year.

Rationale for thesis submission in alternative format

Alternative format is optimal for the present thesis, as it consists of three pieces of work, discretized in separate chapters, each one being a continuation and extension of the previous chapter. The goal of understanding and characterizing protein – protein interactions is maintained and pursued throughout the thesis. This work can be thought of as a “computational microscope” focused on protein – protein interactions, with progressive studies increasing the magnification, and providing finer and finer detail on the subject of study. Thus, the thesis forms a coherent body of work. Chapters 2 and 3 have already been published in reputable journals; chapter 4 will be published in due course. The thesis begins with a general introduction on protein – protein interactions, molecular dynamics, and free energy calculations, explaining why all three are important, and highlighting how they relate to each other and to the work performed. Chapters 2, 3, and 4 examine protein – protein and protein – ligand interactions at an increasing level of detail – from the rapid, large-scale, low-resolution characterization of large datasets of protein – protein interactions, through the higher-level molecular mechanics – Poisson-Boltzmann surface area calculations on a medium-sized dataset of 60 protein – peptide complexes, to the computationally expensive, high-level free energy calculations on a small set of 14 protein – ligand complexes. Finally, the conclusions section explicitly connects the three results chapters, summarizes the main findings of the PhD project, and lays out directions for future work.

... everything that living things do can be understood in terms of the jigglings and wigglings of atoms.

Richard Feynman, 1963

Chapter 1 - Introduction

1.1. Protein – protein interactions

Proteins are the most functionally diverse family of biomolecules. In carrying out their numerous functions, many proteins associate, permanently or transiently, with ions or molecules of varying complexity – from simple inorganic ions, e.g. Ca^{2+}, as is the case with the calcium-binding protein calsequestrin (Slupsky et al., 1987), to biopolymers such as DNA (Newman and Keating, 2003), RNA (Gao et al., 2013; Cawley and Warwicker, 2012), or other proteins (Berggård et al., 2007). Some proteins have the ability to bind more than one type of molecule, often simultaneously, as is the case with zinc-finger-domain-containing proteins. These can concurrently bind zinc ions, which stabilize their fold, as well as nucleic acids, whose biological function is regulated by the protein partner (Berg, 1993).

The interplay between biopolymers is critical in directing and maintaining physiological processes. Homo- and heterocomplexes between biopolymers are involved in every aspect of the cellular life cycle – from packaging the genetic material with histones (in eukaryotes) and histone-like proteins (in prokaryotes) (Dorman and Deighan, 2003) to initiating, regulating, and eventually terminating replication, transcription and translation. Permanent associations between biopolymers, such as the ribosome or the proteasome, carry out physiological functions, such as synthesis and degradation of proteins, respectively, whereas transient interactions tend to be nodes in signaling and regulatory pathways.

Genomic studies have shown that most proteins belong to families of evolutionarily (and, typically, functionally) related molecules (van Dijk et al., 2010). One of the more prevalent mechanisms by which the number of proteins in a given family increases is through gene duplication, which produces a pair (or multiple pairs, if more than one gene is duplicated) of proteins that are described as paralogous (Aiello and Caffrey, 2012). Human protein kinases alone number well over 500; ubiquitin-ligating enzymes also number in the hundreds (Markson et al., 2009). It is believed that these came about via large- and small-scale genetic duplications (Manning et al., 2002). Many key regulatory networks are mediated by the particular case of paralogous inter- and intrafamily protein –

protein interactions, where members of one protein family interact with members of another family or with each other, respectively (Figure 1.1).

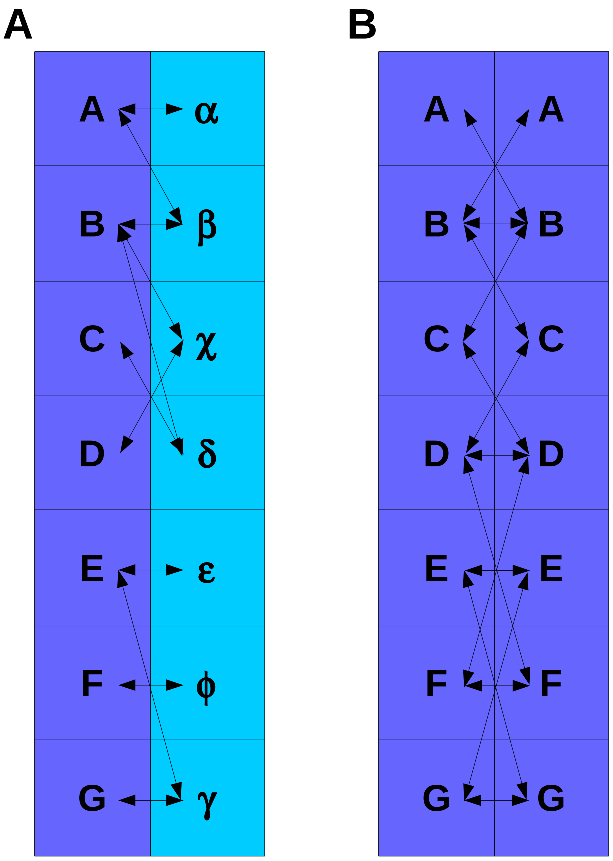


Figure 1.1. Example schematic of paralogous protein – protein interactions. (A) Schematic of paralogous interfamily interactions. Members of one family of proteins, labeled A, B, C, etc., interact with members of another, labeled α, β, χ, etc. An interaction is designated with a two-headed arrow. In the example schematic, protein A interacts with proteins α and β; protein B interacts with proteins β, χ, δ; protein C interacts only with δ, etc. Note that this is a simplified schematic of interfamily interactions, as the two families can differ in size. Moreover, often not all interactions and non-interactions are known. A further complication may be the presence of false positive and/or false negative interactions in published studies. **(B)** Schematic of paralogous intrafamily interactions. Members of one protein family, labeled A, B, C, etc., interact among each other. An interaction is

designated with a two-headed arrow. In the example schematic, protein A interacts with protein B; protein B interacts with itself, protein A, and protein C; C interacts with B and D, etc. As in **(A)**, not all interactions and non-interactions may have been identified and/or there may be false positive and/or false negative interactions.

In this work, the subject of the origins of affinity and specificity in such systems is approached from the standpoint of interface conservation, i.e. a common interface is assumed to exist in between the different pairs of interactors. The mode of binding is presumed to be similar, if not identical, being mediated on a conserved steric scaffold. Moreover, it is presumed that non-interacting pairs also form a similar “interface.” The following protein families are exemplary of such protein – protein interactions.

1.1.1. Basic leucine zipper transcription factors

Basic leucine zipper domain-containing proteins (bZIPs) are eukaryotic transcription factors that control development and stress responses (Gamboa-Meléndez et al., 2013). It is currently believed that the number of bZIPs in the human proteome is 53 (Newman and Keating, 2003). The specificity of bZIP homo- and heterodimerization/oligomerization encodes in itself DNA binding selectivity, i.e. the different bZIP pairings have different DNA-binding specificity. This generates DNA binding diversity, which, in turn, leads to phenotypic diversity. Dimerization occurs via the leucine zipper – a sequence of characteristic heptad repeats in which leucine typically occurs at seven-residue intervals (Carrillo and Privalov, 2010). Human and yeast bZIPs have between 1 and 5 such repeats. Also involved in dimerization are up to 35 more residues located N- and C-terminally to the zipper (Newman and Keating, 2003).

1.1.2. Toxin – antitoxin pairs in bacteria

Many bacteria produce proteinaceous toxins, which they release into their environment (Stanker et al., 2013). Interestingly, many bacterial species need a mechanism to protect themselves from the toxins they produce. A wide-spread mechanism is the toxin – antitoxin system (Dalton and Crosson, 2010). In it, the toxin gene is located next to an antitoxin gene in a single operon, and the two are expressed together. The toxin is a protein, while the antitoxin can be either a protein or a mRNA (Unterholzner et al., 2013). Three types of toxin – antitoxin systems are known to exist in bacteria. In the most common one – type II – both molecules are proteins (Leplae et al., 2011). In most type II toxin – antitoxin systems, the antitoxin proteins interact on a 1:1 basis with their corresponding toxin proteins, i.e. each toxin is recognized and neutralized only by its cognate antitoxin, but not any of its paralogues in the genome (Ahidjo et al. 2011; Dalton and Crosson, 2010).

1.1.3. Bcl-2-family proteins

The intrafamily interactions among the Bcl-2-like proteins are critical in determining whether or not a cell undergoes apoptosis (Okamoto et al., 2012). The Bcl-2 family encompasses the antiapoptotic members Bcl-2, Bcl-xL, Bcl-w, Mcl-1, and A1, as well the proapoptotic Bax, Bad, Bak, and Bid, among others. All family members are characterized by the presence of at least one Bcl-2 homology (BH) domain. The antiapoptotic ones have four such sequences (BH1-4), as do the proapoptotic Bax, Bak, and Bok (Kvansakul et al., 2008). The four BH domains in these proteins form a characteristic hydrophobic groove, which can bind the BH3 helix of the different family members. The BH3 domain is present in all Bcl-2-family members and is central to their apoptosis-regulating activity (Kelekar and Thompson 1998; Lee et al., 2011). Current understanding of apoptosis posits that in preapoptotic cells, proapoptotic proteins are sequestered by antiapotpotic ones. When an apoptosis signal reaches the cell, the proapoptotic proteins dissociate from their antiapoptotic partners and bind via the BH3 domain to the hydrophobic groove on the surface of Bak and Bax. Bak and Bax become activated, oligomerize, and form pores in the outer mitochondrial membrane, which releases cytochrome *c* upon permeation (Suzuki et al., 2000). That, in turn, activates the cellular caspases and results in apopotosis.

1.1.4. Ubiquitin-conjugating – ubiquitin-ligating enzymes

Protein – protein interactions mediate posttranslational modifications such as phosphorylation (Skerker et al., 2008), acetylation, ubiquitination and sumoylation (Cui et al., 2013). The latter two often determine the fate of the proteins involved and, like other posttranslational modifications, regulate physiological processes. The transfer of free ubiquitin to a protein substrate in the cell occurs through a complex series of interactions. First, ubiquitin, a small protein of 74 amino acid residues, (Hunt, 1977) is activated by a ubiquitin-activating enzyme (E1). Next, the activated-ubiquitin – E1 complex is recognized by a ubiquitin-conjugating enzyme (E2). The E2 accepts the activated ubiquitin from the E1 and is, in turn, recognized by a ubiquitin ligase (E3). The E3 enzyme transfers the ubiquitin molecule to a protein targeted for ubiquitination (Kar et al., 2012). The number of enzymes in the system increases downstream by over an order of magnitude - only two E1 enzymes have been identified in human, whereas the E2s number in the tens, and the E3s – in the hundreds (Kamadurai et al., 2009), Furthermore, many pathogenic bacteria deliver virulence factors that act as E3 enzymes in order to "hijack" the cellular machinery (Lin et al, 2012). Ubiquitination can be carried out in a number of ways which direct the protein for proteasomal degradation or signal cell cycle progression or DNA repair (Chen et al., 2013). Polyubiquitinated proteins are targeted for proteasomal degradation (Wojcikiewic et al., 2003), whereas monoubuiquitinated ones are involved in various cellular functions, e.g. regulation of gene expression (monoubiquinated histones (Osley et al., 2006)) and cell signaling (Zhang et al., 2013).

1.2. Methods for the detection of protein – protein interactions

Information on protein – protein interactions can be obtained from high-throughput and low-throughput assays. The high-throughput ones allow the identification of a large number of interactions in a short period of time, albeit with greater uncertainty. Low-throughput methods provide greater detail and confidence about the protein – protein interactions, although they can detect a limited number of interactions at a time. Following is an overview of the genetic, biochemical, and physical methods used to identify and/or validate protein – protein interactions.

1.2.1. Genetic methods

1.2.1.1. Yeast two-hybrid method (Y2H)

The yeast two-hybrid (Y2H) system was first proposed in the early 1990s (Chien et al., 1991). It is based on the circumstance that many transcription factors have a two-active-domain structure (Latchman, 1997), one domain being responsible for binding to a promoter (labeled binding domain, BD), and the other – for activating transcription (labeled activating domain, AD). If the two domains are separated and fused to two distinct proteins, a physical interaction between the latter would potentially bring together the domains necessary for transcription, which, in turn, would manifest itself in the activation of a specific reporter gene. Currently, the most widely used reporter gene is *LacZ*, which encodes the β-galactosidase enzyme (Shoemaker and Panchenko, 2007). The fusion of the BD and AD domains to separate polypeptides is achieved via DNA recombinant technologies – plasmids, containing the domains and proteins of interest, are constructed (Berggård et al., 2007). The basic principle can be subjected to numerous modifications, making it applicable to a wide variety of tasks – from identifying interactions in between sets of paralogs (van Wijk et al., 2009) to identification of interacting pairs within an entire proteome (Li et al., 2004). Its versatility, relatively low cost, *in vivo* applicability, amenability to numerous tasks and modifications, and high-throughput (Berggård et al., 2007) have made the Y2H assay “the system of choice for detecting protein – protein interactions” (Legrain and Selig, 2000). However, it is not devoid of shortcomings, the principal one being the fairly high rate of false positives and negatives, which can be attributed to a number of factors – differences in protein folding or post-translational modifications between yeast and higher organisms, inherent insensitivity to non-binary interactions, and difficulties with membrane proteins, which, when expressed as fusion proteins, tend to aggregate within the yeast nucleus. Y2H results must be further verified with literature searches or more sensitive methods (Shoemaker and Panchenko, 2007).

1.2.1.2. Phage display

Because of its high throughput, phage display is currently a prominent method for identifying protein – protein interactions on a large scale. It is particularly well suited to the needs of immunologists and cell biologists, who are especially interested in protein – peptide interactions (Kushwaha et al., 2013). In essence, phage display is the cloning of a DNA strand into the coding strand of a coating protein of a bacteriophage (Bábíčková et al., 2013). The cloning is performed so that the protein of interest is inserted in between the N- and C-termini of the original coating protein. Expression of the bait protein on the surface of a virion/virus enables it to interact with a high number of proteins, present in the surrounding medium. Originally, a filamentous phage was used (Smith, 1985), but subsequent modifications of the technique have also employed other expression vectors and/or particles. At present, whole DNA libraries can be constructed, where each phage carries a different protein/peptide (Bábíčková et al., 2013). A bait protein can be immobilized and screened against an entire phage library – the interacting phages remain attached to the bait, while non-interacting ones are easily washed away. This makes phage display well suited to the search for protein/peptide drugs/vaccines (Hamzeh-Mivehroud et al., 2013) and in antibody diagnostics (Hairul Bahara et al., 2013). Phage display can also be adapted to the search for small molecule drugs (Takakusagi et al., 2010).

1.2.1.3. Gene coexpression

Another way to infer functional and physical interactions between proteins is through studying gene expression and identifying genes that are consistently expressed together. This approach is greatly facilitated by whole-genome sequencing, RNA sequencing, and the use of microarrays (Tohge and Fernie, 2012). Studies have shown that interacting proteins are coexpressed with a greater degree of correlation than random, non-interacting pairs (Jansen, 2002; Stuart et al., 2003). The highest correlation has been observed between genes coding permanent complexes, e.g. the proteasome. It has also been shown that the coexpressed genes coevolve (Shoemaker and Panchenko, 2007).

1.2.2. Biochemical methods

1.2.2.1. Tandem affinity purification (TAP)

The tandem affinity purification (TAP) method was initially developed to rapidly isolate native protein complexes from yeast cells (Rigaut et al., 1999). It was further extended to isolation and purification of a single complex, if so desired (Puig et al., 2001). The technique is also known as TAP tag because it involves tagging a protein with two IgG binding domains of protein A (ProtA), isolated from *Staphylococcus,* and a calmodulin binding peptide (CBP). These are separated by a tobacco etch virus (TEV) protease cleavage site (Rigaut et al., 1999). The target protein can be tagged C- or Nterminally. Once the tag is fused to the bait protein, the two are introduced into the host cell, where the bait interacts with its native binders under physiological conditions. It is recommended that protein expression is sustained at or near natural levels, otherwise over-expressed proteins may associate with non-native partners (Puig et al., 2001). In TAP tagging, the desired complex(es) is (are) isolated in two consecutive steps, which increases selectivity and reduces non-specific interactions in comparison with single-step methods, e.g. the yeast two-hybrid method (Li, 2011). In the original set-up of Rigaut and coauthors, the protein(s) - bait – CBP – ProtA complex is immobilized onto an IgG matrix. After washing-out the contaminants, the TEV cleavage site is split with the TEV protease, thus releasing the protein(s) – bait – CBP complex. The latter subsequently binds to a calmodulin resin, from which the protein(s) – bait complex is finally released through washing and elution (Rigaut et al., 1999). It is then ready for subsequent analysis, typically a mass spectrometric one. The technique has also been modified and applied to higher eukaryotes, where protein recovery is on average much lower than in yeast. This is said to be one of the limitations of the TAP tag. Over 30 different tags have been developed, some of them applicable in mammalian cells and specifically designed to overcome this problem (Li, 2011). Nevertheless, TAP tagging, like the yeast two-hybrid test, produces many false positives and negatives (Shoemaker and Panchenko, 2007). Wherever possible, results should be further verified.

1.2.2.2. Co-immunoprecipitation (co-IP)

The co-immunoprecipitation technique is based on retrieving a protein complex with an antibody, specific to one of the components in the complex (Ngounou-Wetie et al., 2013). This allows the identification of previously unknown binders to a known protein. Co-IP is a popular technique for identifying physiologically relevant interactions. Depending on the system under study, it can be used to detect interactions *in vivo* or *in vitro*. The antibodies may be immobilized onto a solid phase such as magnetic microbeads or agarose or may be added directly to the protein mixture. The use of antibodies provides high specificity, which is the reason some authors consider it to be one of the best techniques for confirming putative protein – protein interactions (Velasco-García and Vargas-Martínez, 2012). An added benefit of using magnetic beads is the possibility of automatizing the process, which reduces costs and manual labor and makes the technique applicable to high-throughput studies.

1.2.3. Physical methods

1.2.3.1. Mass spectrometry

In the context of protein – protein interactions, mass spectrometry is primarily used to identify the individual components of a complex, rather than to detect complexation, as is the case with TAP tagging or co-immunoprecipitation (Wang et al., 2013a). However, recent advances have seen its successful implementation to the study of a particularly challenging area of the protein interactome – membrane proteins (Schey et al., 2013). Also, the applicability of the technique has been extended to large-scale studies of protein – protein interactions (Ewing et al., 2007). The core of the technique is an analysis of compounds based upon a mass/charge ratio. Proteins are converted to a gas phase and ionized using one of two major approaches – electrospray ionization or matrix assisted laser desorption ionization. In the mass spectrometer, the gaseous ions pass through a magnetic field, which deflects them from their straight-line trajectory to an extent depending on their mass/charge ratio. The ions reach a detector within the mass spectrometer, which discerns between the different particles and their deflection from the original trajectories. Several algorithms have been developed to analyze the retrieved spectra and to convert these into compound composition or sequence, as is the case with proteins. Intact protein complexes, including membrane ones, have also been studied with mass spectrometry (Schey et al., 2013).

1.2.3.2. Nuclear magnetic resonance (NMR) spectroscopy

In NMR spectroscopy, the shift in resonance frequency of atomic nuclei in a magnetic field, relative to a standard, is measured and used to deduce the three-dimensional structures of biological macromolecules and their complexes (Cavalli et al., 2007; Guntert, 2009). Most NMR studies in structural biology are performed in solution, which has the distinct advantage of avoiding crystallization artifacts. Solid state NMR (ssNMR) is less prevalent, but has the capability of tackling challenging targets, such as membrane proteins, where it has made important contributions (Weingarth and Baldus, 2013). With the advent of in-cell NMR, magnetic resonance spectroscopy is becoming even more attractive, as it is now able to provide structural information on proteins in their native state within the living cell, where macromolecular crowding affects protein NMR detectability (Wang et al., 2011). NMR spectroscopy has also been applied to the study of protein folding and misfolding across a wide range of timescales, and has even been used to observe the emergence of the nascent protein from the ribosome in the course of its synthesis (Waudby et al., 2013).

1.2.3.3. X-ray crystallography

X-ray crystallography provides atomic-level detail on the structure of small molecules, macromolecules, and macromolecular complexes by measuring how they scatter X-rays when arranged in a crystal lattice (Su et al., 2015). It yields detailed information on ion, ligand, and cofactor binding to proteins (Moraes et al., 2014). Despite its high resolution, X-ray crystallography has several distinct disadvantages. Determining a protein's 3D structure requires a great amount of purified protein and its complete sequence (Palmer and Niwa, 2003). Growing a protein crystal, and in particular a membrane-protein one, is a challenging task. Moreover, it is not the only bottleneck – protein overexpression, purification, and stabilization have also proved difficult for numerous proteins (Hunter et al., 2011). Finally, the obtained structures may have crystallization artifacts, as, due to the very nature of the technique, structures are resolved from a crystal, which is very different from the well hydrated state in *in vitro* experiments, and the macromolecular crowding in the intracellular environment. Crystallization artifacts may also appear due to sample contamination (Niedzialkowska et al., 2016). Nevertheless, X-ray crystallography presently dominates the field of structural biology (Shi, 2014).

1.2.3.4. Transmission electron microscopy (TEM) and cryo-electron microscopy (cryo-EM)

As their names imply, in electron microscopy techniques, samples are irradiated with a beam of electrons to produce 2- or 3-dimensional images. A notable shortcoming of transmission electron microscopy is its low resolution, which limited the applicability of electron microscopy in the field of structural biology. It was later found that cryoprotection of samples with liquid nitrogen enhances their stability and improves resolution (Nogales, 2016). This has made the technique extremely versatile – from providing a near-atomic-resolution structure of different membrane proteins (Henderson et al., 1990; Collins et al., 2017) to solving the structures of entire virus particles (Fernandez-Leiro and

Scheres, 2016; Liu et al., 2016). This versatility and success in resolving highly complex or challenging structures are fueling a cryo-EM-based revolution, presently ongoing in structural biology (Callaway, 2015).

1.3. Databases

The vastness of protein – protein interaction data necessitates its systematization and classification. Therefore, numerous databases have been created to facilitate the storage, retrieval, and usage of information on protein – protein interactions. Only the most commonly used databases will be briefly reviewed.

1.3.1. Protein Data Bank (PDB)

The Worldwide Protein Data Bank (wwPDB) (Berman et al., 2003) and its member organizations - The Protein Data Bank in Europe (PDBe) (http://www.ebi.ac.uk/pdbe/) (Velankar et al., 2010; Velankar et al., 2012), the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) (Berman et al., 2000), and the Protein Data Bank Japan (PDBj) (http://pdbj.org/) (Kinjo et al., 2012) are the repository where newly-solved macromolecular structures are deposited. The great number of protein – protein structures in the PDB has made the data bank an authoritative resource on protein – protein interactions in its own right. Moreover, the multitude of structures it contains and homology models derived thereof are a starting point for studies, aiming to gain mechanistic understanding of the workings of complex biomacromolecules. Finally, the great and ever increasing number of single-macromolecule- and macromolecular-complex structures in the PDB can be used to model and construct protein – protein interaction networks on a large scale (Tuncbag et al., 2017; Ivanov et al., 2017). The PDB is a primary database, i.e. it does not source its content from other databases. It is, rather, a source of input for several secondary or derivative databases, such as the Structural classification of proteins (SCOP) or Structural classification of protein – protein interfaces (SCOPPI, see below), which generate their contents from entries in the PDB.

As of September 2017, the PDB contains over 132 000 3D structures of proteins and complexes between proteins and/or other biopolymers or ligands of low molecular weight. Around 90% of the structures have been solved by X-ray diffraction, around 9% - by NMR, around 1% - by electron microscopy, with the remaining structures solved by other techniques or some combination of techniques. As well as freely available, curated, and annotated structural information, the PDB contains information on the functional features of the respective proteins, web based tools for the visualization of the 3D structures, tools aiding drug design, and cross-references and links to other databases. Recently, it has been argued that deposition of a 3D structure into the PDB should be made at the time of submission of the corresponding article for peer review, to ensure prepublication validation and the quality of the PDB archive (Joosten et al., 2013).

1.3.2. Uniprot

Another major web-based resource is the Universal Protein Knowledgebase - Uniprot (http://www.uniprot.org/). Uniprot was formed in 2003 by the merger of several smaller databases with the aim to "provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information" (Apweiler et al., 2004). Uniprot provides information on protein sequence, structure and structural features, binding partners, functional characteristics of the protein and its location in the genome, and has cross-references and links to other databases. The Uniprot Knowledgebase (UniprotKB) has a manually curated section – Swiss-prot - as well as an automatically annotated one - TrEMBL. The latter is not subject to author review and curation. TrEMBL, however, can be very useful in the absence of experimentally validated data. Uniprot has two additional sections – Uniref and Uniparc, which contain sequence clusters and archives, respectively.

1.3.3. Structural classification of protein – protein interfaces (SCOPPI)

Specifically dedicated to the matter of systematizing the numerous protein – protein interactions within the proteome by structural features is the SCOPPI database (http://scoppi.biotec.tu-dresden.de/scoppi/). SCOPPI characterizes and classifies the interfaces between interacting proteins by geometric criteria (Winter et al., 2006). At present, SCOPPI has over 4000 distinct types of interfaces. The database provides information on all aspects of the interprotein interaction – the interface type, the atoms and residues from each binding partner involved, and their respective counts, the type of the interaction (permanent or transient), an assessment of its strength, the contact surface area (in Å^2), the contact volume (in Å^3), and the change in accessible surface area upon complexation (ΔASA). A query in SCOPPI produces all the domains which interact with a given query domain in table form, systematized by domain – domain interaction. Also provided are sequence alignments of the concerned domains and the various aspects of the interdomain interface.

1.3.4. Search tool for recurring instances of neighboring genes (STRING)

STRING is a web-based resource which infers functional association between two genes and their respective products on the basis of colocalization in genomes. A functional association between the encoded proteins, hence, may imply a physical one (Snel et al., 2000). From its launch in 2000 to the present day, STRING has greatly increased its coverage of genomes and proteins (von Mering, 2003; von Mering et al., 2005; von Mering et al., 2007; Jensen et al., 2009; Szklarczyk et al., 2011; Franceschini et al., 2013). The database contains and accounts for a large number of experimentally verified interprotein interactions. Interactions are visualized in a graph-like manner, in which each protein (gene) is a node and each interaction is an edge. Predicted interactors are also listed below the graph. STRING can be accessed at http://string-db.org/.

1.3.5. Structural Classification of Proteins (SCOP)

SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) categorizes proteins in a hierarchical fashion according to the structural motifs present in them. The domain is treated as the unit of classification in SCOP. From that starting point, the evolutionary and functional relationships between different proteins can be extracted using SCOP (Murzin et al., 1995). Domains are grouped by species and classified into families, superfamilies, folds and classes (Lo Conte et al., 2002), resulting in the following SCOP hierarchy levels: Species, Protein, Family, Superfamily, Fold and Class (Andreeva et al., 2008). SCOP also integrates data from and cross referencing with multiple other databases at the level of domain entries, facilitating characterization of known and novel protein structures (Andreeva et al., 2004).

1.4. Intermolecular forces and interactions

1.4.1. Coulombic interactions

In the context of protein – protein interactions, it is worth briefly considering what forces operate in between bound proteins, and in molecular recognition in general. A common approach is to divide intermolecular interactions into electrostatic ones, operating between charged particles, and van der Waals interactions between noncharged species. Permanent charges, i.e. electric monopoles, interact with each other in accordance with Coulomb's law, which stipulates that the force between two charged particles *i* and *j* is inversely proportional to the square of the distance between them (r^{-2}):

$$F = \frac{q_i q_j}{4\pi\varepsilon r^2}, \quad equation \quad 1.1.$$

where q_i and q_j are the respective partial charges of particles *i* and *j*, *r* is the Euclidean distance between them, and ε is the dielectric permittivity. This type of interactions underpin ionic bonding in ionic crystals (Pitzer, 1960).

1.4.2. Multipole interactions

Forces involving permanent or induced dipoles/multipoles have a similar functional form, but a different exponent in the distance term, depending on the type of species involved (ion – dipole, dipole – dipole, etc.) and whether the dipoles/multipoles are fixed or freely rotating. For example, charge – fixed dipole interactions have an r^{-3} dependence; freely rotating dipole – fixed dipole interactions have an r^{-4} dependence; Keesom forces operate between freely rotating, permanent di-, quadri-, and multipoles and are inversely proportional to the seventh power of distance (r^{-7}); Debye forces operate between a freely rotating dipole/multipole and an induced multipole (r^{-7}); London dispersion forces arise as a result of the formation of instantaneous multipoles in nonpolar moieties (r^{-7}) (Israelachvili, 2011). Additionally, interactions involving di-/multipoles depend on their relative orientations. Different authors give differing definitions of the term “van der Waals forces.” It is universally agreed that van der Waals forces include London dispersion interactions; some authors also include Keesom and Debye forces as well (French, 2000). What is universally accepted is their key characteristics - they are weak, nondirectional, nonsaturatable, additive, and short-ranged, the latter being the result of the high power dependence of distance. Although van der Waals interactions are short-ranged (< 6 Å) and orders of magnitude weaker than covalent bonds (0.5 – 2 kcal/mol *vs* ~ 200 kcal/mol), their multitude and additivity make them important determinants of molecular properties and bonding. They give rise to the so-called “hydrophobic” effect, where water encloses hydrophobic solutes in a “cage” of hydrogen-bonded water molecules to maximize solvent hydrogen bonding and minimize solute – solvent contact area (Chandler, 2005). This also minimizes the loss of rototranslational freedom water molecules undergo. At room temperature, addition of nonpolar solutes in a high-dielectric medium is associated with an energetically favorable enthalpic term (positive ΔH), stemming from van der Waals interactions, outweighed by an unfavorable entropic term ($T\Delta S < 0$, i.e. $\Delta S < 0$ – a loss of entropy). The hydrophobic effect is considered to be the driving force behind protein folding (Pace and Shirley, 1996).

It must be stressed that there is no single, unambiguous classification of intermolecular forces. Rather, a force that has the same physical origin may be long-ranged as well as short-ranged, isotropic or directional, strong or weak. Few of the forces and phenomena are unequivocal in nature and origin. For example, the interactions between charges, ions, permanent di-, quadri-, octopoles, etc. are purely electrostatic in origin, arising from the Coulomb force (Israelachvili, 2011). Conversely, although van der Waals forces are typically thought of as “nonelectrostatic,” fundamentally, they also arise from electron behavior or, more precisely, from the asymmetric distribution of electron density in atoms or molecules. Some authors view van der Waals forces as a quantum mechanical phenomenon (Israelachvili, 2011), despite the fact that simple expressions for the van der Waals force between macroscopic objects have long been derived (Hamaker, 1937). Indeed, several theories of varying sophistication for the calculation of van der Waals forces have been proposed, some accounting for many-body interactions and retardation, others ignoring them (Derjaguin, 1934; Hamaker, 1937; Lifshitz, 1956; Langbein, 1970). Moreover, often there is no sharp divide between the “different types” of interactions, the different categorizations often being chosen and used for convenience, rather than theoretical rigor. What follows is a brief overview of the interactions, most often discussed in studies of intermolecular recognition and association.

1.4.3. Hydrogen bonds, salt bridges, ion – π interactions, π – π interactions, and halogen bonds

The hydrogen bond is a type of electrostatic interaction that also exhibits features of a covalent bond – it is directional and involves a fixed number of participants, i.e. can be thought of as having a valence. Hydrogen bonds form between a lone electron pair of an electronegative atom and a hydrogen atom, bonded to another electronegative atom. Hydrogen bonds are typically below 3 Å in length between donor and acceptor (Torshin et al., 2002) and usually vary in strength between 2 and 7 kcal/mol (Larson and Mcmahon, 1984), making them stronger than van der Waals interactions, but weaker than covalent bonds. Hydrogen bonds have a more complicated distance-dependence that can be roughly approximated as r^{-2} (Israelachvili, 2011). Their multitude and directionality make them important determinants of structure in proteins and nucleic acids, as well as in macromolecular association. Moreover, water molecules and ions often form bridging hydrogen bonds with two or more residues simultaneously. Bridging waters and ions in the interfaces between macromolecules have been noted to be important contributors to affinity and specificity in certain protein – protein associations (Wojdyla et al., 2012). Another common interaction in and in between proteins is the salt bridge, where a protonated, positively charged residue enters into an electrostatic interaction with a deprotonated, negatively charged residue. Typically, under physiological conditions (T ~ 300 K, pH ~ 7.4), arginine, lysine, glutamic or aspartic acid are involved, although in certain cases histidine and tyrosine can also participate in salt bridges. Salt bridges can be viewed as a type of hydrogen bond or being part-hydrogen bond and are comparable in strength (3 – 7 kcal/mol) to the typical hydrogen bond found in proteins, although they can be somewhat longer – up to 4 Å (Anderson et al., 1990). When two salt bridges become conjugated through a common, central residue, a salt-linked triad is formed. The coupled salt bridges elicit a synergistic effect on each other – each has a greater energy when the other is absent, i.e. the free energy reduction upon forming a triad is greater than the sum of the individual, uncoupled salt bridges (Horovitz et al., 1990). Salt-linked triads and other cooperative interactions, involving three or more residues, have been shown to contribute to protein folding (Horovitz and Fersht, 1992) and protein – protein recognition and binding (Ivanov et al., 2016).

The diversity of monomers, making up biopolymers, and forces, operating in between them, translates into an even greater diversity of interactions within and in between those very polymers. Apart from the aforementioned hydrogen bonds and salt bridges, common interactions in the case of biomacromolecules are cation – π, anion – π, and π – π (aromatic) stacking interactions. All three are underpinned by the presence of an electric quadrupole in aromatic systems – an increase in electron density above and below the plane of the aromatic ring and a decrease in the plain itself. The quadrupole's partial charges can then interact with the (partial) charges of other moieties (Ma and Dougherty, 2012). Cation – π interactions can involve amino acid side chains, as well as alkali and alkaline metals. They are particularly strong (~ 5 – 15 kcal/mol, depending on distance and relative orientation), compared to other non-bonded interactions, as forming them incurs a high desolvation penalty for only one of the partners – the cation – as opposed to salt bridges, where both moieties need to be desolvated for the bridge to be formed (Gallivan and Dougherty, 2000). Cation – π interactions are largely electrostatics driven and have a roughly r^{-2} distance dependence, with binding free energies correlating well with the electrostatic potential of the aromatic ring and the charge density of the cation (Ma and Dougherty, 2012; Mecozzi et al., 1996; Dougherty, 1996). Anion – π interactions are comparable in strength and behavior to cation – π interactions (Gil-Ramirez et al., 2008), but seem to appear less often in protein structures deposited in the data banks (Gromiha et al., 2009; Gromiha et al., 2011). They are, however, often observed in inter- and intramolecular interactions involving drug and drug-like compounds (Wang and Wang, 2013).

Stacking interactions are underpinned by dispersion forces, complemented to a varying degree of electrostatics (Huber et al., 2014). The two basic orientations are T-shaped stacking, where the planes of the aromatic rings are normal to each other, and parallel stacking. In the former case, the positive partial charges on the hydrogen atoms of one aromatic ring interact with the π electron cloud of the other ring, lying at a nearly 90 degree angle. In the case of parallel stacking of two benzene rings, one right below the other (sandwiched stacking), an energetically unfavorable overlap of π electron clouds leads to a displacement of one ring relative to the other (shifted stacking). In the trivial case of two benzene molecules, the T-shaped and shifted stacking orientations are energy minima, separated by a saddle point (sandwiched stacking) and connected by multiple intermediate points, corresponding to different relative orientations between the molecules (Sinnokrot et al., 2002). When more complex compounds are involved, the energy landscape becomes more complicated, depending on the substituents (Hunter and Sanders, 1990) and dipole moments of the compounds (Huber et al., 2014). Base-stacking interactions are critical to DNA stability (Yakovchuk et al., 2006). Moreover, stacking contributes significantly to protein folding, stability, and binding and, consequently, is frequently observed in protein crystal structures (McGaughey et al., 1998). Figure 1.2 below exemplifies the types of interactions discussed previously.

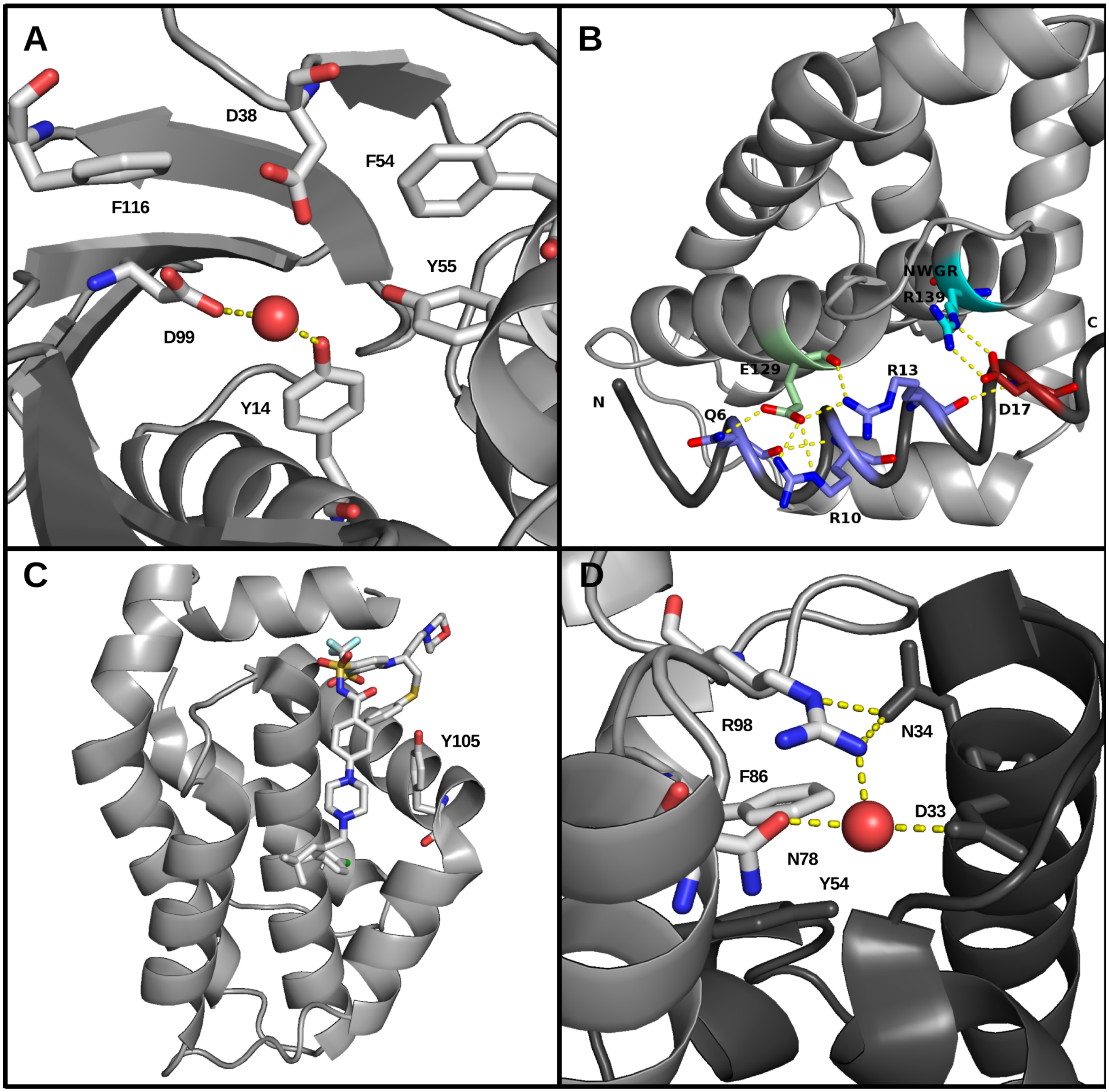


Figure 1.2. Types of intra- and intermolecular interactions. (A) Hydrogen bonding, anion – π, and π – π stacking interactions in the crystal structure of steroid Δ^5-3-isomerase from *Comamonas*

testosteroni, PDB ID: 8CHO (Cho et al., 1998). Oxygen atoms from D99 and Y14, in stick representation, make intermolecular hydrogen bonds, shown as dotted lines, to a water molecule in sphere representation. The oxygen atoms from D38 are equidistant from the phenyl rings of F54 and F116 and are involved in anion – π bonding with them (Cho et al., 1998). F54 is also involved in an intramolecular T-stacking interaction with Y55. Backbone atoms are in cartoon representation, key residues are in stick representation and explicitly labeled. **(B)** Intermolecular hydrogen bonding, salt-bridging, and a salt-linked triad identified from an all-atom simulation of the Bcl-xL protein, shown in gray, bound to the Bad peptide, shown in dark gray (Ivanov et al., 2016). The nitrogen atom of the Q6 side chain of Bad forms a hydrogen bond, shown as a dotted line, to the side chain oxygen of E129 of Bcl-xL. D17 of Bad forms a salt bridge with R139 of Bcl-xL, as do R10 and R13, interacting with E129 and forming a salt-linked triad; salt bridges are also designated with dashed lines; also labeled are the peptide termini. R139 belongs to the NWGR motif, characteristic of Bcl-xL and its homologs. **(C)** π – π stacking interactions in the crystal structure of Navitoclax (ABT-263) bound to Bcl-2, PDB ID: 4LVT (Souers et al., 2013). ABT-263 forms a parallel intramolecular stacking interaction, as well as an angled, T-shaped stacking with Y105 of Bcl-2. Backbone atoms are in cartoon representation, ABT-263 and Y105 are in stick representation with carbon atoms in white, oxygen in red, nitrogen in blue, sulfur in yellow, and fluorine in gray. **(D)** Hydrogen bonding, π – π stacking, a cation – π interaction, and a bridging water molecule, located in the interface between the *E. coli* ceaB toxin, shown in dark gray, and its cognate antitoxin, shown in gray, from the 3U43 crystal structure (Wojdyla et al., 2012). D33 from the toxin, and R98 and the side chain oxygen of N78 from the antitoxin make hydrogen bonds, marked with dashed lines, to a water molecule, in sphere representation, located in the interface between the two proteins. Moreover, the oxygen atom of N34 is involved in intermolecular hydrogen bonding, labeled with dashed lines, with the side chain of R98; R98 and F86 form an intramolecular cation – π interaction; F86 is also involved in an intermolecular parallel stacking interaction with Y54 from the toxin. Backbones are shown as cartoons, key residues are in stick representation and labeled. Hydrogens are absent from all panels; protein numbering corresponds to canonical Uniprot (Apweiler et al., 2004) numbering for all molecules, except Bad, for which the numbering of the simulated sequence starts from 1 (Ivanov et al., 2016).

An important, but greatly underreported, interaction is the so-called halogen bond. Indeed, only in 2013 did the International Union for Pure and Applied Chemistry (IUPAC) recommend a formal definition, although experimental evidence of halogen bonding has been available for over 200 years and crystallographic evidence of it has been mounting for decades (Cavallo et al., 2014). The IUPAC definition is: “A halogen bond occurs when there is evidence of a net attractive interaction between an electrophilic region associated with a halogen atom in a molecular entity and a nucleophilic region in another, or the same, molecular entity.” (Desiraju et al., 2013). The σ-hole model of halogen bond formation was proposed by Clark and coworkers (2007), who showed that halogens X, covalently bonded to an electronegative atom or moiety R, have a region of higher electron density, forming a belt orthogonal to the covalent bond and centered on the halogen, and a region of lower electron density - a “σ-hole” or “crown” - on the elongation of the covalent bond, distal to R, where the electrostatic potential is frequently positive. The electropositive region, offset by around 1.6 – 1.8 Å from the halogen (Harder et al., 2016), is then free to interact with nucleophilic moieties N (a R-X···N schematic is often employed to represent a halogen bond). It is likely that the tendency to think of halogens as nucleophiles, rather than electrophiles, is the primary reason halogen bonds have gone unrecognized for so long, despite the fact that around 40% of drugs on the market or in clinical trials contain halogens, which are likely to be involved in binding to the target (Ho, 2015). The electron distribution imposes strict geometric restrictions on halogen bonds – the R-X-N angle is typically an almost ideal 180^o, unlike hydrogen bonds, which display a greater angular variance; moreover, like hydrogen bonds, halogen bonds are often smaller in length than the sum of the van der Waals radii of the atoms involved. When biomacromolecules are involved, however, they may impose unusual structural constraints on the bonds, leading to deviations from the ideal geometry (Ho, 2015). The more polarizable the halogen and the more electronegative R is, the more electropositive the σ-hole is. Depending on the potentials of the crown and the nucleophile and their polarizabilities, halogen bonds can range greatly in energy – from -1.9 to - 33 kcal/mol, potentially even more (Politzer et al., 2015).

Sometimes the terms “bromine” or “chlorine” or “iodine” bond are used, but this serves only to specify the halogen. There is no fundamental difference between those interactions, nor between halogen and hydrogen bonds – these all arise from electrostatic interactions between electronegative and electropositive moieties; they differ only in their ideal geometries. Indeed, going back to the

R-X···N model with an electronegative disc, normal to the R-X bond and centered on the halogen, it is clear that a hydrogen nucleus would approach the disc laterally, at a 90° angle to the R-X bond, whereas a nucleophile would approach the disc frontally, from the side of the σ-hole, at a 180° angle to the R-X bond. It is apparent that these are fundamentally electrostatic interactions, which differ only by the participants involved and their geometries.

1.5. Protein – protein interfaces

Deviations from the genetically predetermined innate protein interactome are often deleterious, manifesting themselves in a broad spectrum of conditions, ranging from minor symptoms to debilitating syndromes (Gonzalez and Kann, 2012; Rodriguez-Soca et al., 2010). For example, overexpression of the antiapoptotic proteins Bcl-xL and Bcl-2 overwhelms the cell's proapototic defences, facilitating malignant proliferation (Park et al., 2013). Conversely, mutations that abrogate binding between pro- and antiapotptic proteins shift the cellular balance toward premature cell death, and give rise to or are associated with degenerative diseases (Bouillet et al., 2001; Akhtar et al., 2004). Point mutations in leucine zipper transcription factors can lead to altered dimerization and DNA binding, resulting in a great number of documented malignancies (Rodriguez-Martinez et al., 2017).

The fine-tuning of interaction sequences and pathways is striking, considering the tens of thousands of macromolecules in any given cell and organism at any one time, and the astronomical number of potential combinations of these (Berggård et al., 2007). Even slight deviations from the innate interactome can cause pathology or be lethal. It is natural to then ask “How is this specificity achieved?” An obvious starting point is to look at the interfaces, the parts of the molecules that immediately carry out the interaction(s), and look for the answers there.

Firstly, it must be stressed that considerable differences exist between permanent and transient protein – protein complexes, the latter being the focus of this work. Permanent or obligate complexes, such as the proteasome, are large, cylindrical or globular structures, composed of subunits, which can not exist independently of each other (Mintseris and Weng, 2005). The interfaces in such complexes are large (up to 10 000 Å^2), heavily dominated by hydrophobic residues and resemble the interior of globular proteins. Conversely, transient or nonobligate interactions involve partners that exist in isolation as well as in complex form, and, consequently, have evolved adaptations to the aqueous cellular environment. Thus, interfaces in transient interactions are smaller and less dominated by hydrophobic residues than in obligate ones (Scott et al., 2016), presumably because it is unfavorable to expose hydrophobic moieties to the high-dielectric medium upon complex dissociation (Jones and Thornton, 1996; Sheinerman et al., 2000). Moreover, nonobligate interfaces tend to be less twisted than permanent ones (Ozbabacan et al., 2011).

Subtle differences have been observed between interfaces in transient heterocomplexes and the remainder of the constituent molecules, and these can give hints and leads to better understanding of interprotein interaction and association (Keskin et al., 2008). In the case of enzymes, a good candidate for the binding site is the largest cleft on the molecule's surface. This approach, however, is not universally applicable to protein – protein interactions – dimer binding sites seem to be quite flat (Ma et al., 2003). Structural analyses have shown that a typical interface has a surface of around 1600 Å^2 (Lo Conte et al., 1999), which in most instances is less than the maximal possible buried surface area upon complexation (Ma et al., 2003). Also, interfaces in transient interactions tend to be smaller than in permanent ones (Aiello and Caffrey, 2012). When geometric features are insufficiently discriminatory, a good lead may be amino acid composition. In a study of six types of protein – protein interfaces, Ofran and Rost showed that it is possible to predict interface type from amino acid composition alone. Their results also demonstrated that distinguishable differences in sequence and residue – residue preferences exist between “interactions of residues within the same structural domain and between different domains, between permanent and transient interfaces, and between interactions associating homo-oligomers and hetero-oligomers” (Ofran and Rost, 2003). Conservation of tryptophan and, to a lesser extent, phenylalanine and methionine on the protein surface is strongly suggestive of a binding site. Structural studies have shown that the most common pairs of residues found in protein – protein interfaces involve charged and aromatic side chains, as these can form a multitude of interactions in between in each other – hydrogen bonds, salt bridges, salt-linked triads, cation – π, anion – π, T-shaped stacking, parallel shifted stacking, and bridging interactions (Gromiha et al., 2009; Gromiha et al., 2011). Furthermore, studies have shown that protein binding sites consist of tightly packed, structurally conserved regions – “hot spots” – with a significant energetic contribution to binding. These seem to be particularly enriched in tryptophan, tyrosine and arginine (Ma et al., 2003).

Charged and polar residues constitute a significant portion of protein – protein interfaces in transient complexes, and complex formation results in the burial of a significant number of those

(Sheinerman et al., 2000). A study on a broad spectrum of nonobligate protein – protein complexes revealed that the average or “standard sized” (1600 ± 400 Å^2) interface has around 10 intermolecular hydrogen bonds, that almost every third hydrogen bond involves a charged residue, and that 13% of interfaces involve two charged residues (Lo Conte et al., 1999; Sheinerman et al., 2000). A study on interfaces by Skrabanek and coauthors (2008) demonstrated that the largest or second largest hydrophobic patch on the solvent accessible surface is involved in multimeric interfaces in 90% of the complexes they examined. A division of the interface into a core and rim has been suggested and used to successfully differentiate between biological interfaces and nonspecific crystal packing ones (Keskin et al., 2008; Janin et al., 2008; Chartron et al., 2012). Ma and coauthors suggest that polar residues at the interface cores confer rigidity, which reduces the entropic loss upon binding, while the surrounding residues form a “flexible cushion” (Ma et al., 2003). Finally, in a study of paralogous protein interfaces and their evolution, Aiello and Caffrey find that functionally diverged interfaces possess more subfamily specific residues, which are usually located at the rim. The authors propose that binding affinity is determined mainly at the hub, whereas specificity is determined at the rim (Aiello and Caffrey, 2012).

1.6. Molecular dynamics (MD) simulations and free energy techniques

Unlike the structures deposited in the PDB (Berman et al., 2003) and the homology models (Bordoli et al., 2009) derived thereof, biological molecules, and molecules in general, are not static (Childers et al., 2017). Quite the contrary, they are extremely dynamic, typically highly flexible entities, constantly colliding with each other and their surroundings, in a perpetual, never-ending motion, where particle velocities change on the order of femtoseconds (10^{-15} s or a quadrillionth, a millionth of a billionth of a second), obeying the Maxwell-Boltzmann distribution in equilibria at all times and temperatures. It has long been recognized that for many proteins dynamics are essential to function. More generally, the dynamics of biological macromolecules are often intricately connected to their functions. For example, the “lock-and-key” model of protein – ligand interactions, where a rigid protein binds a rigid ligand, proposed by Emil Fischer in 1894, has been superseded by the induced fit model, proposed by Daniel Koshland in 1958, where a flexible ligand and a flexible protein elicit structural changes in each other, facilitating a mutual fit between the two, and, ultimately, binding

(Koshland, 1994). More recent work on protein – ligand interactions has shown that many proteins sample many different conformations – active and inactive – with agonists and antagonists stabilizing the former and the latter, respectively (Yanamala et al., 2008; Kohlhoff et al., 2014). Many more pertinent examples may be given, so perhaps Richard Feynman's eloquent summary of the subject is best suited to a brief review such as this - “... everything that living things do can be understood in terms of the jigglings and wigglings of atoms.”

Molecular dynamics is the method best suited to the study of those very jigglings and wigglings. More precisely, it offers the highest ratio of model fidelity/computational cost out of any *in silico* technique presently available. Although the behavior of matter is ultimately governed by the principles of quantum mechanics, or perhaps even string theory (Mukhi, 2011), only Newtonian mechanics-based descriptions of biological macromolecules are presently feasible (Hansson et al., 2002; Durrant and McCammon, 2011). This is because the computational workload in quantum mechanical models scales with the number of electrons, typically some high power of it (3 – 10, i.e. $N_{electrons}^{3\text{ - }10}$, potentially even higher), whereas in classical models it scales with the number of atoms and/or ions, usually to the second power ($N_{atoms}^{\sim 2}$), making the former applicable to only very small systems (Young, 2001). Apart from the size of the systems, the other key consideration is the timescale of the processes of interest. Biological processes occur on a vastly ranging timescale – from the femto- to picosecond librations of side chains (He et al., 2011) up to hours, and even more, for folding of certain slow-folding proteins (Schuler and Hofmann, 2013) - a scale spanning more than 18 orders of magnitude. Biological systems of interest involve thousands, up to billions of particles and need to be studied on a timescale of picoseconds to microseconds and more. Currently, tackling such large systems over periods of time exceeding 100 ns is only possible with molecular dynamics. For this reason, it has gained great prominence is the study of protein – small molecule ligand/peptide/protein/other type of biopolymer interactions. Any attempt to review the relevant literature, comprising tens of thousands of papers, if not hundreds of thousands, is doomed to be hopelessly incomplete. What fallows is merely a brief description of the technique highlighting its primary strengths and weaknesses.

In molecular dynamics, systems of interest, such as biological macromolecules or complexes between biological macromolecules and/or ligands/cofactors, are simulated *in silico* to gain insight into their properties or understand or predict experimentally-measurable properties of the molecules in question. Being an atomic resolution or near-atomic resolution method, MD is used to gain mechanistic understanding of processes and phenomena, which other methods are not capable of providing. Molecular dynamics began as a method for the simulation of atomic-scale systems, comprising several dozen atoms, at an atomistic level of resolution (Alder and Waiwright, 1957). The computational power needed to simulate and study even the simplest and fastest of biologically relevant phenomena became available decades later – in 1977, McCammon and coworkers were the first to simulate an entire protein molecule and report protein dynamics on picosecond timescales (McCammon et al., 1977). While it was initially applied to biological systems of several hundred atoms, advances in computing capacity have extended its applicability to particles of a much greater size, e.g. ribosomes and viruses (Ode et al., 2012; Huber et al., 2017).

The starting point for molecular dynamics simulations in structural biology is typically an X-ray or NMR-structure or a homology model of a macromolecule, possibly complexed with (an)other (macro)molecule(s). Depending on the computational resources available and the level of model fidelity desired, the choice between implicit and explicit solvation is made. As the name implies, implicit models dispense with a detailed representation of the solvent and counterions, substituting them with a continuous medium of uniform dielectric permittivity, averaging the solvent degrees of freedom (Patodia et al., 2014). This allows a significant speedup of calculations compared to explicit solvation, where solvent molecules and counterions constitute 80-90% of the particles in the computational system, i.e. the vast majority of computation is spent on species, which typically are of little interest. The advantages of explicit solvation are that it affords a more realistic description of solute – solvent hydrogen bonds and salt bridges and is less prone to sampling unphysical states (Zhang et al., 2017). Moreover, due to the very nature of the model, it allows a more accurate account of bridging water molecules or ions, which often constitute an integral part of protein – ligand and protein – protein interactions (Wojdyla et al., 2012). Typically, the choice of solvation model is made in accord with the choice of potential energy function, the latter of which will govern the progression of the system through time. The potential energy function or “force field” is the mathematical model used to compute the potential energy of a system given a configuration. A typical force field, exemplified by the Amber potential energy function, is given below. The potential energy E_{pot} (also labeled as V) depends on the system configuration, expressed as the Cartesian coordinates of the N constituent particles (Cornell et al., 1995):

$$E_{pot}(r^N) = \sum_{bonds} k_b (r_{ij} - r_{ij0})^2 + \sum_{angles} k_a (\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{n} \frac{V_n}{2} [1 + \cos(n\varphi_k - \gamma_{nk})] + \sum_{i<j} [4\zeta [(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}] + \frac{q_i q_j}{\varepsilon r_{ij}}] \quad equation \;\; 1.2 .$$

A potential energy function has bonded and non-bonded terms. As the names imply, they account for the potential energy arising from chemically bonded and non-bonded atoms, respectively. Bonded interactions, in turn, have terms accounting for bonds between two atoms (bonds term), between three atoms (angles term), and between four atoms (dihedrals term). Non-bonded interactions are computed from van der Waals and electrostatic terms, with or without an explicit hydrogen-bonding term (Weiner et al., 1983). In equation 1.2, the bonds term has a Hooke law representation, with k_b being the force constant (the "stiffness" of a virtual spring, connecting atoms *i* and *j*), r_{ij} being the distance between the atoms, and r_{ij0} being the minimum-energy distance between them. The angles term has an analogous form and interpretation, this time for sets of three bonded atoms – k_a is the force constant, θ is the angle between the atoms, and θ_0 is the minimum-energy angle between them. In the dihedrals term, the summation runs over *k* dihedrals with periodicities of *n*, V_n is the *n*-fold Fourier component (Cornell et al., 1995) (A_{nk} or, equivalently, $V_n/2$ is the dihedral force constant for the *k*th dihedral and the *n*th multiplicity), φ_k is the *k*th dihedral angle, and γ_{nk} is the phase constant for the *k*th dihedral angle and the *n*th multiplicity (Hopkins and Roitberg, 2014). In equation 1.2, van der Waals interactions are represented with the Lennard-Jones potential where ζ is the potential well depth, σ is the zero-potential distance, and r_{ij} is the distance between atoms *i* and *j*. $(\frac{\sigma}{r_{ij}})^{12}$ is a repulsive, short-ranged term that reflects the Pauli exclusion principle, $(\frac{\sigma}{r_{ij}})^{6}$ is an attractive term representing dispersion forces. Van der Waals interactions are typically modeled with the Lennard-Jones 6-12 potential due to its computational efficiency and mathematical simplicity, but there is no physical basis for using a 6-12 potential, and other schemes, such as a 6-8 potential, have been proposed (White, 1997). The electrostatic energy is computed with the familiar Coulomb potential; notably, the dielectric permittivity ε here is distance-dependent (Weiner et al., 1983).

Obtaining the exact values for the parameters in the potential energy function is done in various ways – either from quantum mechanical calculations or by fitting to experimental observables, e.g.

solvation energies (Shirts et al., 2003), enthalpies of evaporation or melting, etc. (Warshel and Lifson, 1970). Reparametrizations and improvements over time have led to several versions or “flavours” of the most popular force fields – Amber (Maier et al., 2015), GROMOS (Schmid et al., 2011), CHARMM (MacKerell et al., 2004), OPLS (Harder et al., 2016), and NAMD (Phillips et al., 2005). It is crucial that any refinements or additions to a force field are done in a consistent fashion to the overall parametrization procedure. Notably, in force fields which use a 6-12 Lennard-Jones potential and are non-polarizable, i.e. where charges are time-invariant, the dispersion term in the 6-12 potential is parameterized so as to account only for London dispersion interactions. Interactions involving permanent (Keesom) and induced (Debye) di-, quadri-, octo-, etc.-pole interactions, are accounted for by accordingly parameterizing atomic partial charges and parameters in the Coulomb potential (Patodia et al., 2014). Due to the difficulties of parameterizing a polarizable force field and the greatly increased computational cost associated with utilizing it, molecular dynamics is presently dominated by non-polarizable force fields that include only monopoles and dipoles, but not higher-order-poles (Maier et al., 2015; Schmid et al., 2011; MacKerell et al., 2004).

Force fields can also be used to perform Monte Carlo (MC) simulations, where conformational space is sampled by making random changes to particle positions by varying, for example, three-atom angles, four-atom (dihedral) angles, etc., rather than using the more deterministic approach of molecular dynamics. Although MC simulations are conceptually simpler than MD simulations (they do not require evaluating forces), they are used less often. The reason is that time is not a factor in traditional MC techniques and time-progression of properties is not evaluated. Thus, MC simulations can be used to compute equilibrium properties, such as solvation energies (Jorgensen and Ravimohan, 1985) or free energies of binding (Essex et al., 1997), but not dynamic properties, such as viscosity, diffusion, phase changes, and kinetics (Young, 2001).

As previously stated, van der Waals interactions decay as the inverse sixth power of the distance between particles (r^{-6}), whereas electrostatic forces decay as the inverse second power (r^{-2}). In practice, this means that van der Waals forces decay rapidly and interactions beyond 10 Å can be ignored, whereas electrostatic interactions must be accounted for in a more sophisticated fashion. The “brute-force” approach of computing the electrostatic potential between all pairs of particles in the computational system would entail a scaling of the computational effort with the second power of the number of atoms (N^2). This is prohibitively expensive for even medium-sized systems (20 000 – 40 000 particles). This necessitates devising a more sophisticated solution, and, indeed, an elegant scheme, tackling both the computational cost and another problem in molecular dynamics simulations – boundary artifacts – has been devised and implemented. The problem of boundary artifacts arises from the inability to simulate macroscopic quantities of matter (~10^{23} particles), where in bulk conditions, only a minute fraction of the constituent particles are within 2 nm of a real phase boundary and is influenced by it. Conversely, in systems which are computationally tractable at present (most often systems with 50 000 – 100 000 particles, occupying a volume smaller than 1000 nm^3), a much greater proportion of the constituent particles is in proximity to a virtual boundary, a virtual vacuum. This would cause excessive surface tension and behavior, appreciably different from bulk conditions, the latter being the desired and emulated state. Moreover, as previously mentioned, evaluating all Coulomb forces even in such systems, around 10^{18} times smaller than macroscopic quantities, is prohibitively difficult. Thus, the scheme of "periodic boundary conditions" has been devised, where the computational system of interest is placed in a space-filling box, surrounded by translated copies of itself (Adams et al., 1979). This tends to introduce a certain amount of order and correlations, beyond what is observed in real liquids. These artifacts, however, have been found to be far less significant and far more acceptable than the artifacts stemming from the excessive surface tension in the equivalent non-periodic systems (Darden et al., 1998). In a periodic system, the problem of the N^2 scaling of Coulomb interactions can then be circumvented by introducing a distance cut-off, below which the electrostatic potential is evaluated as per the Coulomb expression and summed for all pairs of particles. Setting potentials to zero at the cut-off distance introduces discontinuities and rapid changes in forces near the cut-off radius. This is addressed by multiplying the potential by a switch function or by adding to it a shift function, ensuring a smooth change in potential with distance (van der Spoel and van Maaren, 2006). Beyond the cut-off, forces and energies may be estimated in several ways. In the so-called *reaction field* method, the Coulomb interaction is modified by setting the dielectric permittivity of the region beyond the cut-off to a uniform value, such that the potential becomes zero at the cut-off distance (Tironi et al., 1995). Alternatively, in the so-called *lattice-sum* methods, beyond the cut-off, the summation is transformed from a summation in real space to a summation in Fourier space, which converges much more rapidly. The technique was first proposed by Paul Ewald and, consequently, bears his name – Ewald summation (Ewald, 1921). As it scales as $N^{3/2}$ or in certain cases even as N^2, it has been superseded by an alternative technique, similar in spirit, but scaling as $N*log(N)$, named particle-mesh Ewald (PME) (Darden et al., 1993).

Determining whether pairs of particles fall within a certain cut-off is itself an N^2 task. Thus, further algorithmic modifications are needed to achieve a speedup of computations. With the so-called *group cut-off* scheme, two groups of particles have their van der Waals interaction computed only if their geometric centers, centers of mass or some other features fall within the cut-off limit, for example the oxygen atoms of water molecules. This elicits a 9-fold (3 x 3) reduction in computations for water molecules. In the more popular Verlet cut-off scheme, atoms are grouped in dynamic clusters, from which pair lists of interacting particles are constructed. The pair lists are buffered, i.e. the pair-list cut-off is larger than the interaction cut-off (Verlet, 1967). The lists are updated every *n* steps, usually around 10, which reduces the computational cost to $\sqrt[5]{N^3}$ ($N^{3/5}$). Further gains can be made by using twin-range cut-offs, i.e. by calculating van der Waals forces, which vary more slowly than electrostatic forces, less often (de Vlieg et al., 1989).

Once a force field and solvation model are chosen, bearing in mind the timescale of the process under study, particles are randomly assigned initial velocities, in accordance with the Maxwell-Boltzmann distribution for the starting temperature, and the system begins to evolve in time. Forces acting on each particle are obtained as the negative of the derivative of the potential with respect to position, and are then summed to provide the resultant force acting on each particle. Particle positions are then propagated in the direction of the resultant forces, incrementing the simulation time by a predetermined amount – the “time step” – and the entire process is repeated, typically millions of times. The equations of motion are typically solved with the *leap-frog* or *Verlet* algorithms. The leap-frog algorithm uses positions at time t and velocities from half a time step back ($t – \Delta t/2$) to compute particle coordinates at time $t + \Delta t$ and velocities at time $t + \Delta t/2$, i.e. there is a constant offset between positions and momenta of $\Delta t/2$, which are constantly leap-frogging over each other, hence the name (Berendsen et al., 1984). In the Verlet algorithm, positions and accelerations are evaluated synchronously, i.e. there is no offset like in the leap-frog algorithm (Verlet, 1967). The algorithm, as well as the Verlet cut-off scheme and the corresponding Verlet lists, are named after French physicist Loup Verlet, who first implemented them in molecular dynamics simulations, although the integration scheme has been discovered and then rediscovered several times before the work of Verlet (Ziegel et al., 2007). Both the Verlet and leapfrog integrators are computationally efficient, time-reversible, and conserve energy.

One of the key parameters in molecular dynamics simulations is the size of the time step. It has been shown theoretically and confirmed through decades of experience that the time step must be at least an order of magnitude smaller than the fastest process occurring in the system. For typical all-atom simulations at room temperature, this is the vibration of bonds involving hydrogen atoms. In bulk water at room temperature, stretching motions have been shown to occur at a frequency of around 100 THz (teraherz), i.e. around 10^{14} times per second (Perakis et al., 2016). As per the previous requirement, the time step in such simulations is set to 1 femtosecond – 10^{-15} s. A doubling of the time step can be achieved by constraining bonds involving hydrogen (Ciccotti and Ryckaert, 1986), which typically does not compromise the fidelity of the simulations. Further increases in time step length can be obtained by merging several atoms, e.g. the hydrogens with the carbon atom in a -CH_2- group, in a common bead and parameterizing the force field for such larger "building blocks." Thus, the highest-frequency process in such a system is no longer the vibration of bonds involving hydrogen, as these are no longer present, but the vibration of much heavier bead-bead bonds, which is usually 10 – 20 times slower. Coarse-graining the system in this way not only allows an increase in the time step, but also reduces the number of particles by a factor of 2 – 4 with respect to the equivalent atomistic representation. This is the basis of coarse-grained force fields, such as MARTINI (Marrink et al., 2007), which allow a significant increase in the timescales accessible with MD, albeit at the cost of atomistic resolution.

Despite its many appealing features, molecular dynamics is not without deficiencies which must always be carefully considered before, during, and after performing an MD-based study. Over the decades of experience, three main sources of error have been identified (Childers et al., 2017): the potential energy function; the finite, often insufficient length and sampling of the simulations; and the numerical errors arising from the finite-precision representation of coordinates and velocities. Multiple validation studies have been performed, demonstrating that different force fields generally perform rather well in reproducing properties such as secondary structure content in proteins (Cino et al., 2012), chemical shift parameters (Beauchamp et al., 2012; Gu et al., 2014; O'Brien et al., 2016), and solvation energies (Shirts et al., 2003). These studies also demonstrated that no single force field is definitively "better" than the others. Typically, small to insignificant differences in performance are observed, depending on the parameter being tested. Although overall performance is encouraging, this is a result of decades of work in a rather small chemical space – the twenty coded amino acids, eight nucleotides, a small set of sugars, and low-molecular weight compounds. It must be kept in mind that shortcomings have also been identified (O'Brien et al., 2016), and that extra caution must be exercised when parameterizing and examining novel chemical entities.

The error stemming from finite sampling is dependent on the context of the study. For example, a computational search of cryptic pockets in a protein target may not sample all biologically relevant conformations, where such pockets appear, but any pockets already identified are certainly not "wrong," they are valid targets for ligand binding, provided their appearance is not an artifact of the force field or the simulation protocol. In free energy calculations (Kollman, 1993), however, not sampling conformations the macromolecule visits *in vitro* is very likely to lead to a wrong answer, i.e. a significant discrepancy between computational and experimental values for the free energy of a given process. Moreover, even in the hypothetical case of a "perfect" potential energy function and "perfect" sampling, long simulations tend to accrue round-off errors, which are hard to estimate. Finally, it must be stressed that molecular dynamics' greatest weakness is that estimating the error of the method is very hard, practically impossible, from within the method itself (Karplus and McCammon, 2002). In a free energy calculation of a novel ligand, for example, it is very hard to be sure if the ligand is correctly parametrized and all relevant conformations are exhaustively sampled. For instance, despite several successes with simulating novel fluorescent membrane probes, it is still hard to know if the conformation within the membrane obtained from MD is truly biologically relevant (Loura and Ramalho, 2011). For the foreseeable future, at least, experiment will be the ultimate judge of a force field's fidelity and utility. In the unfortunate, but common, case of nonmarginal discrepancies between computational predictions and experimental observations, it can be hard to determine if the discrepancies are a result of force field parameters, insufficient sampling, both, and/or other factors. This initiates the (typically iterative) process of obtaining more and/or longer simulations, potentially optimizing certain parameters, and comparing to experiment. Any error estimates (Flyvbjerg and Petersen, 1989) coming from within a simulation must be treated with great caution. A better estimator is the standard deviation or the standard error of mean between several independent replicas. Small intrareplica error estimates, combined with large interreplica differences, are strongly suggestive of undersampling in the individual trajectories, i.e. the different simulations become trapped in different regions of phase space, most likely adjacent potential wells separated by a low energy barrier (Faraldo-

Gómez et al., 2004).

In the early days of molecular dynamics simulations, the quantum nature of matter was completely disregarded because of limited computational power, which rendered these simulations inapplicable to situations where an explicit account of electrons and their properties was required, e.g. chemical reactions. Subsequently, the development of more sophisticated theoretical treatments and increased computing capabilities have led to the emergence of hybrid models, combining density functional theory (DFT) and molecular mechanics – DFT/MM – where the electronically relevant part of the system is described with DFT, whereas the remainder is described by a classical force field, e.g. in simulating enzymatic reactions (Quesne and de Visser, 2012; de Visser et al., 2014; Li et al., 2017). Thus, atomistic and coarse-grained molecular mechanical and hybrid DFT/MM models have become applicable to a bewildering array of problems, of which only a few particularly notable success stories will be briefly mentioned here.

One area of drug design where molecular dynamics excels and even outperforms experimental techniques is its ability to identify cryptic pockets and allosteric sites in target proteins, which have remained occluded in published structural studies (Durrant and McCammon, 2011; Oleinikovas et al., 2016). One such success story is the identification of a cryptic trench in HIV integrase through all-atom MD simulations (Schames et al., 2004), which ultimately led to the development of raltegravir – the first of a novel class of antiretrovirals – the integrase inhibitors. Frembgen-Kesner and Elcock have shown (2006) that explicit solvent MD simulations reveal a 10 Å shift of the F169 side chain of p38 MAP kinase, exposing a cryptic pocket in the presence of BIRB 796 – a novel ligand, later named doramapimod, which was evaluated in clinical trials, sponsored by Boehringer Ingelheim, for the treatment of inflammatory diseases. Cryptic pockets and allosteric binding sites have been identified in many other (now potentially druggable) proteins, including the ZAP-70 kinase (Huber et al., 2015), β-lactamase, interleukin-2 (IL-2), Polo-like kinase-1 (PLK1) (Oleinikovas et al., 2016), p53 (Wassman et al., 2013), the MDM2 (Tan et al., 2016) and eIF4E oncoproteins (Lama et al., 2015), and the β_1 and β_2 adrenergic receptors (Ivetac and McCammon, 2010).

Another area where experimental techniques struggle, but molecular dynamics excels, is the study of membrane proteins. These are severely underrepresented in the PDB, because the high hydrophobicity of most membrane proteins makes them prone to aggregation and hinders crystallization (Biggin and Bond, 2008). Moreover, obtaining X-ray structures requires crystallizing proteins which evolution has designed to exist in a lipid environment. Thus, molecular dynamics has filled in many gaps in scientific understanding of the field of membrane proteins and biological membranes in general. Simulations have successfully been used to analyze and predict lipid and lipopolysaccharide binding sites on membrane proteins (Ortiz-Suarez and Bond, 2016; Chavent et al., 2016); obtain mechanistic understanding of the negative cooperativity observed in epidermal growth factor receptor signaling (Arkhipov et al., 2014) and oligomerization (Needham et al., 2016); predict kinetics-altering mutations, distal to the binding site, in the adenosine A_{2A} receptor (Guo et al., 2016); unravel the inner workings of multiple immunity-regulating relay systems at the surface of cellular membranes (Berglund et al., 2015; Garzón et al., 2013); and predict the positioning of fluorescent probes in membranes (Loura and Ramalho, 2011).

A second class of physiologically important molecules, underrepresented in the structural data banks due to difficulties with crystallization, are carbohydrates (Sattelle and Almond, 2014a; Sattelle et al., 2015). Again, molecular dynamics has filled many of the gaps left behind by experiment. For example, multi-microsecond all-atom simulations of amylose have been used to calculate helix–coil, glycosidic linkage, and ring exchange (sugar puckering) rates that previous X-ray and NMR studies have likely overlooked (Sattelle and Almond, 2014b). Sugar puckering and polymer shape have been shown to be tightly linked to heparin bioactivity (Sattelle et al., 2013). Heparin is clinically used and commercially available as a lifesaving anticoagulant (Szajek et al., 2016). The cellular glycome constitutes a vast, largely untapped reservoir of drug targets and biomaterials (Gabius et al., 2004; Muthana et al., 2012). The “coming of age” of structural glycomics through computational prediction of carbohydrate conformational populations promises to reach deep into that reservoir, “pulling out” novel chemical entities even in the absence of crystallographic data (Sattelle and Almond, 2014b). Where structural or structure – activity data is available, a synergistic effect is to be expected (Blundell et al., 2013).

Multiscale simulation has enabled the study of much larger and more complex biological systems on the million-to-billion atom scale, where many biological processes occur (Perilla et al., 2015). Large, complex systems that have only recently become tractable are the cell envelopes of Gram-negative bacteria (Boags et al., 2017), ribosomes, and entire viruses, where remarkable gains have been made. Wang and coworkers performed all-atom MD simulations on a 10 million atom rabbit haemorhagic disease virus to improve fitting to the crystal structure, which led to the development of a potential vaccine (Wang et al., 2013b). Analogous work with HIV provides ample opportunity to target the virus in previously unexploited ways (Perilla and Schulten, 2017). Multiscale simulations of ribosomes have shown that base-flipping of the ribosome underlies the mechanism of action of certain antiobiotics. Moreover, multiscale techniques have made important contributions to the understanding of bioenergy systems and biofuels (Perilla et al., 2015); information transfer within DNA through mechanical stress (Sutthibutpong et al., 2016); and the inner workings of ATPases – the enzymes, which produce the “fuel” for all living cells – adenosine triphosphate (Richardson et al., 2014).

A particularly noteworthy subject in molecular dynamics is free energy calculations (Pohorille et al., 2010; Klimovich et al., 2015). In stark contrast to the previous examples, here, molecular dynamics presently lags far behind experimental techniques, as calculating free energies is orders of magnitude slower and much more unreliable than measuring them. Thus, experimental measurements are necessary even if calculations are already available. Given that calculations are much slower, more expensive, unreliable, and warrant in themselves a subsequent experimental study, it would appear that a peculiar, perhaps even paradoxical, situation exists in the field – a low yield, highly unreliable method continues to attract researchers. Yet, this is neither coincidental, nor misguided - free energy calculations are an important tool in the computational chemistry toolbox (Boyce et al., 2009; Kosloff et al., 2011; Sliwoski et al., 2014). When Jorgensen and Ravimohan published the first free energy calculation of perturbing ethane into methanol in 1985 (Jorgensen and Ravimohan, 1985), the high accuracy of the results provided a lot of impetus for further work in the field. It was hoped that free energy calculations and, more broadly, computer-aided drug design (CADD), will rapidly deliver pharmacotherapeutic solutions for a multitude of diseases (Sliwoski et al., 2014). While CADD has certainly facilitated fascinating breakthroughs in modern medicine, as previously pointed out, e.g. in antiretroviral therapy, it has fallen short of the expectations of the eighties and nineties (Bennett et al., 1998; Flower et al., 2010). This is primarily the result of the vastness of the chemical and conformational spaces that need to be traversed in a typical *in silico* campaign and the complexities and ambiguities inherent in parameterizing docking functions and molecular dynamics force fields (Hoffmann et al., 2016; Pan et al., 2014; Homeyer et al., 2014; Meng et al., 2016). Nevertheless, success stories in the field have already provided a taste of the reward that is to come from meticulously parameterizing and improving new and existing drug design tools and enhanced sampling techniques (Doshi and Hamelberg, 2015; Urano and Okamoto, 2015; Hospital et al., 2015; Burusco et al., 2015; Bernardi et al., 2014).

In principle, computed free energies should always equal experimental ones, within the limits of computational error and the inherent uncertainty of the experimental measurement. In practice, however, there exists a multitude of free energy techniques of varying theoretical sophistication and computational cost (Michel et al., 2010). Generally, the reliability of computational results is proportional to the computational cost of the technique, provided that guidelines and good practices are observed (Klimovich et al., 2015). What follows is a brief overview of the most popular techniques in an increasing order of sophistication and computational cost. While most of these are defined to be performed in the NVT ensemble (constant number of particles, volume, and temperature), and yield the Helmholtz free energy (ΔF or ΔA), most biological processes occur under constant pressure and temperature. Thus, simulations are typically performed in the NPT ensemble (constant number of particles, pressure, and temperature), yielding the Gibbs free energy (ΔG), under the (usually well justified) assumption that the pressure-volume work component (pΔV) for most processes of interest is negligible (Gilson and Zhou, 2007). Thus, ΔG is used throughout the following sections.

1.6.1. Linear interaction energy (LIE) method

The LIE method is conceptually the simplest and computationally cheapest of the MD sampling-based methods. It is an end-point or end-state method, i.e. only the bound and unbound states are considered and used to derive the free energy of binding. In the general case of binging between a protein and ligand, the free energy of binding is computed as the energy difference between the protein – ligand complex and the ligand free in solution (Su et al., 2007):

$$\Delta G_{bind} = \beta\left(\langle E^{ele}_{bound} \rangle - \langle E^{ele}_{unbound} \rangle\right) + \alpha\left(\langle E^{vdW}_{bound} \rangle - \langle E^{vDW}_{unbound} \rangle\right) + \gamma , \quad equation \ \ 1.3.$$

where $E^{ele}_{bound}, E^{ele}_{unbound}, E^{vdW}_{bound}, E^{vDW}_{unbound}$ are the electrostatic energies in the bound and unbound state, and the van der Waals energies in the bound and unbound states, respectively, the angle brackets $\langle ... \rangle$ designate Boltzmann-weighted ensemble averages, and α, β, and γ are fitting parameters; these have been shown to be system-dependent (Gilson and Zhou, 2007). A key assumption of this method is that polar and nonpolar contributions to the solvation free energies of small molecules scale linearly with their intermolecular interaction energies and surface areas, respectively. No simulation of the free protein is carried out, as entropic and reorganization effects are implicitly accounted for in the scaled surface term (Almlöf et al., 2004).

1.6.2. Molecular mechanics – Generalized Born (Poisson-Boltzmann) surface area (MM-GB(PB)SA)

MM-GB(PB)SA calculations are an end state, post-processing method that allows for the explicit account of the free-receptor state. Here, gas-phase or molecular mechanics energies, computed from the force field, are combined with solvation free energies, comprising a polar and nonpolar component, computed with the Generalized Born or Poisson-Boltzmann method, and molecular surface (or volume), respectively (Genheden and Ryde, 2015; Tan et al. 2007):

$$\Delta G_{bind}=\left(\langle E_{PL}\rangle-\langle E_P\rangle-\langle E_L\rangle\right)+\left(\langle \Delta G_{PL}^{solv}\rangle-\langle \Delta G_P^{solv}\rangle-\langle \Delta G_L^{solv}\rangle\right)+T\Delta S \quad equation \;\; 1.4 .$$

Here, E_{PL}, E_P, E_L are the gas-phase or molecular mechanics energies of the protein – ligand complex, the protein, and the ligand, respectively, $\Delta G_{PL}^{solv}, \Delta G_P^{solv}, \Delta G_L^{solv}$ are the respective solvation free energies, the angle brackets indicate Boltzmann-weighted ensemble averages, and TΔS is an optionally computed entropic term. The protein and ligand terms may be calculated from the complex trajectory or from separate simulations of those components – the so-called “one trajectory” and “three trajectory” approach (Miller et al., 2012). Entropy may be calculated with normal mode analysis or the quasiharmonic approximation, although these have been shown to have deficiencies, as they do not account for anharmonic effects (Hou et al., 2010). The Poisson-Boltzmann equation provides the electrostatic potential of a macromolecule in ionic solution (Warwicker and Watson, 1982; Warwicker, 1986) and has a solid theoretical justification, although solving it comes at a relatively high computational cost (Paquet and Viktor, 2015). The computationally cheaper Generalized Born model has been developed as an approximation to the Poisson-Boltzmann method. In the GB formalism, polar solvation energies are computed from pairwise summations over charge – charge interactions, scaled in accord with effective atomic burial or the “Born radius” of the atoms (Koehl, 2006).

1.6.3. Free energy perturbation (FEP)

FEP is conceptually the simplest of the so-called “alchemical” transformation methods, which are based on the malleability of the potential energy function (Gumbart et al., 2012). In free energy perturbation, a system of interest, termed the “reference” system or state, e.g. a Na^+ ion bound to a protein, is simulated and its potential energy is evaluated. Simultaneously, the potential energy of the “transformed” system is also evaluated, e.g., a Li^+ ion bound to the same protein, based on the system configurations from the reference simulation. The free energy change of going from the reference to the transformed state (A → B, termed the “forward” transformation) is then evaluated as:

$$\Delta G = G_B - G_A = -k_B T \ln \langle e^{-\frac{E_B - E_A}{k_B T}} \rangle_A, \quad \textit{equation} \ 1.5.$$

where k_B is Boltzmann's constant, T is the absolute temperature, and $\langle ... \rangle_A$ designates a Boltzmann-weighted ensemble average from the simulation of the reference state. The process can be performed in the other, “reverse” direction, as well, giving FEP “directionality.” FEP is thought to have been introduced in 1954 by Robert Zwanzig (Zwanzig, 1954), although much of the theory had already been developed by Lev Landau in 1938 (Landau et al., 1960).

1.6.4. Thermodynamic integration (TI)

Thermodynamic integration allows estimating the free energy difference between more heterogeneous systems than what is typically performed in FEP, e.g. between two congeneric ligands bound to the same macromolecule, albeit at a greater computational cost than FEP. The two states of interest are coupled through a nonphysical coordinate, typically denoted as λ. For $\lambda = 0$, the system is in the reference or starting state, and is described by the corresponding potential energy function (labeled V_0 or V_A). For $\lambda = 1$, the system is in the transformed or end state, which has a corresponding potential (V_1 or V_B). For the intermediate, unphysical, mixed states between 0 and 1, the potential is:

$$V(\lambda)=f(\lambda)V_1+[1-f(\lambda)]V_0,\ \ equation\ \ 1.6.$$

In the most trivial case, f(λ) is simply equal to λ, which is referred to as linear mixing. The free energy difference is derived to be (Steinbrecher et al., 2011):

$$\Delta G=\int_0^1 \langle \frac{\partial V}{\partial \lambda} \rangle_\lambda d\lambda\ \ equation\ \ 1.7.$$

Here, $\langle ... \rangle_\lambda$ denotes a Boltzmann-weighted ensemble average from a simulation sampling at a particular value of λ. The extra sampling at the transformed and intermediate states needed to construct the $\langle \partial U/\partial\lambda \rangle_\lambda vs$ λ curve presents a 10-or-more-fold increase in computation time as compared to a one-state FEP calculation, although the additional simulations allow for a more accurate representation of the transformed and intermediate states, which are then sampled. FEP calculations can also be discretized in smaller windows and then numerically integrated, analogously to TI. However, the other key disadvantage of FEP is that it takes the logarithm of an averaged exponential of energies, making a small number of large $|E_B - E_A|$ values disproportionately influential on the final ΔG result (Pohorille et al., 2010).

1.6.5. Bennett acceptance ratio (BAR) and multistate Bennett acceptance ratio (MBAR)

First proposed in 1976 by Charles H. Bennett, BAR is an extension of FEP where bidirectional transformations are optimized to minimize the variance of ΔG by introducing a function that weights the contributions of neighboring ensembles *i* and *j* between 0 and 1 (Bennett, 1976). The free energy change is obtained by solving numerically the implicit function for the potential difference *ΔV*:

$$\frac{1}{\langle 1+e^{\frac{\Delta V_{ij}-C}{k_b T}} \rangle_i}=\frac{1}{\langle 1+e^{-\frac{\Delta V_{ji}-C}{k_b T}} \rangle_j},\ \ equation\ \ 1.8.$$

where *C* is a system-dependent constant to be determined self-consistently (Klimovich et al., 2015). Depending on the choice of *C*, several less popular “flavors” of BAR have been developed, e.g. the unoptimized Bennett acceptance ratio (UBAR) and the range-based Bennett acceptance ratio (RBAR). The BAR technique has been further extended by Shirts and Chodera (2008) by computing energy differences not only between neighboring λ states, as in BAR, but between all sampled λ states, making use of all available data. This produces a statistically optimal estimator of the free energy, which has been corroborated in benchmarking studies, demonstrating that MBAR yields more accurate free energy estimates than LIE, MM-PBSA, TI, and BAR (Paliwal and Shirts, 2011), although in certain situations TI and BAR can produce results nearly identical to MBAR (Ruiter et al., 2013).

1.7. Aims

Gene duplication is a key mechanism in the expansion and diversification of protein families, leading to an abundance of inter- and intrafamily interactions (see Figure 1.1) in crucial regulatory pathways throughout all domains of life. Subsequent to gene duplication, binding specificity diverges at greatly differing rates, with potential binding partners often being highly similar with only subtle differences between binders and non-binders (Aiello and Caffrey, 2012). The aims of this work are to unravel the specificity determinants in such interactions, i.e. the mechanisms by which a protein selects binding partners from a pool of closely related candidates, and to attain a more comprehensive understanding of paralogous protein – protein interactions - their energetics and how the energetics aspect of binding relates to sequence, structure, and their mutual evolution. These goals are pursued through detailed, atomistic molecular dynamics studies, as well as large-scale bioinformatics analysis, which have offered both fine detail, as well as an overall view of the subject, bringing together previous and newly reported herein findings into a single, comprehensive framework of inter- and intrafamily protein – protein interactions. The Bcl-2-intrafamily interactions, which by now have become a highly targeted interaction in cancer therapy, are examined as a test case in great detail, facilitating drug design in this particular area. More generally, the broad range of regulatory protein – protein interaction systems examined and the binding and specificity framework described herein offer vast scope for future medicinal chemistry research, aiming to manipulate such interactions to a therapeutic end in a myriad of diseases, underpinned by aberrant paralogous protein – protein binding.

1.8. References

Acuner Ozbabacan, S.E. et al., 2011. Transient protein – protein interactions. *Protein Engineering, Design and Selection*, 24(9), pp.635–648.

Adams, D.J., Adams, E.M. and Hills, G.J., 1979. The computer simulation of polar liquids. *Molecular Physics*, 38(2), pp.387–400.

Ahidjo, B.A. et al., 2011. VapC Toxins From Mycobacterium tuberculosis Are Rribonucleases that Differentially Inhibit Growth and Are Neutralized by Cognate VapB Antitoxins. *PloS one*, 6(6), p.e21738.

Aiello, D. and Caffrey, D.R., 2012. Evolution of specific protein-protein interaction sites following gene duplication. *Journal of molecular biology*, 423(2), pp.257–72.

Akhtar, R.S., Ness, J.M. and Roth, K.A., 2004. Bcl-2 family regulation of neuronal development and neurodegeneration. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1644(2), pp.189–203.

Alder, B.J. and Waiwright, T.E., 1957. Phase Transition for a Hard Sphere System. *J. Chem. Phys.*, 27(1957), p.1208.

Almlöf, M., Brandsdal, B.O. and Åqvist, J., 2004. Binding affinity prediction with different force fields: Examination of the linear interaction energy method. *Journal of Computational Chemistry*, 25(10), pp.1242–1254.

Anderson, D.E., Becktel, W.J. and Dahlquist, F.W., 1990. pH-Induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, 29(9), pp.2403–2408.

Andreeva, A. et al., 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic acids research*, 36(Database issue), pp.D419-25.

Andreeva, A. et al., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research*, 32(Database issue), pp.D226-229.

Apweiler, R. et al., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue), pp.D115-9.

Arkhipov, A. et al., 2014. Membrane Interaction of Bound Ligands Contributes to the Negative Binding Cooperativity of the EGF Receptor. *PLoS computational biology*, 10(7), p.e1003742.

Bábíčková, J. et al., 2013. In vivo phage display - A discovery tool in molecular biomedicine. *Biotechnology advances*, 31, pp.1247–1259.

Beauchamp, K.A. et al., 2012. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation*, 8(4), pp.1409–1414.

Bennett, B.C. et al., 1998. Pharmaceuticals and Forests Medicinal Resources of the Tropical Forest: Biodiversity and Its Importance to Human Health. *BioScience*, 48(3), pp.213–214.

Bennett, C.H., 1976. Efficient estimation of free energy differences from Monte Carlo data. *Journal of*

Computational Physics, 22(2), pp.245–268.

Berendsen, H.J.C. et al., 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(1984), pp.3684–3690.

Berg, J.M., 1993. Zinc-finger proteins. *Current Opinion in Structural Biology*, 3(1), pp.11–16.

Berggård, T., Linse, S. and James, P., 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16), pp.2833–42.

Berglund, N.A. et al., 2015. The role of protein-protein interactions in Toll-like receptor function. *Progress in Biophysics and Molecular Biology*, 119(1), pp.72–83.

Berman, H., Henrick, K. and Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. *Nature structural biology*, 10(12), p.980.

Berman, H.M. et al., 2000. The Protein Data Bank. *Nucleic acids research*, 28(1), pp.235–42.

Bernardi, R.C., Melo, M.C.R. and Schulten, K., 2014. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *BBA - General Subjects*, 1850(5), pp.872–877.

Biggin, P.C. and Bond, P.J., 2008. Molecular dynamics simulations of membrane proteins. In *Methods Mol Biol*. pp. 147–60.

Blundell, C.D., Packer, M.J. and Almond, A., 2013. Quantification of free ligand conformational preferences by NMR and their relationship to the bioactive conformation. *Bioorganic and Medicinal Chemistry*, 21(17), pp.4976–4987.

Boags, A. et al., 2017. Progress in Molecular Dynamics Simulations of Gram-Negative Bacterial Cell Envelopes. *The Journal of Physical Chemistry Letters*, 8(11), pp.2513–2518.

Bordoli, L. et al., 2009. Protein structure homology modeling using SWISS-MODEL workspace. *Nature protocols*, 4(1), pp.1–13.

Bouillet, P. et al., 2001. Degenerative Disorders Caused by Bcl-2 Deficiency Prevented by Loss of Its BH3-Only Antagonist Bim. *Developmental Cell*, 1(5), pp.645–653.

Boyce, S.E. et al., 2009. Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. *Journal of Molecular Biology*, 394(4), pp.747–763.

Burusco, K.K. et al., 2015. Free Energy Calculations using a Swarm-Enhanced Sampling Molecular Dynamics Approach. *ChemPhysChem*, 16(15), pp.3233–3241.

Callaway, E., 2015. The Revolution Will Not Be Crystallized. *Nature*, 525, pp.172–174.

Carrillo, R.J. and Privalov, P.L., 2010. Unfolding of bZIP dimers formed by the ATF-2 and c-Jun transcription factors is not a simple two-state transition. *Biophysical chemistry*, 151(3), pp.149–54.

Cavalli, A. et al., 2007. Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23), pp.9615–20.

Cavallo, G. et al., 2014. Halogen Bond: A Long Overlooked Interaction. In: Metrangolo, P., Resnati, G.

(eds) Halogen Bonding I. *Topics in Current Chemistry*, 358, pp.1-17

Cawley, A. and Warwicker, J., 2012. eIF4E-binding protein regulation of mRNAs with differential 5'-UTR secondary structure: a polyelectrostatic model for a component of protein-mRNA interactions. *Nucleic acids research*, 40(16), pp.7666–75.

Chandler, D., 2005. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature*, 437(7059), pp.640–647.

Chartron, J.W., VanderVelde, D.G. and Clemons, W.M., 2012. Structures of the Sgt2/SGTA dimerization domain with the Get5/UBL4A UBL domain reveal an interaction that forms a conserved dynamic interface. *Cell reports*, 2(6), pp.1620–32.

Chavent, M., Duncan, A.L. and Sansom, M.S.P., 2016. Molecular dynamics simulations of membrane proteins and their interactions: From nanoscale to mesoscale. *Current Opinion in Structural Biology*, 40, pp.8–16.

Chen, Z. et al., 2013. Ubiquitination-Induced Fluorescence Complementation (UiFC) for Detection of K48 Ubiquitin Chains In Vitro and in Live Cells. *PloS one*, 8(9), p.e73482.

Chien, C.T. et al., 1991. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21), pp.9578–82.

Childers, M.C. et al., 2017. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.*, 2(1), pp.9–33.

Cho, H.S. et al., 1998. Crystal Structure and Enzyme Mechanism of D5-3-isomerase from Comamonas testosteroni. *Biochemistry*, 37(23), pp.8325–8330.

Ciccotti, G. and Ryckaert, J.P., 1986. Molecular dynamics simulation of rigid molecules. *Computer Physics Reports*, 4(6), pp.346–392.

Cino, E. A., Choy, W. and Karttunen, M., 2012. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *Journal of chemical theory and computation*, 8(8), pp.2725–2740.

Clark, T. et al., 2007. Halogen bonding: the σ-hole. *Journal of Molecular Modeling*, 13(2), pp.291–296

Collins, R.F. et al., 2017. Full-length, Oligomeric Structure of Wzz Determined by Cryoelectron Microscopy Reveals Insights into Membrane-Bound States. *Structure*, 25(5), p.806–815.e3.

Lo Conte, L. et al., 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic acids research*, 30(1), pp.264–7.

Lo Conte, L., Chothia, C. and Janin, J., 1999. The Atomic Structure of Protein-Protein Recognition Sites. *Journal of molecular biology*, 285(5), pp.2177–98.

Cornell, W.D. et al., 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19), pp.5179–5197.

Cui, Z. et al., 2013. Regulation of cardiac proteasomes by ubiquitination, SUMOylation, and beyond. *Journal of molecular and cellular cardiology*, 72(6), pp.32–42.

Dalton, K.M. and Crosson, S., 2010. A Conserved Mode of Protein Recognition and Binding in a ParD-ParE Toxin-Antitoxin Complex. *Biochemistry*, 49(10), pp.2205–15.

Darden, T., Pearlman, D. and Pedersen, L.G., 1998. Ionic charging free energies: Spherical versus periodic boundary conditions. *Journal of Chemical Physics*, 109(24), pp.10921–10935.

Darden, T., York, D. and Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(1993), p.10089.

Derjaguin, B., 1934. Untersuchungen über die Reibung und Adhäsion, IV. *Kolloid-Zeitschrift*, 69(2), pp.155-164

Desiraju, G.R. et al., 2013. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure*

Appl Chem, 85(8), pp.1711–1713

van Dijk, A.D.J. et al., 2010. Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. *PLoS computational biology*, 6(11), p.e1001017.

Dorman, C.J. and Deighan, P., 2003. Regulation of gene expression by histone-like proteins in bacteria. *Current Opinion in Genetics and Development*, 13(2), pp.179–184.

Doshi, U. and Hamelberg, D., 2015. Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics. *Biochimica et Biophysica Acta - General Subjects*, 1850(5), pp.878–888.

Dougherty, D.A., 1996. Cation-π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science*, 271(5246), pp.163–168.

Durrant, J. and McCammon, J.A., 2011. Molecular dynamics simulations and drug discovery. *BMC biology*, 9(71), pp.1–9.

Essex, J.W. et al., 1997. Monte Carlo Simulations for Proteins: Binding Affinities for Trypsin−Benzamidine Complexes via Free-Energy Perturbations. *Journal of Physical Chemistry B*, 101(46), pp.9663–9669.

Ewald, P.P., 1921. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik,* 369(3), pp.253–287.

Ewing, R.M. et al., 2007. Large-scale mapping of human protein – protein interactions by mass spectrometry. *Molecular systems biology*, (89), pp.1–17.

Faraldo-Gómez, J.D. et al., 2004. Conformational Sampling and Dynamics of Membrane Proteins From 10-Nanosecond Computer Simulations. *Proteins: Structure, Function and Genetics*, 57(4), pp.783–791.

Fernandez-Leiro, R. and Scheres, S.H.W., 2016. Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537(7620), pp.339–346.

Flower, D.R. et al., 2010. Computer aided selection of candidate vaccine antigens. *Immunome*

Research, 6(SUPPL. 2), pp.1–16.

Flyvbjerg, H. and Petersen, H.G., 1989. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1), pp.461–466.

Franceschini, A. et al., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue), pp.D808-15.

Frembgen-Kesner, T. and Elcock, A.H., 2006. Computational Sampling of a Cryptic Drug Binding Site in a Protein Receptor: Explicit Solvent Molecular Dynamics and Inhibitor Docking to p38 MAP Kinase. *Journal of Molecular Biology*, 359(1), pp.202–214.

French, R.H., 2000. Origins and Applications of London Dispersion Forces and Hamaker Constants in Ceramics. *Journal of the American Ceramic Society*, 83(9), pp.2117–2146.

Gabius, H.J. et al., 2004. Chemical biology of the sugar code. *ChemBioChem*, 5(6), pp.740–764.

Gallivan, J.P. and Dougherty, D.A., 2000. A computational study of cation-π interactions vs salt bridges in aqueous media: Implications for protein engineering. *Journal of the American Chemical Society*, 122(5), pp.870–874.

Gamboa-Meléndez, H., Huerta, A.I. and Judelson, H.S., 2013. bZIP transcription factors in the oomycete phytophthora infestans with novel DNA-binding domains are involved in defense against oxidative stress. *Eukaryotic cell*, 12(10), pp.1403–12.

Gao, S. et al., 2013. Zinc Finger 280B Regulates sGCα1 and p53 in Prostate Cancer Cells. *PloS one*, 8(11), p.e78766.

Garzón, D. et al., 2013. Dynamics of the antigen-binding grooves in CD1 proteins: Reversible hydrophobic collapse in the lipid-free state. *Journal of Biological Chemistry*, 288(27), pp.19528–19536.

Genheden, S. and Ryde, U., 2015. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5), pp.1–13.

Gil-Ramirez, G. et al., 2008. Quantitative Evaluation of Anion-π Interactions in Solution. *Angewandte Chemie - International Edition*, 47(22), pp.4114–4118.

Gilson, M.K. and Zhou, H.-X., 2007. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36, pp.21–42.

Gonzalez, M.W. and Kann, M.G., 2012. Chapter 4: Protein Interactions and Disease. *PLoS Computational Biology*, 8(12), p.e1002819.

Gromiha, M.M. et al., 2011. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome science*, 9(Suppl 1), p.S13

Gromiha, M.M., Yokota, K. and Fukui, K., 2009. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Molecular Biosystems*, 5(12), pp.1779–1786.

Gu, Y., Li, D.-W. and Bruschweiler, R., 2014. NMR Order Parameter Determination from Long MD

Trajectories for Objective Comparison with Experiment. *Journal of Chemical Theory and Computation*, 10(6), pp.2599–2607.

Gumbart, J., Roux, B. and Chipot, C., 2012. Standard binding free energies from computer simulations: What is the best strategy? *Journal of Chemical Theory and Computation*, 9(1), pp.794–802.

Guntert, P., 2009. Automated structure determination from NMR spectra. *European biophysics journal: EBJ*, 38(2), pp.129–43.

Guo, D. et al., 2016. Molecular Basis of Ligand Dissociation from the Adenosine A2A Receptor. *Molecular pharmacology*, 89(5), pp.485–491.

Hairul Bahara, N.H. et al., 2013. Phage display antibodies for diagnostic applications. *Biologicals: journal of the International Association of Biological Standardization*, 41(4), pp.209–16.

Hamaker, H.C., 1937. The London—van der Waals attraction between spherical particles. *Physica*, 4(10), pp.1058-1072

Hamzeh-Mivehroud, M. et al., 2013. Phage display as a technology delivering on the promise of peptide drug discovery. *Drug discovery today*, 18(23–24), pp.1144–1157.

Hansson, T., Oostenbrink, C. and van Gunsteren, W.F., 2002. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2), pp.190–196.

Harder, E. et al., 2016. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of Chemical Theory and Computation*, 12(1), pp.281-296

He, Y. et al., 2011. Evidence of protein collective motions on the picosecond timescale. *Biophysical Journal*, 100(4), pp.1058–1065.

Henderson, R. et al., 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology*, 213(4), pp.899–929.

Ho, P.S., 2015. Biomolecular Halogen Bonds. In: Metrangolo, P., Resnati, G. (eds) Halogen Bonding I. *Topics in Current Chemistry*, 358, pp. 241-276

Hockney, R.W., Goel, S.P. and Eastwood, J.W., 1974. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2), pp.148–158.

Hoffmann, J., Wrabl, J.O. and Hilser, V.J., 2016. The role of negative selection in protein evolution revealed through the energetics of the native state ensemble. *Proteins: Structure, Function, and Bioinformatics*, 84(4), pp.435–447.

Homeyer, N. et al., 2014. Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context. *Journal of Chemical Theory and Computation*, 10(8), pp.1–31.

Hopkins, C.W. and Roitberg, A.E., 2014. Fitting of dihedral terms in classical force fields as an analytic linear least-squares problem. *Journal of Chemical Information and Modeling*, 54(7), pp.1978–1986.

Horovitz, A. et al., 1990. Strength and co-operativity of contributions to surface salt bridges to protein stability. *Journal of Molecular Biology*, 216(4), pp.1031–1044.

Horovitz, A. and Fersht, a. R., 1992. Co-operative Interactions during Protein Folding. *Journal of Molecular Biology*, 224(3), pp.733–740.

Hospital, A. et al., 2015. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8(10), pp.37–47.

Hou, T. et al., 2010. Assessing the Performance of the MM/PBSA and MM/GBSA Methods . I . The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model*, 51(1), pp.69–82.

Huber, R.G. et al., 2014. Heteroaromatic π-Stacking Energy Landscapes. *Journal of chemical information and modeling*, 54(5), pp.1371–1379.

Huber, R.G. et al., 2017. Multiscale molecular dynamics simulation approaches to the structure and dynamics of viruses. *Progress in Biophysics and Molecular Biology*, 128, pp.121–132.

Huber, R.G., Fan, H. and Bond, P.J., 2015. The Structural Basis for Activation and Inhibition of ZAP-70 Kinase Domain. *PLoS Computational Biology*, 11(10), pp.1–16.

Hunt, L.T., 1977. Amino-terminal sequence identity of ubiquitin and the nonhistone component of nuclear protein A24, 74(2), pp.650–655.

Hunter, C.A. and Sanders, J.K.M., 1990. The Nature of π-π Interactions. *Journal of the American Chemical Society*, 112(14), pp.5525–5534.

Hunter, M.S. et al., 2011. X-ray diffraction from membrane protein nanocrystals. *Biophysical journal*, 100(1), pp.198–206.

Israelachvili, J., 2011. *Intermolecular and Surface Forces.* Third edition. Oxford, UK: Elsevier

Ivanov, S.M. et al., 2016. Energetics and dynamics across the Bcl-2-regulated apoptotic pathway reveal distinct evolutionary determinants of specificity and affinity. *Structure*, 24(11), pp.2024–2033.

Ivanov, S.M. et al., 2017. Protein – protein interactions in Paralogues: Electrostatics modulates specificity on a conserved steric scaffold. *PLoS ONE*, 12(10), pp.e0185928

Ivetac, A. and Andrew McCammon, J., 2010. Mapping the druggable allosteric space of g-protein coupled receptors: A fragment-based molecular dynamics approach. *Chemical Biology and Drug Design*, 76(3), pp.201–217.

Janin, J., Bahadur, R.P. and Chakrabarti, P., 2008. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2), pp.133–80.

Jansen, R., 2002. Relating whole-genome expression data with protein - protein interactions. *Genome Res.*, 12(1), pp.37–46.

Jensen, L.J. et al., 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(Database issue), pp.D412-6.

Jones, S. and Thornton, J.M., 1996. Review Principles of protein-protein interactions. *Proceedings of the national Academy of sciences*, 93(1), pp.13–20.

Joosten, R.P. et al., 2013. Timely deposition of macromolecular structures is necessary for peer review. *Acta Crystallographica Section D Biological Crystallography*, 69(12), pp.2293–2295.

Jorgensen, W. and Ravimohan, C., 1985. Monte Carlo simulation of differences in free energies of hydration. *J Chem Phys*, 83(6), pp.3050–3054.

Kamadurai, H.B. et al., 2009. Insights into ubiquitin transfer cascades from a structure of a UbcH5B~Ubiquitin-HECT(NEDD4L) complex. *Molecular cell*, 36(6), pp.1095–102.

Kar, G. et al., 2012. Human Proteome-scale Structural Modeling of E2-E3 Interactions Exploiting Interface Motifs. *Journal of proteome research*, 11(2), pp.1196–207.

Karplus, M. and McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9), pp.646–652.

Kelekar, a and Thompson, C.B., 1998. Bcl-2-family proteins: the role of the BH3 domain in apoptosis. *Trends in cell biology*, 8(8), pp.324–30.

Keskin, O. et al., 2008. Principles of Protein-Protein Interactions: What are the Preferred Ways for Proteins to Interact? *Chemical reviews*, 108(4), pp.1225–44.

Kinjo, A.R. et al., 2012. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic acids research*, 40(Database issue), pp.D453-60.

Klimovich, P. V., Shirts, M.R. and Mobley, D.L., 2015. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5), pp.397–411.

Koehl, P., 2006. Electrostatics calculations: latest methodological advances. *Current Opinion in Structural Biology*, 16(2), pp.142–151.

Kohlhoff, K.J. et al., 2014. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature chemistry*, 6(1), pp.15–21.

Kollman, P., 1993. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chemical Reviews*, 93(7), pp.2395–2417.

Koshland, D.E., 1994. The key–lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.*, 33(510), pp.2375–2378.

Kosloff, M. et al., 2011. Integrating energy calculations with functional assays to decipher the specificity of G protein-RGS protein interactions. *Nature structural and molecular biology*, 18(7), pp.846–53.

Kushwaha, R., Payne, C.M. and Downie, a B., 2013. Uses of phage display in agriculture: a review of food-related protein-protein interactions discovered by biopanning over diverse baits. *Computational and mathematical methods in medicine*, 2013, p.653759.

Kvansakul, M. et al., 2008. Vaccinia virus anti-apoptotic F1L is a novel Bcl-2-like domain-swapped dimer that binds a highly selective subset of BH3-containing death ligands. *Cell death and differentiation*, 15, pp.1564–1571.

Lama, D. et al., 2015. Gating by Tryptophan 73 Exposes a Cryptic Pocket at the Protein-Binding

Interface of the Oncogenic eIF4E Protein. *Biochemistry*, 54(42), pp.6535–6544.

Landau, P.D. and Wang, F., 2004. A new approach to Monte Carlo simulations in statistical physics. *Brazilian Journal of Physics*, 34(2), pp.354–362.

Langbein, D., 1970. Retarded Dispersion Energy between Macroscopic Bodies. *Phys. Rev. B*, 2(8), pp.3371

Landau, Lifshitz and Ross, J., 1960. *Statistical Physics.* Third edition. New York: Elsevier

Larson, J.W. and Mcmahon, T.B., 1984. Gas-Phase Bihalide and Pseudobihalide Ions. An Ion Cyclotron Resonance Determination of Hydrogen Bond Energies in XHY^- Species (X, Y = F, Cl, Br, CN). *Inorganic Chemistry*, 23(14), pp.2029–2033.

Latchman, S., 1997. Transcription Factors: An Overview. *Journal of Cell Biology*, 29(12), pp.1305-1312

Lee, E.F. et al., 2011. Crystal structure of a BCL-W domain-swapped dimer: implications for the function of BCL-2 family proteins. *Structure (London, England: 1993)*, 19(10), pp.1467–76.

Legrain, P. and Selig, L., 2000. Genome-wide protein interaction maps using two-hybrid systems. *FEBS letters*, 480(1), pp.32–6.

Leplae, R. et al., 2011. Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. *Nucleic acids research*, 39(13), pp.5513–25.

Li, S. et al., 2004. A map of the interactome network of the metazoan C. elegans. *Science (New York)*, 303(5657), pp.540–3.

Li, X.X. et al., 2017. Reactivity Patterns of (Protonated) Compound II and Compound I of Cytochrome P450: Which is the Better Oxidant? *Chemistry - A European Journal*, 23(26), pp.6406–6418.

Li, Y., 2011. The tandem affinity purification technology: an overview. *Biotechnology letters*, 33(8), pp.1487–99.

Lifshitz, E.M., 1954. The theory of molecular attractive forces between solids. *Soviet physics*, 2(1), pp.73-83

Lin, D.Y., Diao, J. and Chen, J., 2012. Crystal structures of two bacterial HECT-like E3 ligases in complex with a human E2 reveal atomic details of pathogen-host interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6), pp.1925–30.

Liu, Z. et al., 2016. Structure determination of a human virus by the combination of cryo-EM and X-ray crystallography. *Biophysics Reports*, 2(2–4), pp.55–68.

Loura, L.M.S. and Ramalho, J.P.P., 2011. Recent developments in molecular dynamics simulations of fluorescent membrane probes. *Molecules*, 16(7), pp.5437–5452.

Ma, B. et al., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp.5772–7.

Ma, J.C. and Dougherty, D.A., 2012. The Cation-π Interaction. *Chemical reviews*, 97(5), pp.1303–

1324.

MacKerell, A.D., Feig, M. and Brooks, C.L., 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation. *Journal of Computational Chemistry*, 25(11), pp.1400–1415.

Maier, J.A. et al., 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), pp.3696–3713.

Manning, G. et al., 2002. The Protein Kinase Complement of the Human Genome. *Science (New York, N.Y.)*, 298(5600), pp.1912–34.

Markson, G. et al., 2009. Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. *Genome Research*, 19(10), pp.1905–1911.

Marrink, S.J. et al., 2007. The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry. B*, 111(27), pp.7812–24.

McCammon, J.A., Gelin, Bruce R. and Karplus, M., 1977. Dynamics of folded proteins. *Nature*, 267, pp.585–590.

McGaughey, G.B., Gagnes, M. and Rappe, A.K., 1998. π-Stacking Interactions: Alive and Well in Proteins. *Journal of Biological Chemistry*, 273(25), pp.15458–15463.

Mecozzi, S., West, A.P. and Dougherty, D.A., 1996. Cation-π interactions in simple aromatics: Electrostatics provide a predictive tool. *Journal of the American Chemical Society*, 118(9), pp.2307–2308.

Meng, Y. et al., 2016. Transition path theory analysis of c-Src kinase activation. *Proceedings of the National Academy of Sciences*, 113(33), pp.9193–9198.

Mering, C. V., 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1), pp.258–261.

von Mering, C. et al., 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(Database issue), pp.D433-7.

von Mering, C. et al., 2007. STRING 7 - recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, 35(Database issue), pp.D358-62.

Michel, J., Foloppe, N. and Essex, J.W., 2010. Rigorous free energy calculations in structure-based drug design. *Molecular Informatics*, 29(8–9), pp.570–578.

Miller, B.R. et al., 2012. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.*, 8(9), pp.3314–3321.

Mintseris, J. and Weng, Z., 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31), pp.10930–10935.

Moraes, I. et al., 2014. Membrane protein structure determination - The next generation. *Biochimica et* *biophysica acta*, 1838(1PA), pp.78–87.

Mukhi, S., 2011. String theory: a perspective over the last 25 years. *Classical and Quantum Gravity*, 28(15), p.153001.

Murzin, A.G. et al., 1995. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of molecular biology*, 247(7), pp.536–540.

Muthana, S.M., Campbell, C.T. and Gildersleeve, J.C., 2012. Modifications of glycans: Biological significance and therapeutic opportunities. *ACS Chemical Biology*, 7(1), pp.31–43.

Needham, S.R. et al., 2016. EGFR oligomerization organizes kinase-active dimers into competent signalling platforms. *Nature communications*, 7, p.13307.

Newman, J.R.S. and Keating, A.E., 2003. Comprehensive Identification of Human bZIP Interactions with Coiled-Coil Arrays. *Science*, 300(5628), pp.2097–101.

Ngounou Wetie, A.G. et al., 2013. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and molecular life sciences : CMLS*.

Niedzialkowska, E. et al., 2016. Protein purification and crystallization artifacts: The tale usually not told. *Protein Science*, 25(3), pp.720–733.

Nogales, E., 2016. The development of cryo-EM into a mainstream structural biology technique. *Nat Meth*, 13(1), pp.24–27.

O’Brien, E.S., Wand, A.J. and Sharp, K.A., 2016. On the ability of molecular dynamics force fields to recapitulate NMR derived protein side chain order parameters. *Protein Science*, 25(6), pp.1156–1160.

Ode, H. et al., 2012. Molecular dynamics simulation in virus research. *Frontiers in microbiology*, 19(3), pp.1–9.

Ofran, Y. and Rost, B., 2003. Analysing Six Types of Protein–Protein Interfaces. *Journal of Molecular Biology*, 325(2), pp.377–387.

Okamoto, T. et al., 2012. Sheeppox Virus SPPV14 Encodes a Bcl-2-Like Cell Death Inhibitor That Counters a Distinct Set of Mammalian Proapoptotic Proteins. *Journal of virology*, 86(21), pp.11501–11.

Oleinikovas, V. et al., 2016. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *Journal of the American Chemical Society*, 138(43), pp.14257–14263.

Osley, M.A., Fleming, A.B. and Kao, C.-F., 2006. Histone ubiquitylation and the regulation of transcription. *Results and problems in cell differentiation*, 41(February), pp.47–75.

Pace, C.N. and Shirley, B.A., 1996. Forces contributing proteins of proteins. *The FASEB Journal*, 10(1), pp.75–83.

Paliwal, H. and Shirts, M.R., 2011. A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *Journal of Chemical Theory and Computation*, 7(12), pp.4115–4134.

Palmer, R. a and Niwa, H., 2003. X-ray crystallographic studies of protein-ligand interactions. *Biochemical Society transactions*, 31(Pt 5), pp.973–9.

Pan, A.C. et al., 2014. Assessing the Accuracy of Two Enhanced Sampling Methods Using. *J Chem Theory Comput*, 10(7), pp.2860–2865.

Paquet, E. and Viktor, H.L., 2015. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Research International*, 3(4), pp.1-18

Park, D. et al., 2013. Novel small-molecule inhibitors of Bcl-XL to treat lung cancer. *Cancer Research*, 73(17), pp.5485–5496.

Patodia, S., Bagaria, A. and Chopra, D., 2014. Molecular Dynamics Simulation of Proteins: A Brief Overview. *Journal of Physical Chemistry and Biophysics*, 4(6), pp.4–7.

Perilla, J.R. et al., 2015. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology*, 31, pp.64–74.

Perilla, J.R. and Schulten, K., 2017. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nature communications*, 8(19), p.15959.

Pitzer, K.S., 1960. The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry. *Journal of the American Chemical Society*, 82(15), pp.4121–4121.

Pohorille, A., Jarzynski, C. and Chipot, C., 2010. Good Practices in Free-Energy Calculations. *Journal of Physical Chemistry B*, 114(32), pp.10235–10253.

Puig, O. et al., 2001. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods (San Diego, Calif.)*, 24(3), pp.218–29.

Quesne, M.G. and de Visser, S.P., 2012. Regioselectivity of substrate hydroxylation versus halogenation by a nonheme iron(IV)-oxo complex: possibility of rearrangement pathways. *Journal of Biological Inorganic Chemistry*, 17(6), pp.841–852.

Richardson, R.A. et al., 2014. Understanding the apparent stator-rotor connections in the rotary ATPase family using coarse-grained computer modeling. *Proteins: Structure, Function and Bioinformatics*, 82(12), pp.3298–3311.

Rigaut, G. et al., 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(October), pp.1030–1032.

Rodriguez-Martinez, J.A. et al., 2017. Combinatorial bZIP dimers define complex DNA-binding specificity landscapes. *In review*, pp.1–29.

Rodriguez-Soca, Y. et al., 2010. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *Journal of proteome research*, 9(2), pp.1182–90.

Ruiter, A. De, Boresch, S. and Oostenbrink, C., 2013. Comparison of Thermodynamic Integration and Bennett Acceptance Ratio for Calculating Relative Protein – Ligand Binding Free Energies. *Journal of Computational Chemistry*, 34(12), pp.1024–1034.

Sattelle, B.M. et al., 2015. Proteoglycans and Their Heterogeneous Glycosaminoglycans at the Atomic Scale. *Biomacromolecules*, 16(3), pp.951–961.

Sattelle, B.M. and Almond, A., 2014a. Microsecond kinetics in model single- and double-stranded amylose polymers. *Physical chemistry chemical physics : PCCP*, 16(17), pp.8119–8126.

Sattelle, B.M. and Almond, A., 2014b. Shaping up for structural glycomics: A predictive protocol for oligosaccharide conformational analysis applied to N-linked glycans. *Carbohydrate Research*, 383(100), pp.34–42.

Sattelle, B.M., Shakeri, J. and Almond, A., 2013. Does microsecond sugar ring flexing encode 3D-shape and bioactivity in the heparanome? *Biomacromolecules*, 14(4), pp.1149–1159.

Schames, J.R. et al., 2004. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.*, 47(8), pp.1879–1881.

Schey, K.L., Grey, A.C. and Nicklay, J.J., 2013. Mass spectrometry of membrane proteins: a focus on aquaporins. *Biochemistry*, 52(22), pp.3807–17.

Schmid, N. et al., 2011. Definition and testing of the GROMOS force field versions 54A7 and 54B7. *European Biophysics Journal*, 40(7), pp.843–856.

Schuler, B. and Hofmann, H., 2013. Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Current Opinion in Structural Biology*, 23(1), pp.36–47.

Scott, D.E. et al., 2016. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, 15(8), pp.533–550.

Sheinerman, F.B., Norel, R. and Honig, B., 2000. Electrostatic aspects of protein-protein interactions. *Current opinion in structural biology*, 10(2), pp.153–9.

Shi, Y., 2014. A glimpse of structural biology through X-ray crystallography. *Cell*, 159(5), pp.995–1014.

Shirts, M.R. et al., 2003. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *Journal of Chemical Physics*, 119(11), pp.5740–5761.

Shirts, M.R. and Chodera, J.D., 2008. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12), p.124105.

Shoemaker, B.A. and Panchenko, A.R., 2007. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS computational biology*, 3(3), p.e42.

Sinnokrot, M.O., Valeev, E.F. and Sherrill, C.D., 2002. Estimates of the Ab Initio Limit for π-π Interactions: The benzene dimer. *Journal of the American Chemical Society*, 124(36), pp.10887–10893.

Skerker, J.M. et al., 2008. Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell*, 133, pp.1043–1054.

Skrabanek, L. et al., 2008. Computational prediction of protein-protein interactions. *Molecular* *biotechnology*, 38(1), pp.1–17.

Sliwoski, G. et al., 2014. Computational Methods in Drug Discovery. *Pharmacological reviews*, 66(1), pp.334–95.

Slupsky, J.R. et al., 1987. Characterization of cardiac calsequestrin. *Biochemistry*, 26(20), pp.6539–44.

Smith, G.P., 1985. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, N.Y.)*, 228(4705), pp.1315–7.

Snel, B. et al., 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18), pp.3442–4.

Souers, A.J. et al., 2013. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, 19(2), pp.202–8.

van der Spoel, D. and van Maaren, P.J., 2006. The Origin of Layer Structure Artifacts in Simulations of Liquid Water. *J. Chem. Theory Comput.*, 2(1), pp.1–11.

Stanker, L.H. et al., 2013. A monoclonal antibody based capture ELISA for botulinum neurotoxin serotype B: toxin detection in food. *Toxins*, 5(11), pp.2212–26.

Steinbrecher, T., Joung, I. and Case, D.A., 2011. Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. *Journal of Computational Chemistry*, 32(15), pp.3253–3263.

Stuart, J.M. et al., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643), pp.249–255.

Su, X.-D. et al., 2015. Protein crystallography from the perspective of technology developments. *Crystallography Reviews*, 21(1–2), pp.122–153.

Su, Y. et al., 2007. Linear Interaction Energy (LIE) Models for Ligand Binding in Implicit Solvent: Theory and Application to the Binding of NNRTIs to HIV-1 Reverse Transcriptase in Implicit Solvent. *Journal of Chemical Theory and Computation*, 3(1), pp.256–277.

Sutthibutpong, T. et al., 2016. Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. *Nucleic Acids Research*, 44(19), pp.9121–9130.

Suzuki, M., Youle, R.J. and Tjandra, N., 2000. Structure of Bax: coregulation of dimer formation and intracellular localization. *Cell*, 103(4), pp.645–54.

Szajek, A.Y. et al., 2016. The US regulatory and pharmacopeia response to the global heparin contamination crisis. *Nature Biotechnology*, 34(6), pp.625–630.

Szklarczyk, D. et al., 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue), pp.D561-8.

Takakusagi, Y. et al., 2010. Use of phage display technology for the determination of the targets for small-molecule therapeutics. *Expert opinion on drug discovery*, 5(4), pp.361–89.

Tan, C., Tan, Y.H. and Luo, R., 2007. Implicit nonpolar solvent models. *Journal of Physical Chemistry*

B, 111(2), pp.12263–12274.

Tan, Y.S. et al., 2016. Benzene Probes in Molecular Dynamics Simulations Reveal Novel Binding Sites for Ligand Design. *The Journal of Physical Chemistry Letters*, 7(17), pp.3452–3457.

Tironi, I.G. et al., 1995. A generalized reaction field method for molecular dynamics simulations. *The Journal of Chemical Physics*, 102(13), pp.5451–5459.

Tohge, T. and Fernie, A.R., 2012. Co-expression and co-responses: within and beyond transcription. *Frontiers in plant science*, 3(November), p.248.

Torshin, I.Y., Weber, I.T. and Harrison, R.W., 2002. Geometric criteria of hydrogen bonds in proteins and identification of “bifurcated” hydrogen bonds. *Protein Engineering Design and Selection*, 15(5), pp.359–363.

Tuncbag, N., Keskin, O., Nussinov, R,, Gursoy, A., 2017. Prediction of protein interactions by structural matching: prediction of PPI networks and the effects of mutations on PPIs that combines sequence and structural information. *Methods Mol Biol*, 1558, pp.255–270.

Unterholzner, S.J., Poppenberger, B. and Rozhon, W., 2013. Toxin – antitoxin systems: Biology , identification , and application. *Mobile Genetic Elements*, (October), pp.1–13.

Urano, R. and Okamoto, Y., 2015. Designed-walk replica-exchange method for simulations of complex systems. *Computer Physics Communications*, 196(1992), pp.380–383.

Velankar, S. et al., 2012. PDBe: Protein Data Bank in Europe. *Nucleic acids research*, 40(Database issue), pp.D445-52.

Velankar, S. et al., 2010. PDBe: Protein Data Bank in Europe. *Nucleic acids research*, 38(Database issue), pp.D308-17.

Velasco-García, R. and Vargas-Martínez, R., 2012. The study of protein-protein interactions in bacteria. *Canadian journal of microbiology*, 58(11), pp.1241–57.

Verlet, L., 1967. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review Letters*, 159(5), pp.98–103.

de Visser, S.P. et al., 2014. Computational modelling of oxygenation processes in enzymes and biomimetic model complexes. *Chemical communications (Cambridge, England)*, 50(3), pp.262–82.

de Vlieg, J., Berendsen, H.J.C. and van Gunsteren, W.F., 1989. An NMR-Based Molecular Dynamics Simulation of the Interaction of the lac Repressor Headpiece and Its Operator in Aqueous Solution. *Proteins*, 6(2), pp.104–127.

Wang, D.X. and Wang, M.X., 2013. Anion-π interactions: Generality, binding strength, and structure. *Journal of the American Chemical Society*, 135(2), pp.892–897.

Wang, L. et al., 2013a. Identification of an FHL1 Protein Complex Containing Gamma-Actin and Non-Muscle Myosin IIB by Analysis of Protein-Protein Interactions. *PloS one*, 8(11), p.e79551.

Wang, Q., Zhuravleva, A. and Gierasch, L.M., 2011. Exploring weak, transient protein--protein interactions in crowded in vivo environments by in-cell nuclear magnetic resonance spectroscopy. *Biochemistry*, 50(43), pp.9225–36.

Wang, X. et al., 2013b. Atomic Model of Rabbit Hemorrhagic Disease Virus by Cryo-Electron Microscopy and Crystallography. *PLoS Pathogens*, 9(1), p.e1003132.

Warshel, A. and Lifson, S., 1970. Consistent force field calculations. II. Crystal structures, sublimation energies, molecular and lattice vibrations, molecular conformations, and enthalpies of alkanes. *The Journal of Chemical Physics*, 53(2), p.582.

Warwicker, J., 1986. Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. *Journal of Theoretical Biology*, 121(2), pp.199–210.

Warwicker, J. and Watson, H.C., 1982. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *Journal of Molecular Biology*, 157(4), pp.671–679.

Wassman, C.D. et al., 2013. Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nature communications*, 4, p.1407.

Waudby, C. a et al., 2013. Protein folding on the ribosome studied using NMR spectroscopy. *Progress in nuclear magnetic resonance spectroscopy*, 74, pp.57–75.

Weiner, S.J. et al., 1983. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Society*, 106(3), pp.765–784.

Weingarth, M. and Baldus, M., 2013. Solid-state NMR-based approaches for supramolecular structure elucidation. *Accounts of chemical research*, 46(9), pp.2037–46.

White, D.N., 1997. A computationally efficient alternative to the Buckingham potential for molecular mechanics calculations. *Journal of computer-aided molecular design*, 11(5), pp.517–521.

van Wijk, S.J.L. et al., 2009. A comprehensive framework of E2-RING E3 interactions of the human ubiquitin-proteasome system. *Molecular systems biology*, 5(295), p.295.

Winter, C. et al., 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic acids research*, 34(Database issue), pp.D310-4.

Wojcikiewicz, R.J.H. et al., 2003. Ubiquitination and proteasomal degradation of endogenous and exogenous inositol 1,4,5-trisphosphate receptors in alpha T3-1 anterior pituitary cells. *The Journal of biological chemistry*, 278(2), pp.940–7.

Wojdyla, J.A. et al., 2012. Structure of the ultra-high-affinity colicin E2 DNase-Im2 complex. *Journal of Molecular Biology*, 417(1–2), pp.79–94.

Yakovchuk, P., Protozanova, E. and Frank-Kamenetskii, M.D., 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2), pp.564–574.

Yanamala, N., Tirupula, K.C. and Klein-Seetharaman, J., 2008. Preferential binding of allosteric modulators to active and inactive conformational states of metabotropic glutamate receptors. *BMC bioinformatics*, 9, p.S16.

Young, D.C., 2001. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, New York: John Wiley and Sons

Zhang, J. et al., 2017. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation*, 13(3), pp.1034–1043.

Zhang, Y. et al., 2013. Interaction of presequence peptides with human translocase of inner membrane of mitochondria Tim23. *Biochemical and biophysical research communications*, 437(2), pp.292–9.

Ziegel, E. et al., 2007. *Numerical Recipes: The Art of Scientific Computing.* Third edition. Cambridge, UK: Cambridge University Press

Zwanzig, R.W., 1954. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *Journal of Chemical Physics*, 22(8), pp.1420–1426.

Chapter 2 - Protein – protein interactions in paralogues: electrostatics modulates specificity on a conserved steric scaffold

PAPER 1

Stefan M. Ivanov[1,2], Andrew Cawley[1,#a], Roland G. Huber[2], Peter J. Bond[2,3], and Jim Warwicker[1*]

[1] Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, 131 Princess Street, Manchester, United Kingdom

[2] Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Matrix 07-01, 30 Biopolis Street, Singapore

[3] Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore

[#a] Current Address: Unilever Research and Development, Port Sunlight, Bromborough Road, Wirral, United Kingdom

[*] Corresponding author:
E-mail: jim.warwicker@manchester.ac.uk (JW)

SMI and AC compiled the datasets and performed the calculation of interface descriptors; JW wrote the code for the calculations; all authors analyzed the data and wrote the paper.

Published in *PLoS ONE*, Volume 12, Issue 10, Pages e0185928, 10 October 2017

Abstract

An improved knowledge of protein – protein interactions is essential for better understanding of metabolic and signaling networks, and cellular function. Progress tends to be based on structure determination and predictions using known structures, along with computational methods based on evolutionary information or detailed atomistic descriptions. In the present work, it is hypothesized that for the case of interactions across a common interface, between proteins from a pair of paralogue families or within a family of paralogues, a relatively simple interface description could distinguish between binding and non-binding pairs. Using binding data for several systems, and large-scale comparative modeling based on known template complex structures, it is found that charge – charge interactions (for groups bearing net charge) are generally a better discriminant than buried surface. This is particularly the case for paralogue families that are less divergent, with more reliable comparative modeling. It is then suggested that electrostatic interactions are major determinants of specificity in such systems, an observation that could be used to predict binding partners.

2.1. Introduction

The interplay between biopolymers is critical in directing and maintaining physiological processes. Whilst genome-sequencing projects are providing large amounts of protein sequence data from many organisms, understanding of binding specificity between proteins, and how a protein selects partners from closely related alternatives, remains limited. The majority of work in identifying specificity determinants focuses on the sequences and structures of the proteins involved. Methods for identifying residues that determine specificity face challenges, often due to an absence of suitable experimentally determined structures or the lack of affinity data (Fromer and Shifman, 2009). Where structural models are available, computational predictions of protein – protein interactions focus on aspects of the association such as size, shape, and physicochemical complementarities at the interaction interface (Pechmann et al., 2009; Ritchie, 2008). Increasingly, experimental data is being combined with physicochemical calculations to provide predictions of interfaces and the roles of individual residues at interfaces (Petukh et al., 2015; Xue et al., 2015) and, in turn, experiments are being guided by such calculations (Winter et al., 2012). Sequence, evolutionary, and expression data may also be included in predictions (Keskin et al., 2016). Computational methods can be benchmarked against experimentally determined complexes in community-wide studies (Vajda and Kozakov, 2009; Janin, 2010).

Genomic and proteomic studies have shown that most proteins belong to families of evolutionarily, and often functionally, related molecules (Teichmann and Babu, 2004). The number of proteins in a given family increases through gene duplication and the resulting generation of paralogues. For example, the human genome encodes several hundred protein kinases, which are believed to have arisen through large- and small-scale genetic duplications (Manning et al., 2002). When interactions between proteins in paralogue families are considered, maintaining physiological cellular signaling requires proteins to distinguish between highly similar surfaces. Several approaches have been taken in attempting to rationalize such intricate interactions. Studies have shown that interacting proteins are coexpressed with a greater degree of correlation than random, non-interacting pairs (Stuart et al., 2003). Moreover, it has been shown that coexpressed proteins coevolve (Shoemaker and Panchenko, 2007), with genes with multiple coexpressed partners evolving more slowly than genes with fewer coexpressed partners (Jordan et al., 2004). Structural and bioinformatics studies have shown that protein – protein interfaces can be divided into a core and rim, with the rim being enriched in subfamily-specific residues (Aiello and Caffrey, 2012). There have been attempts to rationalize specificity through computational studies at differing levels of theoretical sophistication. Fong and Keating (2004) have assessed the binding feasibility of different pairs of leucine zipper transcription factors by representing each pair as a multidimensional vector, the entries of which represent the different amino acid pairings from the two opposing chains. Each vector is then multiplied by a vector of corresponding weights for the different pairings. Most interfaces, however, are more complicated than the coiled-coil of a leucine zipper dimer, and are less amenable to such an approach. Atomistic models are, therefore, more prominent in rationalizations of specificity determinants. Calculations of electrostatic interactions with Generalized Born or Poisson-Boltzmann methods, combined with surface area, are often used in molecular mechanics, and have proven successful in identifying specificity determinants and recognition mechanisms (Delgado-Soler et al., 2012; Ivanov et al., 2016). Due to the computational expense of traversing the conformational space of even small protein – protein complexes, such methods are generally more successful in rationalizing protein – small molecule binding than protein – protein binding (Steinbrecher et al., 2006). More computationally expensive higher-level theory calculations, such as density functional theory and quantum mechanics, are almost exclusively carried out on protein – small molecule systems (Fox et al., 2014).

The present report examines specificity in paralogous protein – protein interactions from a structural viewpoint, combining atomistic-level detail, with rapid calculation of electrostatic interactions and surface burial. In computing interfacial properties, an empirical calculation approach is taken, using the solvent accessible surface area (SASA) approach of Lee and Richards (1971) and a Debeye-Hückel computation of charge interactions between groups bearing net charge (Warwicker, 1999). Computed properties are compared between interacting and non-interacting pairs of proteins, identified from literature. This study aims to establish whether these simple interface descriptors discriminate between binding and non-binding pairs in paralogous protein – protein interactions. It has been shown that in certain regulatory systems involving paralogous protein – protein binding, charge interactions modulate specificity on a relatively conserved steric framework (Grigoryan et al., 2009). Further sets of experimental data have been identified, together with structural templates for modeling paralogous complexes, so that this observation can be tested more generally. The simple surface area and electrostatics model allows rapid estimation of interfacial energetics over a wide range of paralogue complexes generated by side chain replacement comparative modeling. It is found that the model for charge mediated specificity persists in numerous systems, although both the effect and the confidence with which it can be assessed decay with sequence divergence. Whilst there are many examples of paralogous protein - protein interactions, corresponding experimental data is limited. Improved modeling of specificity in such interactions will lead to a better understanding of structure – function relationships and protein – protein interaction networks.

2.2. Methods

2.2.1. Sequence alignment and comparative modeling

The key requirements for a system to be included in this study are the availability of binding data, and the presence of at least one representative complex in the protein structural database (Berman et al., 2000). After obtaining a three-dimensional structure of a complex, a multiple sequence alignment is generated between each molecule in the template and the relevant set of paralogues. Sequences were obtained from UniProt (Apweiler et al., 2004). Sequence alignment was performed with the default settings of T-Coffee (Notredame et al., 2000), and used in generating a three-dimensional structure for each possible combination of potential interactors. The comparative modeling pipeline incorporated side chain replacement with fixed backbones. Identical side chains between template and model are maintained in their conformers, while swapped side chains are repacked (Bougouffa and Warwicker, 2008) with an adaptation (Cole and Warwicker, 2002) of a self-consistent mean-field method for rotamer selection from a rotamer library (Koehl and Delarue, 1994). The algorithm performs pairwise packing of rotamers while observing a predefined tolerance for clashes of van der Waals radii. Beyond that tolerance, overlap of atomic van der Waals radii is prohibited subject to a further relaxation that is incremented until a packing solution is found, i.e. with all side chains having at least one allowed rotamer (Cole and Warwicker, 2002).

2.2.2. Buried surface and electrostatic energy calculations

The estimated electrostatic energy of interaction for groups bearing net charge (NetQ) and changes in nonpolar and polar solvent accessible surface areas upon complexation (ΔSASAnp and ΔSASApol) are calculated for all complexes modeled as rigid structures, with the differences for surfaces denoting subtraction of the sum of the component values from the complex value. Each component may be one, or more than one, polypeptide chain (Bougouffa and Warwicker, 2008). Surfaces are calculated using a sphere of radius 1.4 Å rolling on the van der Waals contour of a protein (Lee and Richards, 1971; Cole and Warwicker, 2002). In keeping with the empirical nature of this study, a framework for electrostatic interactions was used that allowed rapid application to multiple comparative models, with simple Debye-Hückel estimation of charge interactions in water at neutral pH and 0.15 M ionic strength (Warwicker, 1999). For each complex, NetQ is computed by summing all interactions between charged groups across proteins (Lys, Arg, N-terminus +1; Asp, Glu, C-terminus -1) and subtracting the sum of charge interactions in the individual proteins from that. Charges q_i and q_j, separated by a distance of r, interact with a potential of

$$\Phi = \frac{q_i q_j e^{-\kappa r}}{4\pi\varepsilon_0 \varepsilon r} \quad equation \;\; 2.1.$$

Here, ε_0 is the zero permittivity of vacuum, ε is the relative permittivity of water (80), and κ is the Debye-Hückel factor at 0.15 M ionic strength (Warwicker, 1999).

2.2.3. Binding data and structural templates

Experimental data obtained from literature is used to separate interactors from non-interactors, which are then coupled with template-based comparative models for the potential interacting pairs. In the case of bZIPs, a dataset of 127 strong interactions, 324 weak interactions, and 1214 non-interactions was assembled from a comprehensive study of leucine zipper dimerization (Newman and Keating, 2003). When reporting their experimental binding results, Newman and Keating (2003) define a strong interaction as having a Z-score [(signal – mean)/estimated standard deviation] of greater than 10, weak interactions as having Z-scores between 2.5 and 10, non-interactions as having Z-scores below 1, and "undetermined" when not meeting any of the above criteria. Leucine zipper sequences were aligned with each other and the template from the first zipper anchoring position. Templates with long helical regions were chosen - 1T2K (Panne et al., 2004) and 1CI6 (Podust et al., 2001).

As part of the ubiquitination pathway, ubiquitin-conjugating enzymes (E2s) interact with ubiquitin-ligating enzymes (E3s). Human E3 ubiquitin ligases are divided into three subgroups depending on the structure of the catalytic domain, the largest group being the RING-type E3s (Li et al., 2008). In a genomic study, 31 human E2s, 17 E2 pseudogenes, and 313 RING-type E3s were identified (van Wijk et al., 2009). A dataset of 329 interactions and 7219 non-interactions was derived. Four template structures of different RING domain lengths were used: 2YHO (36 amino acids, Zhang et al., 2011), 3HCT (40 amino acids, Yin et al., 2009), 1UR6 (44 amino acids, Dominguez et al., 2004), and 4CCG (59 residues, Hodson et al., 2014). A separate study on functional interactions between 22 human E2s and 9 HECT type E3s produced a dataset of 94 interacting and 104 non-interacting pairs (Sheng et al., 2012). All 198 models were generated using the 3JVZ (Kamadurai et al., 2009), 1C4Z (Huang et al., 1999), and 5HPT (Zhang et al., 2016) template structures.

The *Caulobacter crescentus* genome encodes three parE toxins and one pseudogene (parE$_2$), and their corresponding parD antitoxins (Yamaguchi et al., 2011), whereas the relEB family is represented by four toxin – antitoxin pairs (Pandey and Gerdes, 2005). The parED/relEB superfamily toxins and antitoxins interact with each other on a 1:1 basis (Dalton and Crosson, 2010), i.e. each toxin interacts with and is neutralized by its cognate antitoxin only. Thus, there are 3 interacting and 6 non-interacting pairs in the parED system, and 4 interacting and 12 non-interacting pairs in the relEB system. Another toxin – antitoxin system is the *Mycobacterium tuberculosis* vapBC family, comprising 48 vapC toxins that interact on a 1:1 basis with their vapB antitoxins (Ahidjo et al., 2011), which produces 48 interacting and 2256 non-interacting pairs. Complex structures for the toxin – antitoxin pairs were generated by modeling on the 3KXE (Dalton and Crosson, 2010) template for the parED family; 2KC8 (Li et al., 2009) for the relEB family; and 3H87 (Min et al., 2012) and 3DBO (Miallau et al., 2009) for the vapBC family.

Data on BH3 peptide interactions with antiapoptotic proteins, consisting of 48 IC$_{50}$ values, was obtained from solution competition assays on the binding between five antiapoptotic proteins and BH3 peptides from 10 proapoptotic proteins (Okamoto et al., 2012). The 48 complexes were generated with comparative modeling based on the 2XA0 (Ku et al. 2011), 3PL7 (Czabotar et al., 2011), and 1ZY3

(Denisov et al., 2006) templates. Table 1 provides a summary of the examined systems.

system		interactors (N1)	weak interactors	non-interactors (N2)	experimental technique (reference)
bZIPs		127	324	1214	fluorescent peptide arrays (Newman and Keating, 2003)
E2 – RING E3s		329	-	7219	yeast two-hybrid screen (Y2H) (van Wijk et al., 2009)
E2 – HECT E3s		94	-	104	functional screen (Sheng et al., 2012)
Toxins – antitoxins	parE – parD	3	-	6	growth inhibition (Dalton and Crosson, 2010)
	relE – relB	4	-	12	growth inhibition (Dalton and Crosson, 2010)
	vapC – vapB	48	-	2256	growth inhibition and Y2H (Ahidjo et al., 2011)
Bcl-2-intrafamily interactions		43	-	5	solution competition assay (Okamoto et al., 2012)

Table 2.1. Summary of the paralogue systems examined

Before comparative modeling, the binding mode between different pairs of proteins within each system was examined for conservation. Structural and other experimental data demonstrate that the binding modes within the Bcl-2 family and the E2 - E3 system are highly conserved (Day et al., 2008; Czabotar et al., 2007). Generated homology models of Bcl-2 family complexes are in excellent agreement with recently published structures (Rajan et al., 2015; Robin et al., 2015; Kim et al., 2015; Jenson et al., 2017) with Cα RMSDs ~ 0.5 Å. Only in the toxin – antitoxin systems large divergence in sequence and structure was observed, with sequence identities as low as 4% and RMSDs above 3 Å.

Where comparisons are made between sets of calculated properties, statistical significance is assessed with the two-tailed Mann-Whitney U test – a nonparametric test used to determine if samples come from populations with identical distributions (not necessarily Gaussian) (Mann and Whitney, 1947). Use of multiple templates permitted assessing the robustness of the results presented herein.

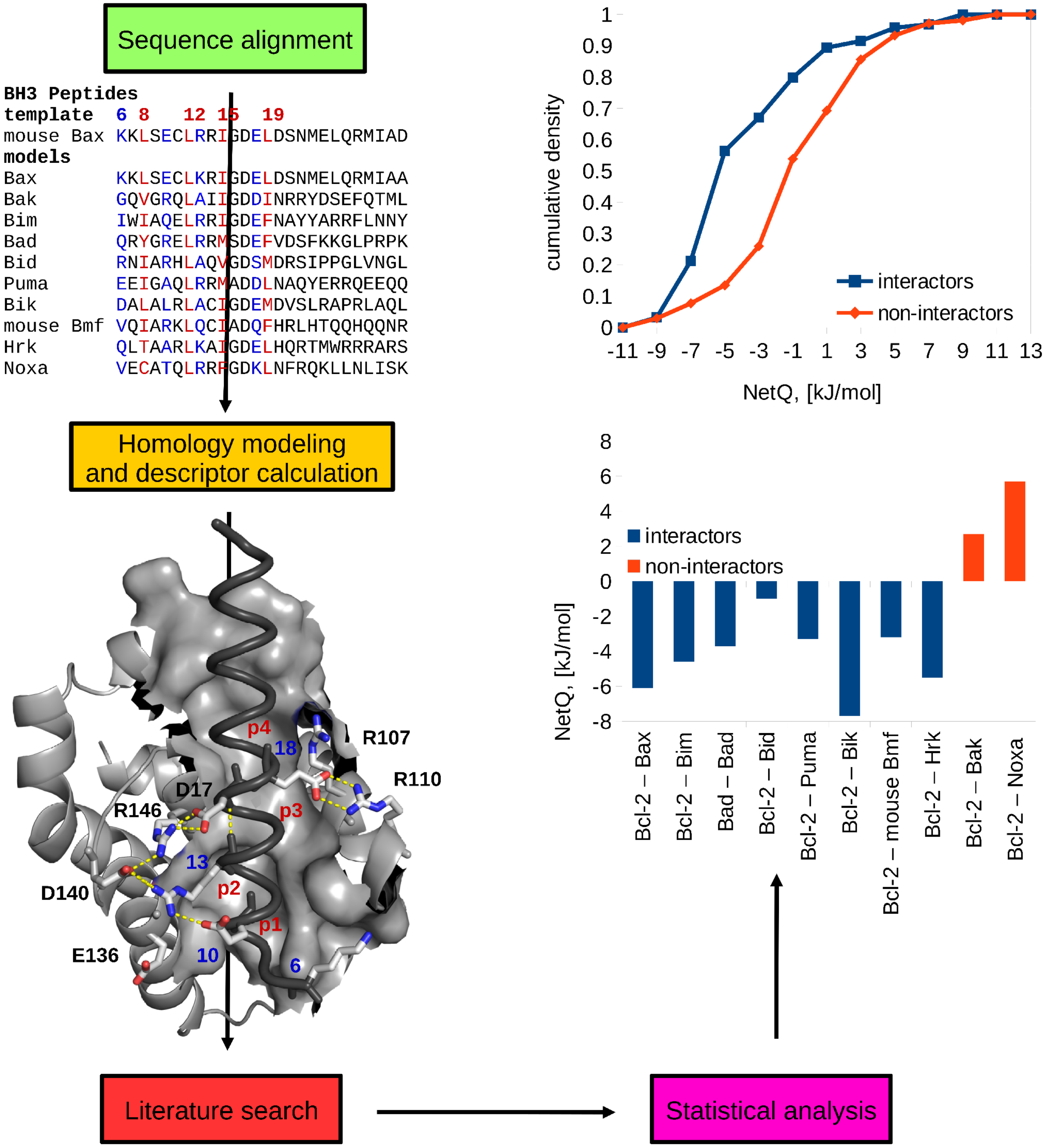
2.3. Results

2.3.1. Workflow

A multiple sequence alignment between paralogues in protein families is used to perform comparative modeling with one or more template structures for a complex. Figure 2.1 shows the procedure for 10 BH3 peptides and a template structure of an antiapoptotic protein bound to a BH3 peptide. For each of the 10 modeled complexes, interface descriptors are computed: interactions of groups bearing net charge (NetQ), change in nonpolar solvent accessible surface area upon complex formation (ΔSASAnp), and change in polar solvent accessible surface area (ΔSASApol). Interacting and non-interacting pairs are identified from literature and interfacial properties are compared between the two groups with appropriate statistical analysis. Results are plotted, for this example (Figure 2.1) as individual values of NetQ for interacting and non-interacting pairs in the Bcl-2 – BH3 peptide set, and also as the cumulative density of NetQ values in a larger dataset.

2.3.2. Basic leucine zipper transcription factors

After performing a multiple sequence alignment, 3-dimensional models of all possible binary combinations of bZIPs were generated. Interfacial properties for the different complexes were calculated and compared between interactors and non-interactors. The electrostatic energy of interaction (NetQ) is more favorable for interactors, (mean M = -5.3, standard deviation SD = 3.9 kJ/mol, number of interacting pairs = N1 = 127) than for non-interactors (M = -2.3, SD = 3.40 kJ/mol, number of non-interacting pairs = N2 = 1214) when modeling on the 1CI6 template. Change in nonpolar solvent accessible surface area is larger in interactors (M = -1681, SD = 99 Å^2) than non-interactors (M = -1633, SD = 95 Å^2), whereas change in buried polar accessible surface area is similar for interactors (M = -473, SD = 112 Å^2) and non-interactors (M = -486, SD = 100 Å^2) (Figure 2.2). The NetQ and ΔSASAnp differences between interactors and non-interactors are significant, with p values of 4.29x10^{-19} and 1.27x10^{-9}, respectively, using the two-tailed Mann-Whitney U test, whereas ΔSASApol is not significantly different (p = 0.88). Notably, weak interactions cluster in between interactions and non-interactions, but only in terms of NetQ values. The ranking of p-values is the same when modeling with the 1T2K template, with Mann-Whitney test p-values for interactors compared with non-interactors of 6.99x10^{-14} for NetQ, 2.71x10^{-10} for ΔSASAnp, and 2.62x10^{-5} for ΔSASApol.



```
BH3 Peptides
template     6 8   12 15  19
mouse Bax    KKLSECLRRIGDELDSNMELQRMIAD
models
Bax          KKLSECLKRIGDELDSNMELQRMIAA
Bak          GQVGRQLAIIGDDINRRYDSEFQTML
Bim          IWIAQELRRIGDEFNAYYARRFLNNY
Bad          QRYGRELRRMSDEFVDSFKKGLPRPK
Bid          RNIARHLAQVGDSMDRSIPPGLVNGL
Puma         EEIGAQLRRMADDLNAQYERRQEEQQ
Bik          DALALRLACIGDEMDVSLRAPRLAQL
mouse Bmf    VQIARKLQCIADQFHRLHTQHQQNR
Hrk          QLTAARLKAIGDELHQRTMWRRRARS
Noxa         VECATQLRRFGDKLNFRQKLLNLISK
```

Figure 2.1. Schematic representation of the workflow. In this example of BH3 peptides potentially binding to the Bcl-2 antiapoptopic protein, multiple sequence alignment feeds into comparative modeling, generation of electrostatic and buried surface area interface descriptors, and subsequent comparison between interactors and non-interactors, as individual complex and cumulative density data. The cumulative density derives from a larger dataset than the sequences shown. Key hydrophobic residues in the sequence alignment are highly conserved and highlighted in red; these four positions fit into conserved hydrophobic pockets on the surface of the protein, labeled in red - p1, p2, p3, and p4. The groove is partly in surface representation and colored in gray to better illustrate the positioning of the pockets. Variable positions of great significance (6, 10, 13, and 18) are highlighted in blue and discussed further in the text. Positions 6, 10, 13, and 18 of the 2XA0 template (Ku et al., 2011) are shown in stick representation, with side chain carbons in white, oxygen in red, and nitrogen in blue; their positions are also labeled with a blue number. Also shown in stick representation and labeled are the invariant aspartic acid in position 17 and key residues from the Bcl-2 protein. D17 from the peptides forms a salt bridge with R146 from the invariant NWGR motif, characteristic of antiapoptotic proteins. Residue numbering corresponds to canonical Uniprot numbering (Apweiler et al., 2004), except for the BH3 peptides, where the numbering corresponds to the numbering of the sequences simulated in Ivanov et al. (2016); all sequences are human except where explicitly stated otherwise. Backbones are in cartoon representation and colored in gray for Bcl-2 and dark gray for Bax. Interside chain salt bridges and backbone hydrogen bonds in the template structure are represented with dashed lines.

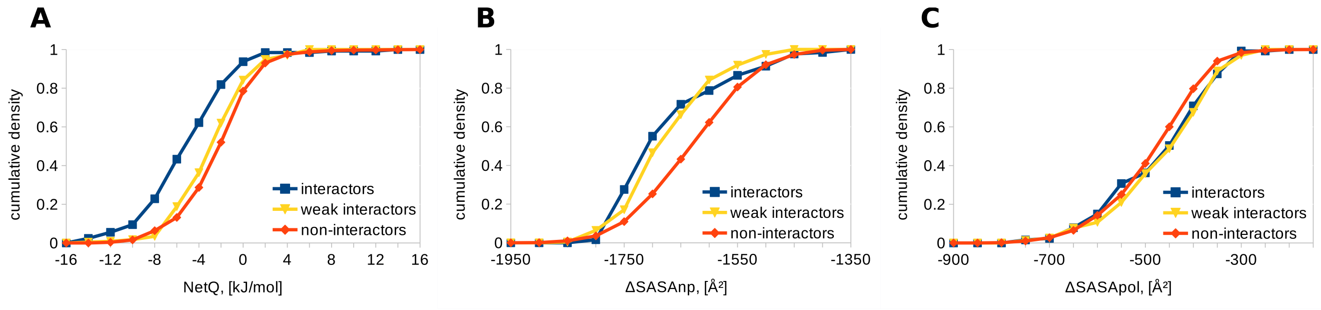


Figure 2.2. Comparison of interfaces for bZIPs. Cumulative densities for interactors, non-interactors, and weak interactors are shown, using the 1CI6 template. **(A)** NetQ **(B)** ΔSASAnp **(C)**

ΔSASApol

2.3.3. E2 ubiquitin conjugating enzymes – RING E3 ubiquitin ligases

Ubiquitination contributes to the regulation of many physiological processes (Cui et al., 2013). The transfer of ubiquitin to a protein substrate in the cell occurs through a complex series of interactions involving E1, E2, and E3 enzyme classes, with the number of enzymes in each class increasing by over an order of magnitude along the pathway. E2 enzymes accept activated ubiquitin from E1s and are, in turn, recognized by an E3 ubiquitin ligase. Finally, E3s transfer the ubiquitin to a protein target (Kar et al., 2012). Experimental studies on the ubiquitination pathway have provided insight into the specificity of protein – protein interactions within the system (van Wijk et al., 2009).

The majority of suitable templates in the Protein Data Bank (Berman et al. 2003) represent 36 – 46 residue-long RING domains. Modeling on a template with a RING domain length of 40 amino acids (3HCT) gave all three properties, NetQ, ΔSASAnp, and ΔSASApol, as significantly different between interactors and non-interactors. NetQ for interactors of M = -2.1, SD = 2.3 kJ/mol compares with M = 0.6, SD = 3.0 kJ/mol for non-interactors (N1 = 329, N2 = 7219, Mann-Whitney p = 4.70x10^{-22}). For interactors, ΔSASAnp, M = -661, SD = 77 Å^{2} compares with M = -610, SD = 102 Å^{2} for non-interactors (p = 1.31x10^{-22}). For ΔSASApol, interactors give M = -370, SD = 98 Å^{2} and non-interactors M = -398, SD = 95 Å^{2} (p = 5.76x10^{-9}). The largest available E3 structure suitable to be a template, a 59 residue-long RING domain bound to an E2 enzyme (4CCG), also gave separation for all three properties (Figure 2.3). NetQ for interactors is M = -3.0, SD = 4.1 kJ/mol and for non-interactors, M = -0.9, SD = 4.4 kJ/mol, with Mann-Whitney p = 1.35x10^{-13}. For ΔSASAnp, M = -806, SD = 99 Å^{2} for interactors compares with M = -767, SD = 100 Å^{2} for non-interactors (p = 9.50x10^{-18}). For ΔSASApol, interactors give M = -419, SD = 74 Å^{2} and non-interactors M = -460, SD = 90 Å^{2}, with p = 1.65x10^{-18}. Results were analogous for the 1UR6 and 2YHO templates.

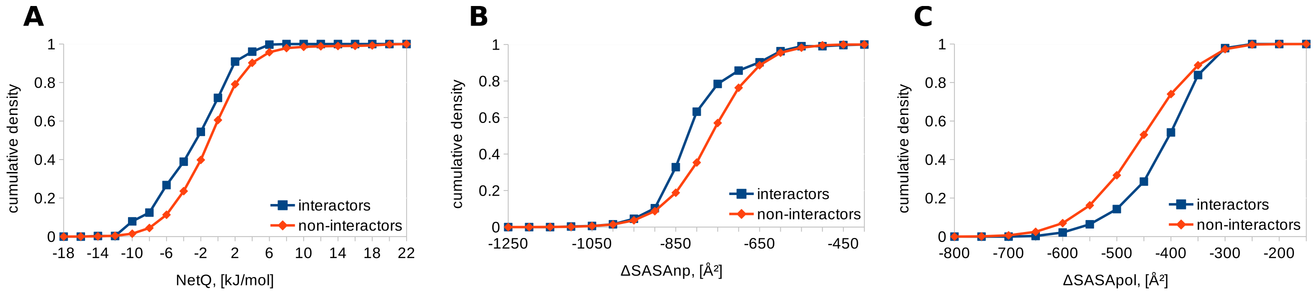


Figure 2.3. Comparison of interfaces for E2 – RING E3 complexes modeled on 4CCG. Cumulative densities for interactors and non-interactors are shown. **(A)** NetQ **(B)** ΔSASAnp **(C)** ΔSASApol

2.3.4. E2 ubiquitin conjugating enzymes – HECT E3 ubiquitin ligases

HECT E3 ubiquitin ligases, like the RING E3s, are involved in transferring ubiquitin from an E2 enzyme to a protein target. A study on functional E2 – HECT E3 interactions provides interaction data (Sheng et al., 2012). Using the 5HPT template (Figure 2.4), NetQ is more favorable for interactors (M = -4.0, SD = 4.1 kJ/mol) than for non-interactors (M = -1.0, SD = 4.0 kJ/mol), which is statistically significant with the two-tailed Mann-Whitney U test (N1 = 94, N2 = 104, p = 6.39x10^{-8}). Buried nonpolar surface is significantly larger in interactors (M = -1198, SD = 86 Å^{2}) than non-interactors (M = -1153, SD = 105 Å2, p = 2x10^{-3}), whereas polar surface is not significantly different (interactors M = -758, SD = 113 Å^{2}, non-interactors M = -751, SD = 123 Å^{2}, p = 0.33); results are analogous with the 1C4Z template. Similar results are also obtained with the 3JVZ template (Figure 2.5), listing interactors versus non-interactors: NetQ, M = -7.1, SD = 6.3 kJ/mol versus M = -2.9, SD = 5.6 kJ/mol, with p = 2.87x10^{-7}; ΔSASAnp, M = -1501, SD = 123 Å^{2} versus M = -1397, SD = 165 Å^{2} with p = 4.35x10^{-6}; ΔSASApol, M = -1159, SD = 177 Å^{2} versus M = -1091, SD = 190 Å^{2}, with p = 5.39x10^{-3}. For the 3JVZ template, unlike 5HPT and 1C4Z, buried polar surface area is also significantly different, possibly because the C-lobe of the HECT domain is positioned differently, capturing different points along the pathway of transferring ubiquitin from the E2 to the E3.

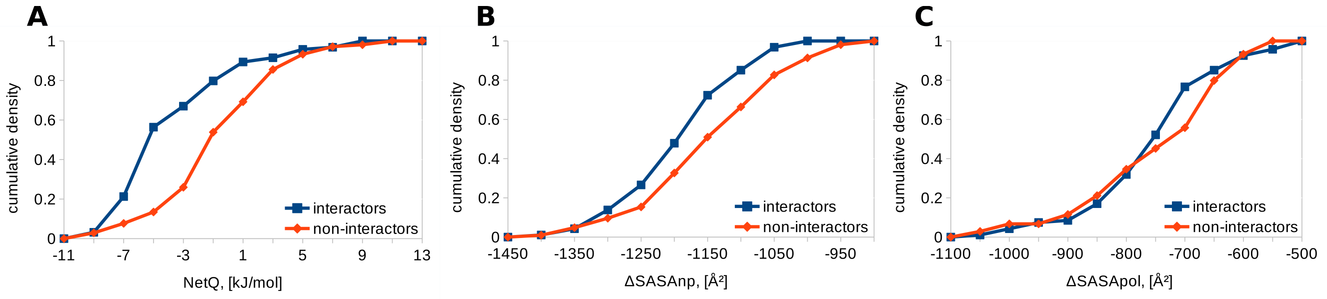


Figure 2.4. Comparison of interfaces for E2 – HECT E3 complexes modeled on 5HPT. Cumulative densities for interactors and non-interactors are shown. **(A)** NetQ **(B)** ΔSASAnp **(C)** ΔSASApol

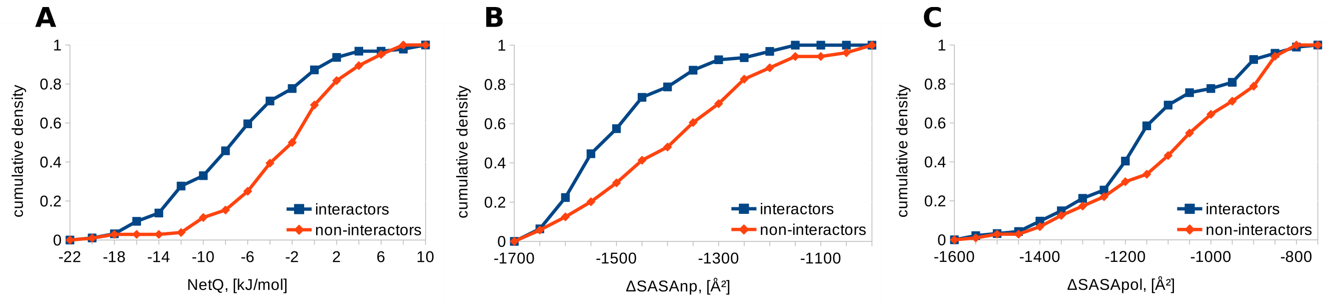


Figure 2.5. Comparison of interfaces for E2 – HECT E3 complexes modeled on 3JVZ. Cumulative densities for interactors and non-interactors are shown. **(A)** NetQ **(B)** ΔSASAnp **(C)** ΔSASApol

2.3.5. Toxin – antitoxin pairs

Specificity data is available for parD – parE pairs in *Caulobacter crescentus* (Dalton and Crosson, 2010), and vapB – vapC pairs for the related vapBC system in *Mycobacterium tuberculosis* (Ahidjo et al., 2011). NetQ, ΔSASApol, and ΔSASAnp are not significantly different between interactors and non-interactors for the vapBC family (N1 = 48, N2 = 2256, Figure 2.6) when modeling on the 3H87 or 3DBO templates. Modeling parE – parD pairs (N1 = 3, N2 = 6) on the 3KXE template, and relE – relB (N1 = 4, N2 = 12) on the 2KC8 template also fails to produce any separation between interactors and non-interactors. Toxin – antitoxin pairs are by far the most divergent system, with sequence identities as low as 4% and Cα RMSDs between 3 and 7 Å between the template structures.

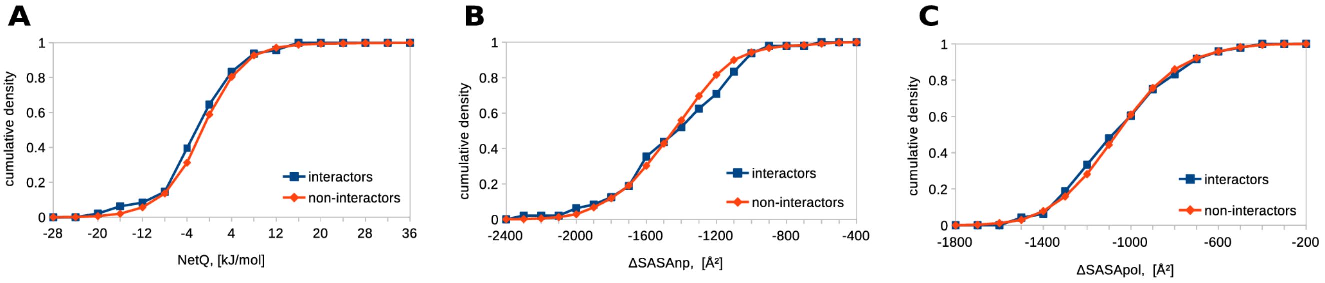


Figure 2.6. Comparison of interfaces for vapC toxin – vapB antitoxin complexes modeled on 3H87. Cumulative densities for interactors and non-interactors are shown. **(A)** NetQ **(B)** ΔSASAnp **(C)** ΔSASApol

2.3.6. Bcl-2-family proteins

Interactions among the Bcl-2-like proteins are crucial in regulating apoptosis. Specificity data is available for a set of 10 BH3 peptides from 8 BH3-only proteins and Bax and Bak (see Figure 2.1) interacting with BH3-binding grooves from 5 antiapoptotic proteins (human Bcl-xL, Bcl-2, Bcl-w, mouse Mcl-1 and A1) (Okamoto et al., 2012). After modeling antiapoptotic protein – BH3 peptide interactions on a template of human Bcl-2 bound to a BH3 peptide (2XA0), and comparing charge interactions and buried surfaces between interacting (N1 = 43) and non-interacting pairs (N2 = 5), the most evident difference is that non-interactors typically have a less favorable NetQ than interactors (p = 0.002, Figure 2.7). Buried surface is less discriminating between interactors and non-interactors (p = 0.131 for ΔSASAnp, p = 1 for ΔSASApol). Results are similar for the 3PL7 and 1ZY3 templates.

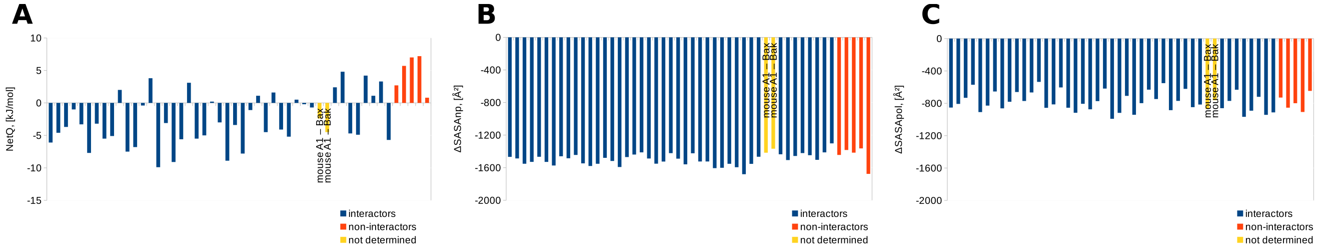


Figure 2.7. Comparison of interfaces for BH3 peptide – hydrophobic groove interactions, modeled on 2XA0. Color-coded histograms for interactors (blue), non-interactors (red), and

interactions that have not been determined (yellow). **(A)** NetQ **(B)** ΔSASAnp **(C)** ΔSASApol

2.4. Discussion

2.4.1. Homology modeling

This study assesses to what extent interactions between groups bearing net charge correlate with specificity for complexes formed by families of paralogous proteins at a common interface. Modeling paralogues on a suitable template and comparing empirical interface properties produces significant separation between interactors and non-interactors in most systems, with electrostatic interactions (between groups bearing net charge) typically being most discriminatory, followed by buried nonpolar surface, with buried polar surface being least discriminatory. It is shown that the results are largely independent of the template, although there is a limit to the template-based modeling with the present methods, demonstrated by the bacterial toxin – antitoxin pairs. These systems have diverged sufficiently to seriously impact the accuracy of the comparative modeling process. For example, the vapB$_2$ – vapC$_2$ and vapB$_5$ – vapC$_5$ pairs have an overall sequence identity of 6% and an RMSD between template structures of 6.6 Å, in contrast to the more typical case in the current work of sequence identities ~45% and RMSDs < 1.5 Å. Extensive sequence divergence, seen particularly in bacterial systems, is likely to provide a challenge for even the most sophisticated comparative modeling tools (Cohen and Schuldiner, 2011). However, the lower sequence divergence seen for proteins in paralogue families in metazoan systems makes them amenable to the comparative studies employed here.

The side chain replacement comparative modeling tool employed in the present work provides no opportunity to model insertions and deletions. Whether such changes can be modeled with sufficient accuracy and speed for large-scale analysis of complexes remains an open question. With the tool employed in this work, one has a choice whether to repack all side chains or to employ a minimalistic repacking of only those side chains that differ between model and template. The minimal repacking scheme has been used, as amino acid conservation could reflect an important role in maintenance of structure (Janin et al., 2008). For example, in the case of the BH3 peptides, 4 conserved hydropobic residues (see the sequence alignment in Figure 2.1) fit into 4 conserved pockets on the antiapoptotic proteins, and an invariant aspartic acid forms a salt-bridge with a conserved arginine from the partner protein (see the template structure in Figure 2.1). In RING E3s, conserved histidine and/or cysteine residues coordinate Zn^{2+} to maintain the native protein structure. It has been found that preserving the template amino acid side chain rotamer is beneficial in maintaining the stability of modeled antiapoptotic protein – BH3 peptide complexes during molecular dynamics simulations (Ivanov et al., 2016).

High throughput experimental data for protein – protein interactions is key for the current study, but this data can be imprecise. For example, the largest dataset used, E2 – RING E3 interactions, derives from a yeast two-hybrid screen (van Wijk et al., 2009). Given the generally low affinity of E2 – E3 interactions (Metzger et al., 2014), the screen may contain false positive and/or false negative data. Additionally, the functional assay used in the E2 – HECT E3 study is not capable of detecting interactions which only extend ubiquitin chains on mono-ubiquitinated targets or require cofactors (Sheng et al., 2012). Further computational study would benefit from more data collection in a variety of paralogue systems.

2.4.2. Specificity determinants and coevolution of sequences

In agreement with previous work (Lo Conte et al., 1999), results presented herein demonstrate that nonpolar surface constitutes the majority of the interface, consistent with it being the dominant contributor to the free energy of binding. The current study suggests that superposed on burial of nonpolar surface, the interactions of groups bearing net charge are a major determinant of binding specificity, for interactions between members of paralogue families. This finding is consistent with the core and rim model of protein interfaces (Chakrabarti and Janin, 2002), which postulates that conservation is greatest at the mostly hydrophobic core (Guharoy and Chakrabarti, 2005). This study indicates that specificity of binding for proteins from paralogue families, at a common interface, is largely, but not entirely, modulated by charge alterations on a relatively conserved steric scaffold. This can be interpreted from the standpoint of the core and rim model and the coevolution of sequences, showing how the two are related and arise from a single underlying phenomenon. This point is now illustrated with a medium-sized dataset – the BH3 peptide – antiapoptotic protein complexes. First, the sequence conservation within the system was assessed by computing the per-position Shannon entropy from the multiple sequence alignment of the 5 antiapoptotic proteins and 10 BH3 peptides; the entropy ranges from 0, when only one amino acid occurs in a given position, to 4.322, when all 20 amino acids are equally represented in that position (Garcia-Boronat et al., 2008). The per-residue entropies were then mapped onto the surface of the Bcl-2 - Bad complex (Supplementary information figure 2.1, see Appendix) to elucidate the pattern of sequence conservation within the system. As seen in SI Figure 2.1 and highlighted in the sequence alignment in Figure 2.1, anchoring hydrophobic residues are highly conserved. This is mirrored by a conservation of the hydrophobic residues forming the four hydrophobic pockets (see SI figure 2.1). This conservation manifests itself in the lack of variance in the ΔSASAnp values – the average for all complexes is -1497 Å^2, with a standard deviation of 78Å^2 - 5% of the mean – and no separation between interactors and non-interactors. This contrasts starkly with the behavior of the protein rim and the peptide residues that contact it, which is now described. The five antiapoptotic proteins under study – Bcl-xL, Bcl-2, Bcl-w, mouse Mcl-1, and mouse A1 (also known as Bfl-1) cluster into two groups based on their sequences and BH3-peptide-binding properties – the first three are highly similar to each other and bind the BH3 sequences of all BH3-only proteins, except Noxa. Conversely, Mcl-1 and A1 are more similar in sequence and binding properties to each other than Bcl-xL, Bcl-2, and Bcl-w – they are potent Noxa binders, but do not bind the Bad BH3 sequence (see Table 3.1). This behavior is mirrored by a set of specificity determinants previously reported (Ivanov et al., 2016). Peptide positions 6, 10, and 13 are solvent exposed and more variable (Figure 2.1 and SI figure 2.1). In the Bcls, they contact a conserved glutamic acid residue – E129/E136/E85 for Bcl-xL, Bcl-2, and Bcl-w, respectively. Additionally, residues 13 and 14 can interact with D133/D140/G89. Thus, peptide binders to the Bcls display a substantial enrichment of positive charge in positions 6, 10, 13, and 14. Conversely, Mcl-1 and A1 have a histidine and lysine, respectively – H233/K77, which correspond to E129/E136/E85 in the Bcls – and aspartic acids – D237 and D81 – in the positions corresponding to D133/D140/G89. Correspondingly, Noxa does not carry positive charges in positions 6 and 10 (rather, a valine and a threonine), but retains arginines in positions 13 and 14. Moreover, Noxa is the only BH3 peptide that has a positively charged residue in position 18 (K18), whereas all other peptides have a negatively charged or neutral residue in this position. This residue contacts residues R100/R107/R56/N204/E47 and R103/R110/R59/T207/K50 in Bcl-xL, Bcl-2, Bcl-w, Mcl-1, and A1, respectively (Figure 2.1 and SI figure 2.1). Apart from the large amount of positive charge in positions 6, 10, and 13, which comes into contact with H233/K77 of Mcl-1 and A1, respectively, Bad binding to Mcl-1 and A1 is also hindered by its residue in position 16 (serine), which is opposed by A142/A149/A98/T247/T91. T247/T91 in Mcl-1 and A1 impose greater steric restrictions than the corresponding alanines in the Bcls and the groove in this region cannot accommodate the bulkier S16 of Bad, whereas all other BH3 peptides carry a glycine or an alanine in this position. E10 of Bak is likely to be tolerated by the Bcls because it can form a salt-linked triad (Ivanov et al., 2016) with R13 and one of the conserved glutamic acids, as seen in the 2XA0 crystal structure (Figure 2.1). This triad likely enhances peptide helicity and stabilizes the peptide into the groove. A similar situation is also likely to occur with Noxa K18 and E47 and K50 of A1. Note that in the 2XA0 structure, the Bak peptide is missing 5 N-terminal residues, compared to the sequences simulated in Ivanov et al. (2016). Thus, the N-terminus is partially unfolded with K6 pointing away from E136. Ample structural and simulation data, however, shows that when those residues are present, the peptide in this region is helical, with positions 6, 10, and 13 interacting with the key residues discussed previously. The sequence variation of the rim and the peptide residues that interact with it is reflected in the great variance of ΔSASApol and NetQ – M = -785, SD = 117 Å^2; M = -2.05, SD = 4.38 kJ/mol – among all complexes, respectively, with NetQ being highly discriminatory between interactors and non-interactors. Thus, ΔSASAnp varies insignificantly (on average, 5% of the mean), whereas ΔSASA polar and NetQ vary considerably (15 and 213%, respectively). The pattern is quite apparent – the histograms for ΔSASAnp in Figure 2.7 trace out a smooth curve, whereas the histograms for NetQ and ΔSASApol trace out much more jagged curves. This is also reflected in the Shannon entropy of the complexes – entropies are lowest in the core of the interface, which forms the hydrophobic pockets; the invariant NWGR motif, characteristic of antiapoptotic proteins; and the peptide positions that contact these residues. Conversely, sequence entropy is high in the rim, mirroring the specificity determinants described above. Thus, a paradigmatic example of the core and rim model and sequence coevolution is presented – a highly conserved hydrophobic core provides affinity, alterations in the rim provide specificity, with substitution patterns correlated between partners – the pocket-forming residues remain invariant, in accord with the pocket residues of the peptides. Positions 6, 10, 13, 14 of the peptides are correlated with E129/E136/E85/H233/K77 and D133/D140/G89/D237/D81 of the proteins, mutations in position 16 are matched by alterations in A142/A149/A98/T247/T91, and position 18 is coupled to R100/R107/R56/N204/E47 and R103/R110/R59/T207/K50, with mutation patterns matched by binding patterns. It is very likely that these correlations are further propagated to the level of cellular function and tissue expression – healthy cells under normal conditions are likely to not express Noxa and Bad simultaneously, as this combination is guaranteed to trigger apoptosis in cells expressing any combination of the five proteins discussed here (Okamoto et al., 2012).

Whereas the importance of the hydrophobic pockets and the conserved aspartic acid in position 17 in this system is widely recognized, the importance of the rim is not, with only anecdotal reports of limited mutational studies (see Ivanov et al. (2016) and the references therein). A particularly interesting example in this regard is that mutating E18 of Bim to a serine diminishes binding to Bcl-xL, whereas phosphorylating that serine residue restores binding – a result of the phosphoryl group interacting with the arginine residues, previously discussed (Kim et al., 2015). Despite the great interest in the Bcl-2 system, the general pattern governing binding has largely evaded detection by the research community. This pattern was elucidated and unraveled only via extensive, detailed free energy calculations and meticulous analysis of the molecular dynamics trajectories and behavior of the peptides and proteins involved (Ivanov et al., 2016). In hindsight, many of those patterns could have been identified and recognized through much simpler and much faster calculations, such as the ones reported herein. While these results are encouraging, a great caveat must be stressed – homology modeling is far less capable of sampling even the narrowest of energy wells in the conformational landscape than molecular dynamics. For example, homology modeling may place a neighboring arginine and glutamic acid in orientations pointing away from each other or into the solvent, rather than pointing toward each other and forming a highly favorable salt bridge. While such conformations are not “wrong” and are certainly sampled in solution, a simple calculation on a static structure such as this will likely fail to detect that this is a favorable pairing. Conversely, a molecular dynamics simulation would allow the two residues to explore different conformations, eventually “finding each other” and falling into the energy well of a favorable interaction, which could then be detected by monitoring the potential energy of the system or the per-residue energies over time (Ivanov et al., 2016). Thus, molecular dynamics simulations are likely to remain the most reliable route in the search of affinity and specificity determinants, at least until more sophisticated homology modeling algorithms become available. This caveat is also likely to be partly responsible for the patterns in statistical significance observed herein - NetQ being discriminatory between interactors and non-interactors, but ΔSASApol being non-discriminatory. Many of the side chain placements performed in this study are likely suboptimal for electrostatic interactions, with side chains of opposite charge pointing away, rather than toward each other. It may be the case that NetQ results, computed through a Debye-Hückel formalism, are less sensitive to such shortcomings than ΔSASA calculations or that the electrostatics calculations reported herein are more accurate than the surface calculations.

It is conceivable that the primordial driving force behind the core and rim mechanism of modulating binding and the corresponding coevolution of sequences is the hydrophobic effect – the “desire” of nonpolar groups to be shielded from high dielectric solvent (Chandler, 2005). This naturally leads to a hydrophobic core. Specificity, then, is modulated to a large degree by charged residues. Indeed, it has been known for some time now that the most frequently occurring residue – residue pairs in protein – protein interfaces involve charged and aromatic residues, due to the multitude of interactions they can form in between each other - hydrogen bonds, salt bridges, salt-linked triads, cation – π, anion – π, T-shaped stacking, parallel shifted stacking, and bridging interactions (Gromiha et al., 2009; Gromiha et al., 2011). While substitution of a large hydrophobic residue with a small one can certainly affect binding specificity (i.e. abrogate binding), in reality it has a much greater effect on affinity than specificity – the large loss of affinity leads to a loss of binding. Conversely, substitutions involving charged and polar residues allow a much finer tuning of specificity. This naturally confines specificity mostly to the hydrophilic rim surrounding the core and shielding it from solvent, and is clearly manifested in the greater Shannon entropies of the rim residues and the greater statistical significance of NetQ than ΔSASA – NetQ has the smallest p value in most of the examined systems and requires the least number of data points to fall below the (commonly accepted) threshold value of 0.05. Notably, in the bZIP system, weak interactors cluster between strong interactors and non-interactors only in terms of NetQ values, but not buried surface, indicating that the former parameter is more amenable to fine tuning when designing specificity. Finally, in the case of paralogous protein – protein interactions, mutations in the partnering sets of proteins become correlated, i.e. the proteins coevolve, to assure physiological signaling.

2.5. Conclusions

More work is needed to conclusively determine the generality of the present findings. While the general ordering in terms of significance and, presumably, importance for specificity is NetQ > ΔSASAnp > ΔSASApol, more detailed analysis, potentially involving simulation studies, is necessary to definitively establish if this pattern is genuine, and make assessments quantitative, rather than categorical. Moreover, patterns are likely to vary, at least to an extent, in between systems. They may also vary depending on the type of interface – single-patch interfaces are known to follow the core and rim model well, whereas multipatch interfaces are likely to be more complicated (Janin et al., 2008). Moreover, the present work pertains exclusively to transient protein – protein interactions; permanent interfaces tend to be much larger, more hydrophobic, and resemble the core of a globular protein (Acuner Ozbabacan et al., 2011). Analyzing a medium-sized set of interactions and complexes with a relatively simple interface allows tracing the coevolution of sequences "by hand." Larger datasets with more complicated interfaces, however, necessitate careful structural and bioinformatics analysis. The empirical modeling pipeline could be trialed with a combination of charge and surface burial, or inclusion of volume-based descriptors (Bougouffa and Warwicker, 2008), and with other features, such as hydrogen bonding, more detailed analysis of buried surface and solvation (Shirts et al., 2003), and alternate analysis of side chain conformers in protein – protein interactions (Beglov et al., 2012). Further work is required to establish the degree to which the present empirical model can be used predictively for interacting and non-interacting pairs, in particular looking at restrictions imposed by divergence at the sequence alignment and comparative modeling stages. In this regard, calculations for A1 – Bax and A1 – Bak binding, which were not present in the original experimental dataset, have been included. The present calculations suggest that these are favorable interactions, which is corroborated by experimental work for A1 – Bax (Zhang et al., 2000) and A1 – Bak (Smits et al., 2008). The benefit of the current study is that a very simple model is employed, so that the effectiveness of charge interactions in contributing to interaction specificity is clearly encoded in the geometry of charge disposition at the interface. This study is designed around variation at a common interface, which yields to the simple model applied, in contrast, for example, to more detailed modeling for design of a new interface (Procko et al., 2013). It could be applied to modeling those parts of protein – protein interaction networks within a cell (Soni and Madhusudhan, 2017; Im et al., 2016;

Tuncbag et al., 2017) that involve interactions between proteins from paralogous families.

2.6. References

Acuner Ozbabacan, S.E. et al., 2011. Transient protein – protein interactions. *Protein Engineering, Design and Selection*, 24(9), pp.635–648.

Ahidjo, B.A. et al., 2011. VapC Toxins From Mycobacterium tuberculosis Are Rribonucleases that Differentially Inhibit Growth and Are Neutralized by Cognate VapB Antitoxins. *PloS one*, 6(6), p.e21738.

Aiello, D. and Caffrey, D.R., 2012. Evolution of specific protein-protein interaction sites following gene duplication. *Journal of molecular biology*, 423(2), pp.257–72.

Apweiler, R. et al., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue), pp.D115-9.

Beglov, D. et al., 2012. Minimal ensembles of side chain conformers for modeling protein-protein interactions. *Proteins: Structure, Function and Bioinformatics*, 80(2), pp.591–601.

Berman, H., Henrick, K. and Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. *Nature structural biology*, 10(12), p.980.

Berman, H.M. et al., 2000. The Protein Data Bank. *Nucleic acids research*, 28(1), pp.235–42.

Bougouffa, S. and Warwicker, J., 2008. Volume-based solvation models out-perform area-based models in combined studies of wild-type and mutated protein-protein interfaces. *BMC bioinformatics*, 9, p.448.

Chakrabarti, P. and Janin, J., 2002. Dissecting protein-protein recognition sites. *Proteins: Structure, Function and Genetics*, 47(3), pp.334–343.

Chandler, D., 2005. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature*, 437(7059), pp.640–647.

Cohen, Y. and Schuldiner, M., 2011. *Network Biology: Methods and Applications.* vol 781, Humana Press

Cole, C. and Warwicker, J., 2002. side chain conformational entropy at protein – protein interfaces. *Protein Science*, 11(12), pp.2860–2870.

Lo Conte, L., Chothia, C. and Janin, J., 1999. The Atomic Structure of Protein-Protein Recognition Sites. *Journal of molecular biology*, 285(5), pp.2177–98.

Cui, Z. et al., 2013. Regulation of cardiac proteasomes by ubiquitination, SUMOylation, and beyond. *Journal of molecular and cellular cardiology*, 72(6), pp.32–42.

Czabotar, P.E. et al., 2011. Mutation to Bax beyond the BH3 Domain Disrupts Interactions with Pro-survival Proteins and Promotes Apoptosis. *Journal of Biological Chemistry*, 286(9), pp.7123–7131.

Czabotar, P.E. et al., 2007. Structural insights into the degradation of Mcl-1 induced by BH3 domains. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15), pp.6217–6222.

Dalton, K.M. and Crosson, S., 2010. A Conserved Mode of Protein Recognition and Binding in a ParD-ParE Toxin-Antitoxin Complex. *Biochemistry*, 49(10), pp.2205–15.

Day, C.L. et al., 2008. Structure of the BH3 Domains from the p53-Inducible BH3-Only Proteins Noxa and Puma in Complex with Mcl-1. *Journal of Molecular Biology*, 380(5), pp.958–971.

Delgado-Soler, L. et al., 2012. Molecular determinants of Bim (BH3) peptide binding to pro-survival proteins. *Journal of chemical information and modeling*, 52(8), pp.2107–2118.

Denisov, A.Y. et al., 2006. Structural Model of the BCL-w - BID Peptide Complex and Its Interactions with Phospholipid Micelles. *Biochemistry*, 45(7), pp.2250–2256.

Dominguez, C. et al., 2004. Structural Model of the UbcH5B/CNOT4 Complex Revealed by Combining NMR, Mutagenesis, and Docking Approaches. *Structure*, 12(4), pp.633–44.

Fong, J.H., Keating, A.E. and Singh, M., 2004. Predicting specificity in bZIP coiled-coil protein interactions. *Genome biology*, 5(2), p.R11.

Fox, S.J. et al., 2014. Density functional theory calculations on entire proteins for free energies of binding: application to a model polar binding site. *Proteins*, 82(12), pp.3335–46.

Fromer, M. and Shifman, J.M., 2009. Tradeoff between stability and multispecificity in the design of promiscuous proteins. *PLoS Computational Biology*, 5(12), p.e1000627.

Garcia-Boronat, M. et al., 2008. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic acids research*, 36(Web Server issue), pp.35–41.

Grigoryan, G., Reinke, A.W. and Keating, A.E., 2009. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, 458(7240), pp.859–864.

Gromiha, M.M. et al., 2011. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome science*, 9(Suppl 1), p.S13.

Gromiha, M.M., Yokota, K. and Fukui, K., 2009. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Molecular bioSystems*, 5(12), pp.1779–1786.

Guharoy, M. and Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15447–52.

Hodson, C. et al., 2014. Structure of the Human FANCL RING-Ube2T Complex Reveals Determinants of Cognate E3-E2 Selection. *Structure (London, England : 1993)*, 22(2), pp.337–44.

Huang, L. et al., 1999. Structure of an E6AP-UbcH7 Complex: Insights into Ubiquitination by the E2-E3 Enzyme Cascade. *Science*, 286(5443), pp.1321–1326.

Im, W. et al., 2016. Challenges in structural approaches to cell modeling. *Journal of Molecular Biology*, 428(15), pp.2943–2964.

Ivanov, S.M. et al., 2016. Energetics and Dynamics Across the Bcl-2-Regulated Apoptotic Pathway Reveal Distinct Evolutionary Determinants of Specificity and Affinity. *Structure*, 24(11), pp.2024–2033.

Ivanov, S.M. et al., 2017. Protein – protein interactions in paralogues: electrostatics modulates specificity on a conserved steric scaffold. *PLoS ONE,* 12(10), pp.e0185928

Janin, J., 2010. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Molecular bioSystems*, 6(12), pp.2351–2362.

Janin, J., Bahadur, R.P. and Chakrabarti, P., 2008. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2), pp.133–80.

Jenson, J.M. et al., 2017. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1. *eLife*, 6, pp.1–23.

Jordan, I.K. et al., 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular Biology and Evolution*, 21(11), pp.2058–2070.

Kamadurai, H.B. et al., 2009. Insights into ubiquitin transfer cascades from a structure of a UbcH5B~Ubiquitin-HECT(NEDD4L) complex. *Molecular cell*, 36(6), pp.1095–102.

Kar, G. et al., 2012. Human Proteome-scale Structural Modeling of E2-E3 Interactions Exploiting Interface Motifs. *Journal of proteome research*, 11(2), pp.1196–207.

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

Kim, J.-S. et al., 2015. Conversion of cell-survival activity of Akt into apoptotic death of cancer cells by two mutations on the BIM BH3 domain. *Cell death and disease*, 6, p.e1804.

Koehl, P. and Delarue, M., 1994. Application of a Self-consistent Mean Field Theory to Predict Protein side chains Conformation and Estimate Their Conformational Entropy. *Journal of Molecular Biology*, 239(2), pp.249–275.

Ku, B. et al., 2011. Evidence that inhibition of BAX activation by BCL-2 involves its tight and preferential interaction with the BH3 domain of BAX. *Cell research*, 21(4), pp.627–41.

Lee, B. and Richards, F.M., 1971. The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of molecular biology*, 55(3), pp.379–400.

Li, G.-Y. et al., 2009. Inhibitory Mechanism of Escherichia coli RelE-RelB Toxin-Antitoxin Module Involves a Helix Displacement Near an mRNA Interferase Active Site. *The Journal of biological chemistry*, 284(21), pp.14628–14636.

Li, W. et al., 2008. Genome-Wide and Functional Annotation of Human E3 Ubiquitin Ligases Identifies MULAN, a Mitochondrial E3 that Regulates the Organelle's Dynamics and Signaling. *PloS one*, 3(1), p.e1487.

Mann, H.B. and Whitney, D.R., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1), pp.50–60.

Manning, G. et al., 2002. The Protein Kinase Complement of the Human Genome. *Science*, 298(5600), pp.1912–34.

Metzger, M.B. et al., 2014. RING-type E3 ligases: Master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1843(1), pp.47–60.

Miallau, L. et al., 2009. Structure and Proposed Activity of a Member of the VapBC Family of Toxin-Antitoxin Systems: VapBC-5 from Mycobacterium tuberculosis. *The Journal of biological chemistry*, 284(1), pp.276–83.

Min, A.B. et al., 2012. The crystal structure of the Rv0301-Rv0300 VapBC-3 toxin-antitoxin complex from M. tuberculosis reveals a Mg^{2+} ion in the active site and a putative RNA-binding site. *Protein science : a publication of the Protein Society*, 21(11), pp.1754–67.

Newman, J.R.S. and Keating, A.E., 2003. Comprehensive Identification of Human bZIP Interactions with Coiled-Coil Arrays. *Science*, 300(5628), pp.2097–101.

Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of molecular biology*, 302(1), pp.205–17.

Okamoto, T. et al., 2012. Sheeppox Virus SPPV14 Encodes a Bcl-2-Like Cell Death Inhibitor That Counters a Distinct Set of Mammalian Proapoptotic Proteins. *Journal of virology*, 86(21), pp.11501–11.

Pandey, D.P. and Gerdes, K., 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic acids research*, 33(3), pp.966–76.

Panne, D., Maniatis, T. and Harrison, S.C., 2004. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer. *The EMBO journal*, 23(22), pp.4384–93.

Pechmann, S. et al., 2009. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proceedings of the National Academy of Sciences*, 106(25), pp.10159–10164.

Petukh, M., Li, M. and Alexov, E., 2015. Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method. *PLoS Computational Biology*, 11(7), pp.1–23.

Podust, L.M., Krezel, A.M. and Kim, Y., 2001. Crystal Structure of the CCAAT Box/Enhancer-binding protein β Activating Transcription Factor-4 Basic Leucine Zipper Heterodimer in the Absence of DNA. *Journal of Biological Chemistry*, 276(1), pp.505–513.

Procko, E. et al., 2013. Computational Design of a Protein-Based Enzyme Inhibitor. *Journal of Molecular Biology*, 425(18), pp.3563–3575.

Rajan, S. et al., 2015. BH3 induced conformational changes in Bcl-Xl revealed by crystal structure and comparative analysis. *Proteins*, 83(7), pp.1262–1272.

Robin A. Y. et al., 2015. Crystal structure of Bax bound to the BH3 peptide of Bim identifies important contacts for interaction. *Cell Death Dis*. 6(7), p.e1809.

Sheng, Y. et al., 2012. A human ubiquitin conjugating enzyme (E2)-HECT E3 Ligase structure-function Screen. *Molecular and Cellular Proteomics*, 11(8), pp.329–341.

Shirts, M.R. et al., 2003. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *Journal of Chemical Physics*, 119(11), pp.5740–5761.

Shoemaker, B.A. and Panchenko, A.R., 2007. Deciphering protein-protein interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS computational biology*, 3(4), p.e43.

Smits, C. et al., 2008. Structural Plasticity Underpins Promiscuous Binding of the Prosurvival Protein A1. *Structure*, 16(5), pp.818–829.

Soni, N. and Madhusudhan, M.S., 2017. Computational modeling of protein assemblies. *Current Opinion in Structural Biology*, 44(6), pp.179–189.

Steinbrecher, T., Case, D.A. and Labahn, A., 2006. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. *Journal of Medicinal Chemistry*, 49(6), pp.1837–1844.

Stuart, J.M. et al., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643), pp.249–255.

Teichmann, S.A. and Babu, M.M., 2004. Gene regulatory network growth by duplication. *Nature genetics*, 36(5), pp.492–496.

Tuncbag, N., Keskin, O., Nussinov, R,, Gursoy, A., 2017. Prediction of Protein Interactions by Structural Matching: Prediction of PPI Networks and the Effects of Mutations on PPIs that Combines Sequence and Structural Information. *Methods Mol Biol*, 1558, pp.255–270.

Vajda, S. and Kozakov, D., 2009. Convergence and combination of methods in protein-protein docking. *Current Opinion in Structural Biology*, 19(2), pp.164–170.

Warwicker, J., 1999. Simplified methods for pKa and acid pH-dependent stability estimation in proteins: Removing dielectric and counterion boundaries. *Protein science: a publication of the Protein Society*, 8(2), pp.418–25.

van Wijk, S.J.L. et al., 2009. A comprehensive framework of E2-RING E3 interactions of the human ubiquitin-proteasome system. *Molecular systems biology*, 5(295), p.295.

Winter, C. et al., 2012. Protein interactions in 3D: From interface evolution to drug discovery. *Journal of Structural Biology*, 179(3), pp.347–358.

Xue, L.C. et al., 2015. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters*, 589(23), pp.3516–3526.

Yamaguchi, Y., Park, J.-H. and Inouye, M., 2011. Toxin-Antitoxin Systems in Bacteria and Archaea. *Annual review of genetics*, 45, pp.61–79.

Yin, Q. et al., 2009. E2 interaction and dimerization in the crystal structure of TRAF6. *Nature structural and molecular biology*, 16(6), pp.658–66.

Zhang, H. et al., 2000. Structural Basis of BFL-1 for Its Interaction with BAX and Its Anti- Apoptotic Action in Mammalian and Yeast Cells. *Journal of Biological Chemistry*, 275(15), pp.11092–11099.

Zhang, L. et al., 2011. The IDOL – UBE2D complex mediates sterol-dependent degradation of the LDL receptor. *Genes and Development*, 25(12), pp.1262–1274.

Zhang, W. et al., 2016. System-Wide Modulation of HECT E3 Ligases with Selective Ubiquitin Variant Probes. *Molecular Cell*, 62(1), pp.121–136.

Chapter 3 - Energetics and dynamics across the Bcl-2-family-dependent apoptosis pathway reveal distinct evolutionary determinants of specificity and affinity

PAPER 2

Stefan M. Ivanov[1,2], Roland G. Huber[1], Jim Warwicker[2], Peter J. Bond[1,3*]

1 Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Matrix 07-01, 30 Biopolis Street, 138671 Singapore

2 Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

3 Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

*corresponding author: peterjb@bii.a-star.edu.sg

SMI assembled the dataset; SMI and RGH performed the calculations; all authors analyzed the data and wrote the paper.

Published in *Structure*, Volume 24, Issue 11, Pages 2024-2033, 1 November 2016

Abstract

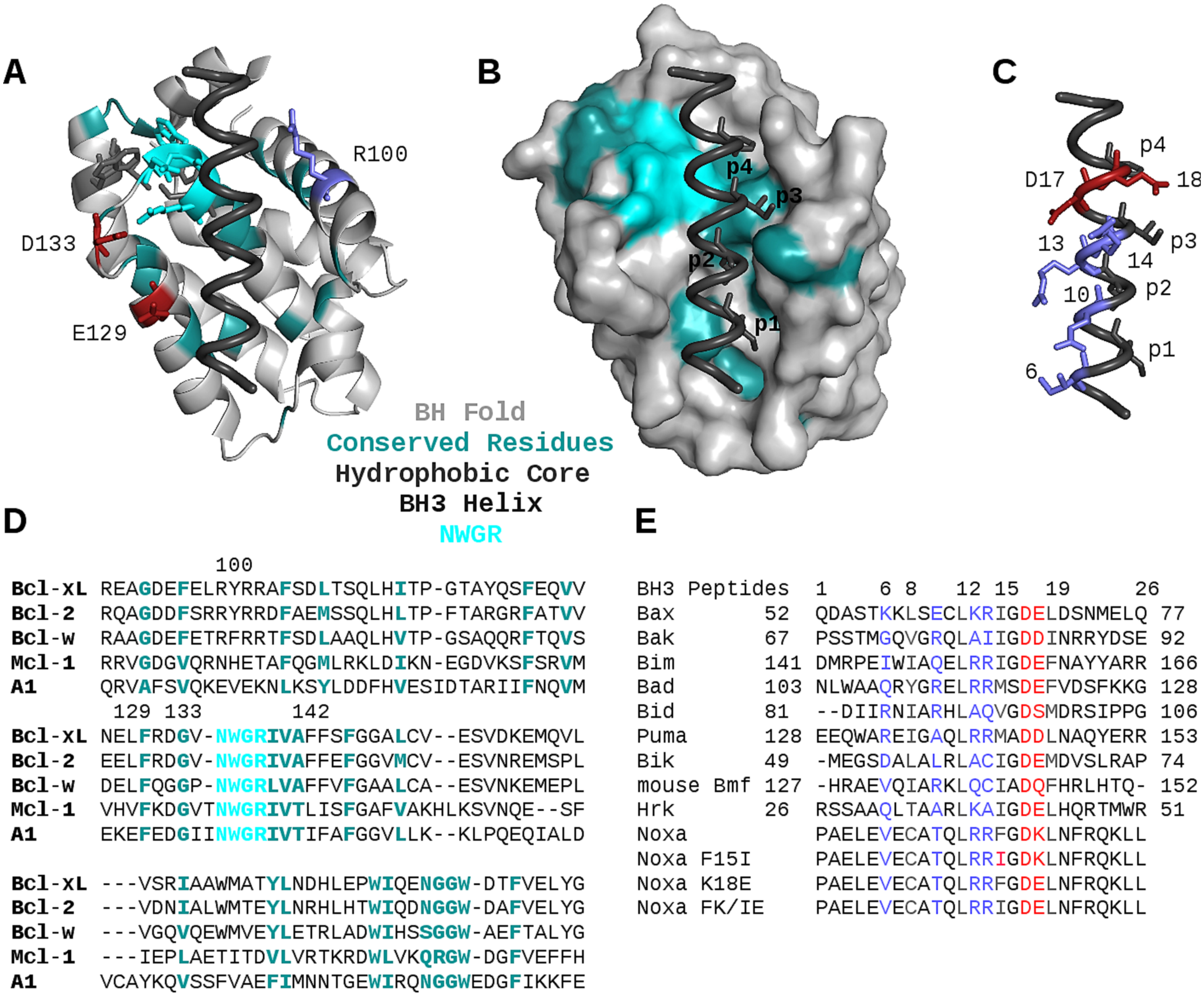
Critical regulatory pathways are replete with instances of intra- and interfamily protein – protein interactions, due to the pervasiveness of gene duplication throughout evolution. Discerning the specificity determinants within these systems has proven a challenging task. Herein, an energetic analysis of the specificity determinants within the Bcl-2 family of proteins (key regulators of the intrinsic apoptotic pathway) via a total of ∼20 μs of simulation of 60 distinct protein – protein complexes is presented. Further, it is shown where affinity and specificity of protein – protein interactions arise across the family, and conclusions are corroborated with extensive experimental evidence. Energy and specificity hot spots that may offer valuable guidance in the design of targeted therapeutics for manipulating the protein – protein interactions within the apoptosis-regulating pathway are identified. Finally, a conceptual framework that allows quantifying the relationship between sequence, structure, and binding energetics is proposed. This approach may represent a general methodology for investigating other paralogous protein – protein interaction sites.

3.1. Introduction

Most proteins belong to families of evolutionarily and functionally related molecules, often arising from gene duplication (Friedman and Hughes, 2001). A classic example of such paralagous proteins are the human kinases, numbering over 500 (Manning et al., 2002). The specificity of biological pathways is thus striking, considering the thousands of potentially interacting macromolecules in a cell at any given time (Berggård et al., 2007). In general, protein interaction sites consist of tightly packed, structurally conserved regions or “hot spots” (Shoemaker and Panchenko, 2007; Ma et al., 2003). Hot spots tend to be enriched in tryptophan, tyrosine and arginine (Ma et al., 2003), and the most frequent residue pairs in the associated protein – protein complexes involve charged and aromatic residues (Gromiha et al., 209; Gromiha et al., 20011). It has been suggested that polar residues at the interface cores confer rigidity, reducing the entropic loss upon binding, while the surrounding residues may form a “flexible cushion.” A study on paralogous protein – protein interfaces led to the proposal that binding affinity is provided mainly by the hub, whereas specificity is determined at the rim. Specificity between paralogs diverges at greatly differing rates, while interfaces evolve more slowly then the rest of the protein (Aiello and Caffrey, 2012). Explaining specificity within families of paralogs is particularly challenging, given that they usually share a common, conserved interface based on a conserved scaffold, for both interacting and non-interacting pairs (van Wijk et al., 2009; Kar et al., 2012).

This work focuses on the mechanisms by which a protein selects binding partners from a pool of closely related candidates, starting with the assumption that the decisive factors determining binding versus non-binding in paralagous protein pairs are alterations in and around a common scaffold. It focuses on the B-cell lymphoma-2 (Bcl-2) family of proteins, due to its great physiological and clinical importance, as well as the abundance of structural and interaction data (Chen et al., 2005). The intrafamily interactions among Bcl-2-like proteins determine whether a cell undergoes apoptosis (Cory et al., 2003). The Bcl-2 family encompasses the antiapoptotic molecules Bcl-xL, Bcl-2, Bcl-w, Mcl-1 and A1 (Cheng et al., 2001), and ~ 15 proapoptotic members. The antiapoptotic proteins have four Bcl-2 homology (BH) regions (BH1-4), as do the proapoptotic Bax and Bak (Kvansakul et al., 2008), which constitute a separate, Bax-like, subfamily. Most proapoptotic members (e.g. Noxa, Hrk, Bid, Puma, Bmf, Bik, and Bim) belong to the BH3-only subfamily (Happo et al., 2012).

Bax, Bak, and the antiapoptotic proteins consist of 7 or 8 amphipathic α-helices, clustered around a central hydrophobic α-helix (Suzuki et al., 2000), forming an exposed hydrophobic groove for binding the BH3 domain of proapoptotic proteins (Figure 3.1) (Petros et al., 2004). The core fold has 85-95 % structural overlap (Nguyen et al., 2011) across deposited structures in the PDB (Berman et al., 2000), and contains highly conserved regions including an invariant NWGR motif at the beginning of helix 5 (Day et al., 2008), and a conserved hydrophobic core which maintains the tryptophan in its position (Figure 3.1A).



```
                100
Bcl-xL REAGDEFELRYRRAFSDLTSQLHITP-GTAYQSFEQVV
Bcl-2  RQAGDDFSRRYRRDFAEMSSQLHLTP-FTARGRFATVV
Bcl-w  RAAGDEFETRFRRTFSDLAAQLHVTP-GSAQQRFTQVS
Mcl-1  RRVGDGVQRNHETAFQGMLRKLDIKN-EGDVKSFSRVM
A1     QRVAFSVQKEVEKNLKSYLDDFHVESIDTARIIFNQVM
        129 133        142
Bcl-xL NELFRDGV-NWGRIVAFFSFGGALCV--ESVDKEMQVL
Bcl-2  EELFRDGV-NWGRIVAFFEFGGVMCV--ESVNREMSPL
Bcl-w  DELFQGGP-NWGRLVAFFVFGAALCA--ESVNKEMEPL
Mcl-1  VHVFKDGVTNWGRIVTLISFGAFVAKHLKSVNQE--SF
A1     EKEFEDGIINWGRIVTIFAFGGVLLK--KLPQEQIALD

Bcl-xL ---VSRIAAWMATYLNDHLEPWIQENGGW-DTFVELYG
Bcl-2  ---VDNIALWMTEYLNRHLHTWIQDNGGW-DAFVELYG
Bcl-w  ---VGQVQEWMVEYLETRLADWIHSSGGW-AEFTALYG
Mcl-1  ---IEPLAETITDVLVRTKRDWLVKQRGW-DGFVEFFH
A1     VCAYKQVSSFVAEFIMNNTGEWIRQNGGWEDGFIKKFE
```

```
BH3 Peptides  1    6 8   12 15  19     26
Bax       52  QDASTKKLSECLKRIGDELDSNMELQ 77
Bak       67  PSSTMGQVGRQLAIIGDDINRRYDSE 92
Bim       141 DMRPEIWIAQELRRIGDEFNAYYARR 166
Bad       103 NLWAAQRYGRELRRMSDEFVDSFKKG 128
Bid       81  --DIIRNIARHLAQVGDSMDRSIPPG 106
Puma      128 EEQWAREIGAQLRRMADDLNAQYERR 153
Bik       49  -MEGSDALALRLACIGDEMDVSLRAP 74
mouse Bmf 127 -HRAEVQIARKLQCIADQFHRLHTQ- 152
Hrk       26  RSSAAQLTAARLKAIGDELHQRTMWR 51
Noxa          PAELEVECATQLRRFGDKLNFRQKLL
Noxa F15I     PAELEVECATQLRRIGDKLNFRQKLL
Noxa K18E     PAELEVECATQLRRFGDELNFRQKLL
Noxa FK/IE    PAELEVECATQLRRIGDELNFRQKLL
```

Figure 3.1. Summary of conserved structure and sequence properties across the Bcl-2-family-dependent apoptosis pathway. (A) Bcl-xL – Bim complex in cartoon representation with key residues

in stick representation – NWGR motif (cyan), hydrophobic core around the tryptophan (dark gray), E129 and D133 (red), and R100 (blue). **(B)** Same complex in surface representation for Bcl-xL; also labeled are the 4 hydrophobic pockets and the peptide residues (in stick representation) that fit into them. **(C)** Bim BH3 peptide with key residues labeled and in stick representation. **(D)** Sequence alignment of the fold-forming portions of the 5 antiapoptotic proteins with the most conserved regions highlighted. Protein residues are referred to by their canonical Uniprot numbering (Apweiler et al., 2004); numbering in the figure corresponds to Bcl-xL. **(E)** Sequence alignment of the BH3 peptides used in this study and their location in the full-length proteins. Pocket residues (positions 8, 12, 15, and 19) are highlighted in gray, positions 6, 10, and 13 are highlighted in blue, and positions 17 and 18 are in red. All sequences are human, except Bmf, which is from mouse. All sequences are identical to the canonical sequences, deposited in Uniprot, except for a single mutation in Hrk (L15I). The sequences are identical to the ones used during pIC_{50} measurements, except for Bax. In the Bax affinity measurements, the authors used 34-mer peptides (Fletcher et al., 2008), whereas a 26-residue-long Bax BH3 peptide was simulated here.

In preapoptotic cells, the BH3 domains of proapoptotic Bak and Bax (Shamas-Din et al., 2011) are bound to the hydrophobic groove on the surface of the antiapoptotic proteins, rendering them inactive (Stewart et al., 2010). When an apoptosis signal reaches the cell, BH3-only proteins outcompete Bak and Bax for their antiapoptotic partners, freeing the formers' BH3 domains, which are then involved in homo- and possibly heterodimerization via a BH3 domain – hydrophobic groove interaction, leading to oligomeric pore formation in the outer mitochondrial membrane and subsequent apopotosis (Happo et al., 2012).

The binding mode between different pairs of proteins within the system is highly similar (Day et al., 2008; Smits et al., 2008; Czabotar et al., 2007): all BH3 peptides have four hydrophobic residues (positions 8, 12, 15, and 19, see Figure 3.1E) that fit into four hydrophobic pockets (labeled p1 – p4, see Figure 3.1B) on the surface of the groove, whilst an absolutely conserved aspartic acid (position 17) in the proapoptotic proteins forms a salt bridge with the arginine of the NWGR motif. Nevertheless, the affinities between the different BH3 peptides and the five antiapoptotic proteins span more than four orders of magnitude – from IC_{50} values below 5 nM to >100 μM (Chen et al., 2005). As all antiapoptotic proteins in a cell must be neutralized for it to undergo apoptosis and not all BH3 peptides are omnibinders, their binding selectivity has implications for peptide-micking drugs that target this interaction (Czabotar et al., 2014).

In order to elucidate the origins of affinity and specificity across a paralogous set of interacting and non-interacting pairs, a computational study of the Bcl-2 family was performed. A total of 60 different complexes were modeled, using a template of a BH3 peptide (or “ligand”) bound to an antiapoptotic protein (or “receptor”), guided by a dataset of experimentally measured affinities (see Table 3.1 and Figure 3.1). For each complex, and also for the constituent isolated ligands/receptors, triplicate MD simulations were carried out (amounting to 180 x 100-ns complex trajectories, 15 x 100-ns receptor trajectories, and 39 x 100-ns ligand trajectories), enabling accurate calculation of the enthalpies of each protein – protein interaction and decomposition on a per-residue basis. It is demonstrated that in antiapoptotic proteins, pockets provide affinity, but not specificity. Energetic recognition patterns are shown to be the most adaptable feature in a hierarchy of structure, sequence, and energy conservation. It is then posited that the groove – BH3 helix case discussed here may be representative of a general pattern on the relationship between structure, sequence, and binding energetics in protein families. Thus, a method to characterize energy and specificity hot spots that can be utilized in targeting paralogous protein – protein interactions is presented.

pIC_{50}, [M]	**Bcl-xL**	**Bcl-2**	**Bcl-w**	**mouse Mcl-1**	**mouse A1**
Bax	6.89	7.00	7.23	7.92	N/A
Bak	7.30	< 6.00	6.30	8.00	N/A
Bim	> 8.30	> 8.30	> 8.30	> 8.30	> 8.30
Bad	8.28	7.80	7.52	< 4.00	4.82
Bid	7.09	5.17	7.40	5.68	8.03
Puma	8.20	> 8.30	8.29	> 8.30	8.24
Bik	7.37	6.08	7.92	5.77	7.24
mouse Bmf	8.01	> 8.30	8.01	5.96	5.74
Hrk	> 8.30	6.49	7.31	6.43	7.34
Noxa	< 4.00	< 4.00	< 4.00	7.22	6.74
Noxa K18E	5.30	< 4.00	4.05	7.46	N/A
Noxa F15I	6.00	< 4.00	5.00	7.60	N/A
Noxa FK/IE	6.96	4.96	6.30	7.62	N/A

Table 3.1. pIC50 values for different BH3 peptide – antiapoptotic protein interactions. All sequences are human, except where explicitly stated otherwise. Bax data is from Fletcher et al. (2008); Bak data is from Willis et al. (2005); the remaining data is from Chen et al. (2005).

3.2. Results

3.2.1. Structural stability of the modeled complexes

Cα RMSD values for all simulations indicated that the complexes were stable (Supplementary information figure 3.1A, see Appendix), along with the core regions of receptor (SI figure 3.1B) and ligand (SI figure 3.1C, D). For the antiapoptotic protein components, structural variability was concentrated primarily in the loops connecting the helices (SI figure 3.1B). Ligands tended to display higher RMSD values, but the increased dynamics originated from the termini (SI figure 3.1C). If RMSDs between positions 8 and 20 are considered, RMSD values generally tend to vary within a small window of around 0.3 Å with mean values between 0.2 – 0.7 Å (SI figure 3.1D). Moreover, RMSD variations between replicas were small for the majority of complexes. When simulated in isolation, the receptors maintained their structure (SI figure 3.1E), whereas the peptides unfolded in agreement with experiment (Chen et al., 2005). In order to optimize the signal/noise ratio for the energy calculations, subsequent analysis was based on the latter 60 ns of each trajectory.

3.2.2. Energetic basis for protein – protein affinities

Molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations were then employed to calculate the enthalpy of binding (ΔH) for each replica, and decompose the results on a per-residue basis. In order to discern the origins of affinity and specificity, the per-residue ΔH contributions across the whole set of simulations were analyzed. Comparisons of the means and variances of per-residue ΔH observed in the trajectory sets for the five receptors, each interacting with the same set of ligands, were made. If a residue consistently contributes a high ΔH value with low variance, this indicates that it is an important site for generating affinity. Conversely, residues that show a high variance across the set of interactions with different ligands are likely to be involved in determining binding specificity. The mean per-residue ΔH values and their variances were then mapped onto the surface of each complex, in order to discern the main contributors to affinity and specificity for the five receptor – ligand sets (Figure 3.2A and 3.2B, respectively).

It is evident that for the Bcls (Bcl-xL, Bcl-2, and Bcl-w) and Mcl-1, affinity originates predominantly from the region around the NWGR motif of the receptor, particularly the arginine, which forms a salt bridge with the aspartic acid in position 17 of the ligand. Furthermore, in the Bcls, there exists a conserved glutamic acid (E129/E136/E85, respectively), which contacts ligand positions 6, 10, and 13, which are typically positively charged or polar (see Figure 3.1). Correspondingly, for the peptides bound to Bcl-xL, Bcl-2, and Bcl-w, it is these three residues, along with D17, that are the greatest contributors to affinity. In Mcl-1 and A1, the glutamic acid has been substituted by H233 and K77, respectively. As evident from Table 3.2, the most conserved residues account for around 45 – 55% of total receptor contribution to binding, with the NWGR motif alone responsible for 25 – 35%.

	Bcl-xL	**Bcl-2**	**Bcl-w**	**mouse Mcl-1**	**mouse A1**
Conserved Residues	53%	44%	48%	55%	56%
NWGR Motif	31%	25%	28%	35%	29%

Table 3.2. Energetic contributions to binding. Energetic contributions to binding (as a percentage of total receptor contribution) for the conserved residues (highlighted in Figure 3.1) and the NWGR motif. Data shown are averages over 39 trajectories for Bcl-xL, Bcl-2, Bcl-w, and Mcl-1 (13 ligands x 3 replicas) and 24 for A1 (8 ligands x 3 replicas).

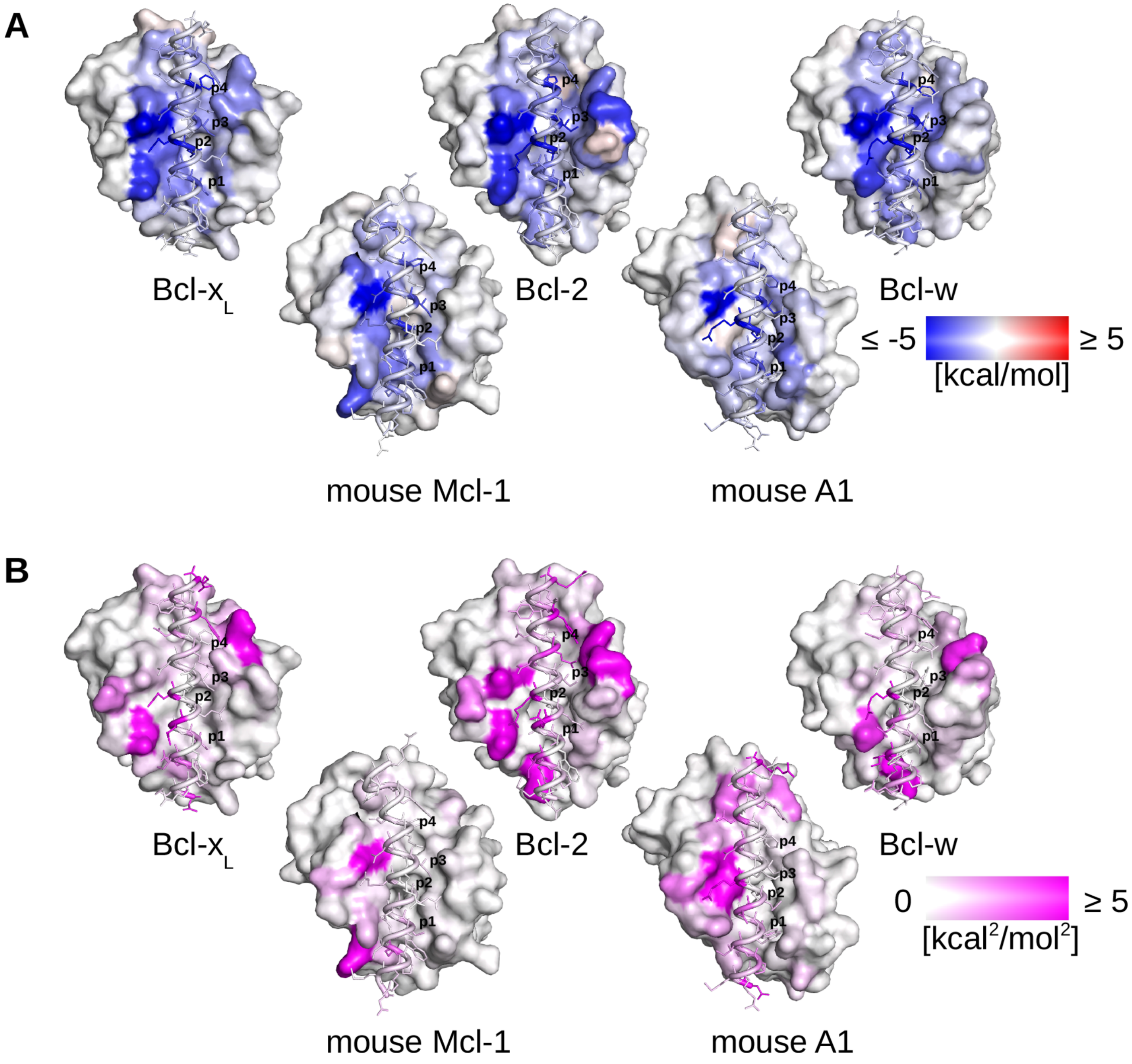


Figure 3.2. Sources of affinity and specificity assessed via energetics analysis, based on protein – peptide complex trajectories. (A) Antiapoptotic protein – BH3 peptide complexes colored by average per-residue ΔH values. **(B)** Antiapoptotic protein – BH3 peptide complexes colored by the variance of per-residue ΔH values. Averages and variance were calculated across 39-trajectory sets for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 (13 ligands x 3 replicas), and across 24 trajectories for mouse A1 (8 ligands x 3 replicas). Ligand N-termini are at the bottom of the figures, C-termini are at the top. ΔH was computed from complex trajectories only.

3.2.3. Energetic basis for protein – protein specificities

For the Bcl receptors specificity is greatest at the rim around pockets 3 and 4, and a patch centered at the conserved glutamate in the receptor, E129/E136/E85 (Figure 3.2B; see Figure 3.1). For the ligands, specificity is highest at the N-terminal half of the peptides, at positions 6, 10, and 13, which contact this patch, and position 18, which contacts the aforementioned rim. In Mcl-1 and A1, the rim is much shallower, especially around pocket 4 (Czabotar et al., 2007), and is a lot less discriminating than in the other antiapoptotic proteins, whereas the NWGR motif and its adjacent residues appear to take on a greater role in determining specificity, as they contact ligand residues 16, 19, and 20. Due to the increased ligand flexibility in the absence of a receptor, the results from MM-PBSA calculations on complex, receptor, and ligand trajectories (the "three-trajectory" approach) point to a greater number of residues being involved in determining specificity then the MM-PBSA data relying solely upon complex trajectories. A complete sampling of ligand conformations in isolation would require orders of magnitude longer dynamics than could typically be accessed computationally. Importantly, however, the results from the three-trajectory MM-PBSA calculations are consistent with the forgoing data on specificity and affinity (SI figure 3.2).

As previously stated, ligand residues 6, 10, and 13 contact a conserved glutamic acid in Bcl-xL, Bcl-2, and Bcl-w (E129/E136/E85, respectively). When two of those positions are positively charged (see Figure 3.1B), this allowed the formation of a highly favorable salt-linked triad (Horovitz et al., 1990) between them and the glutamic acid. Moreover, when the remaining residue is also capable of hydrogen bonding to this glutamic acid, the latter hydrogen bond became coupled to the triad, further strengthening binding (Figure 3.3).

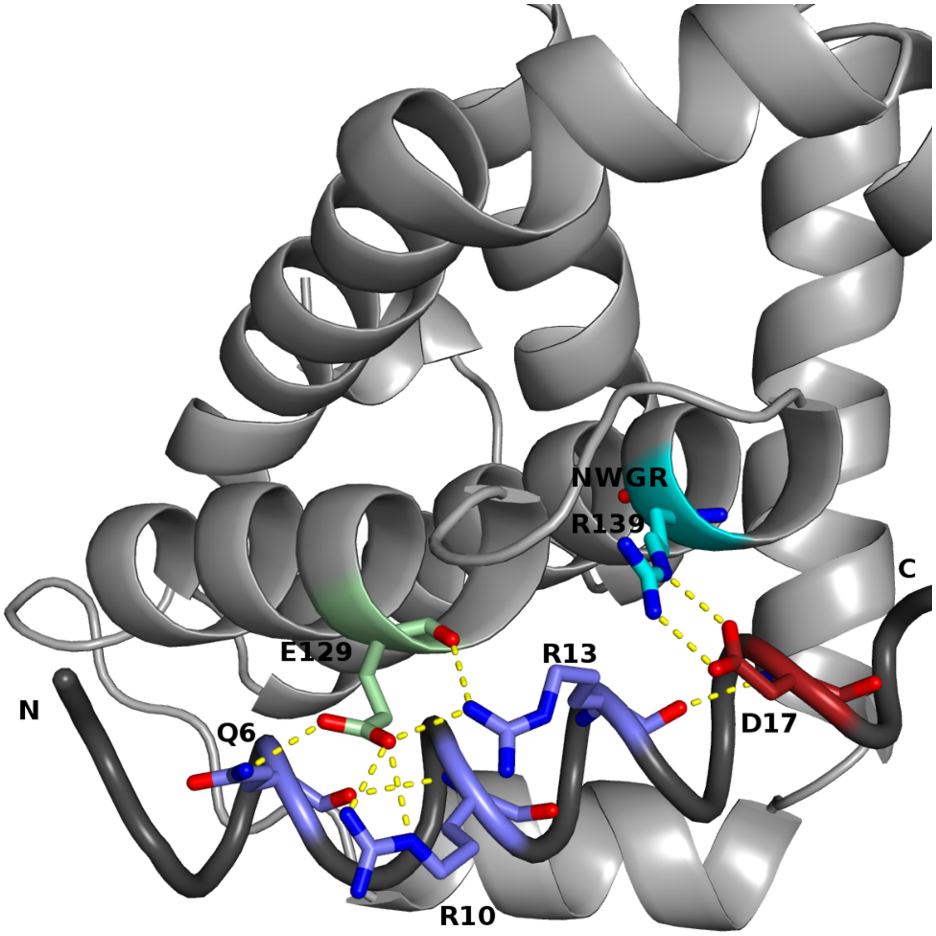


Figure 3.3. Key interactions highlighted in a snapshot from a Bcl-xL – Bad trajectory. The complex is in cartoon representation with Bcl-xL colored gray, Bad colored dark gray, and key residues in stick representation. Q6, R10, and R13 of Bad are in blue, D17 is in red, E129 of Bcl-xL is in green, and R139 (from the NWGR motif) is in cyan, with nitrogen atoms in blue and oxygen atoms in red. Also labeled are the peptide termini. Bcl-xL residue E129 simultaneously forms three salt-linked triads with 6Q, 10R, and R13 of Bad. Additionally, R13 simultaneously hydrogen bonds to the side chain and backbone of E129. The key D – R salt bridge is also present.

Although position 14 remained oriented towards the solvent throughout most of the simulations, it is possible that it may also participate in binding through E129/E136/E85 or D133/D140/G89 (see Figure 3.1). Interestingly, the side chain of R13 in the ligand could simultaneously hydrogen bond to the backbone and side chain of the glutamic acid residue (Figure 3.3). Positively charged residues, especially KR and RR combinations for positions 13 and 14, are commonly found in these positions. In Mcl-1 and A1, the glutamic acid has been substituted by histidine and lysine, respectively, greatly reducing the hydrogen bonding potential between ligand and receptor. Consequently, in the Mcl-1 and A1 – ligand trajectories, R13 could only form hydrogen bonds with receptor backbone atoms, resulting in much less favorable interactions with the antiapoptotic protein. The importance of the 6–10–13 – receptor residue coupling is also reinforced by the fact that all weak binders (i.e., peptides in receptor-ligand complexes with pIC50 < 6 M) have one or more residues in positions 6, 10, or 13 which are incapable of participating in an interaction with this key receptor residue (see Figure 3.1 and Table 3.1).

3.2.4. Energetic correlation analysis

ΔH values for each ligand position were correlated with every other across the five trajectory sets. Correlating ΔH values for peptide positions 1 through 26 among each other reveals that in ligands bound to human Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1, there seem to exist two regions of energetic correlation (Figure 3.4 and SI figure 3.3). The first one extends up to around position 15, which fits into pocket 3. Past that, there is a C-terminal region of somewhat weaker energetic correlation. It is possible that this is due to the 6–10–13 and the 16–19–20 couplings (the latter of which is achieved through the NWGR motif and its adjacent residues coming into contact with the ligand residues), and the clamping effect exerted on the bound peptides by the protein rim. In A1, however, there appears to be an almost uninterrupted region of helix-like energetic correlation spanning most of the peptide length (Figure 3.4B). This is likely because the rim in A1 is much shallower, particularly around pocket 4, making ligand structure and properties more pronounced and important for binding A1 than the other antiapoptotic proteins. This implies that helix stability *per se* would offer greater gains in affinity to A1 than the other proteins. Given that Mcl-1's rim is shallower than those of the Bcls, but less so than A1, it can be anticipated that helix stability will have an effect intermediate in magnitude between those in A1 and the Bcls. Indeed, in two Mcl-1 trajectories and five A1 trajectories, disengagement of ~10 C-terminal peptide residues from the proteins was observed. Although the three-trajectory MM-PBSA results are somewhat harder to interpret, they are consistent with these findings (SI figure 3.3).

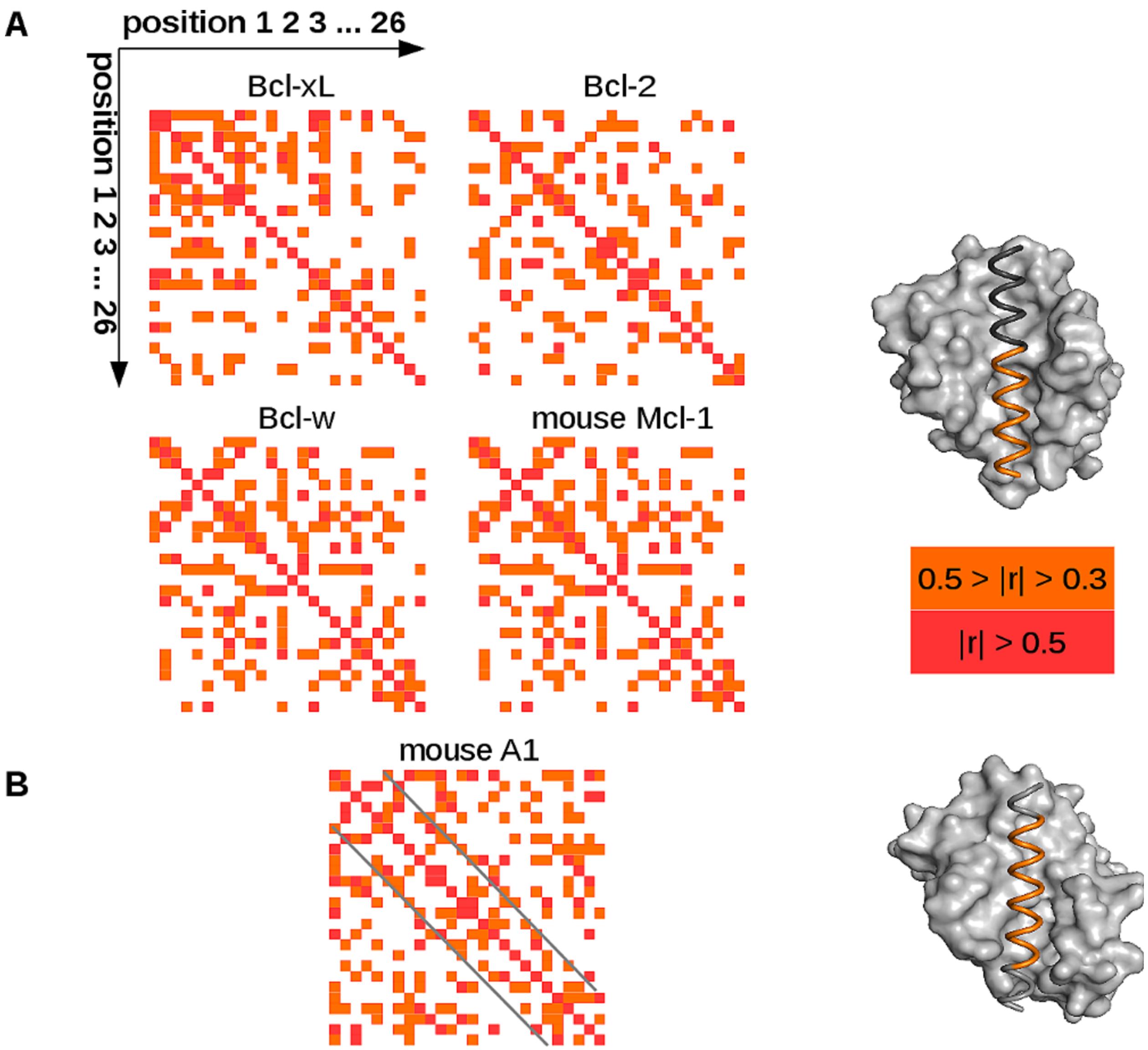


Figure 3.4. Energy correlation analysis, based on protein – peptide complex trajectories. (A) Energy correlation analysis performed among the 26 ligand residues across the four 39-trajectory sets (13 ligands x 3 replicas). BH3 ligands seem to display two regions of energetic correlation – an N-terminal one, spanning up to around position 15 (colored in orange in the structure to the right), and a C-terminal one (colored in gray). **(B)** Energy correlation analysis performed among the 26 ligand residues across the 24-trajectory set for A1 (8 ligands x 3 replicas). BH3 ligands seem to display an almost uninterrupted region of helix-like energetic correlation, spanning most of the peptide length (colored in orange in the structure to the right). ΔH was computed from complex trajectories only.

3.3. Discussion

3.3.1. Comparison to experimental data

In this study, the origins of affinity and specificity within a family of proteins have been systematically investigated by a careful analysis of binding energetics across a diverse set of complexes. Moreover, it is shown how the behavior of ligands differs according to which receptor they are complexed with. A caveat of the present analysis is that it has been performed exclusively on homology models. However, they are in excellent agreement with multiple existing structures (RMSDs ~0.4 – 1 Å), with recently published ones (Robin et al., 20015; Kim et al., 2015; Rajan et al., 2015; Jenson et al., 2017) only reinforcing confidence in the models used. Other potential limitations are the limited sampling afforded by explicit solvent simulations, the fidelity of force field parameters, and the reliability of MM-PBSA results, omitting entropic contributions (Hansen and van Gunsteren, 2014). Nevertheless, the results are in good agreement with multiple experimental studies and provide the first quantitative assessment across the family of the contributions of different regions in each receptor and ligand to binding. For example, most of the receptor residues deemed critical to BH3 peptide binding in an alanine scan study (Campbell et al., 2015), all of which are highly conserved, are prominent contributors to binding in the present energetic analysis. That study and others (Day et al., 2008; Ku et al., 2011; Fletcher et al., 2008) have shown that the D17 – R (from NWGR) interaction is critical in multiple peptide – protein pairs, in accord with results presented herein, which show that typically it is the greatest single contributor to binding. The significance of the 6–10–13 coupling through the E129/E136/E85/H233/K77 residue is clearly demonstrated by the observation that mutating the glutamate in the Bcls is detrimental to BH3 binding, whereas mutating the corresponding histidine in Mcl-1 to alanine strengthens binding to peptides which carry positive charges in positions 6, 10, and/or 13 (Campbell et al., 2015).

Mutating Bim residues 6 and 10 to glutamate strengthens binding to Mcl-1, whereas the I6E mutation weakens binding to Bcl-xL; Q10E has little effect on Bcl-xL binding, likely because of salt-linked triad formation, as shown in Figure 2.1. Mutating Bim positions 13 and 14 to glutamate weakens binding to Mcl-1. This is likely because they contact a highly conserved aspartate located four positions C-terminal to H233 - D237. This aspartate is highly conserved among all antiapoptotic proteins except

Bcl-w (Figure 3.1D). However, the R13E and R14E substitutions practically abolish Bim binding to Bcl-xL (Boersma et al., 2008), suggesting another route to the design of Mcl-1 selective peptides and peptidomimetics (Smits et al., 2008; Lee et al., 2008).

Mutating Bim position 13 to an acidic residue weakens binding to A1, rather than enhancing it (DeBartolo et al., 2012). This is likely due to the aforementioned aspartate (D81 in mouse A1), as well as a unique feature of A1, residue E78, which is involved in forming pocket 2 and is buried in all human and murine A1 – BH3 X-ray structures (Herman et al., 2008). This residue is a leucine in the Bcls (L130 in Bcl-xL, see Figure 3.1D) and a valine in Mcl-1. Indeed, it is the only pocket-forming residue with a high variance in ΔH values (Figure 3.2B and SI figure 3.2B). The simulations reported herein demonstrate that positions 10 and 13 are in greater proximity to D81 and this glutamate, rather than the preceding lysine, and that position 6 appears in a more favorable position to interact with K77. Thus, it is to be anticipated that an acidic residue in position 6 would either strengthen binding to mouse A1 or at least offer greater selectivity for A1 than Bcl-xL, Bcl-2, and Bcl-w. Moreover, it is to be expected that acidic residues in positions 10 and 13 would cause a greater decrease in affinity for the Bcls than A1, opening up an avenue for the design of A1-selective molecules. Finally, these suggested mutations should have an effect on binding affinity towards Mcl-1, which is intermediate in magnitude between A1 and the Bcls.

Bad is the only BH3 sequence that does not bind Mcl-1. Moreover, its affinity for A1 seems to be only slightly above the detection limit of the affinity measurements (see Table 3.1; Chen et al., 2005). This is likely because of the positive charge in 6-10-13 (greatest among all the ligands), which is paired with H233/K77 in Mcl-1/A1, and the peculiarity of Bad residues 16 and 20, which are unique. In particular, all peptides have a glycine or an alanine in position 16, except Bad, which has a serine. Its side chain is in proximity to that of T247/T91 (in Mcl-1 and A1, respectively) and the NWGR motif and several adjacent residues, which helps explain why serine seems to be disfavored at this position whereas alanine and glycine, in particular, are favored. T247/T91, located three positions C-terminal to the NWGR motif, seem to be more restrictive of binding than the corresponding alanines in Bcl-xL, Bcl-2, and Bcl-w (A142/A149/A98, Figure 3.1D), as those proteins better tolerate mutations to serine in peptide position 16. Indeed, the packing in this region is very dense, which is likely the reason mutating position 16 to any other residue weakens binding (DeBartolo et al., 2012) and mutating the glycine from NWGR even to alanine abolishes antiapoptotic activity (Yin et al., 1994; Sedlak et al., 1995). Moreover, Bad has a valine in position 20, unlike any of the other BH3 sequences under study, which have polar or charged residues in this position (D, N, or H). In the present simulations, G245 of Mcl-1 is involved in an intermolecular N-capping interaction with the ligand residue in position 20, helping maintain the ligand tethered to the receptor. Other authors have described this N-capping interaction as well (Day et al., 2008). In the receptor – Bad trajectories, where a valine stands at position 20, however, no such interaction is possible and in two of the Mcl-1 and A1 simulations the C-terminus disengages from the receptor. This led to the breaking of the key D – R salt bridge, which is the reason position 17 and the arginine from NWGR in Mcl-1 and A1 appear variable in terms of energetics. Experimental evidence also demonstrates that the antiapoptotic proteins have a high preference for polar and charged residues in ligand position 20, with Mcl-1 (data not available for A1) being particularly selective for D, E, H, and N (DeBartolo et al., 2012). It is to be expected that A1 will display an identical preference and that this heightened selectivity in Mcl-1 is due to the shallowness of the rim, which makes the NWGR motif and its adjacent residues critical in terms of providing affinity and, as a consequence, specificity.

Results suggest that in Bcl-xL, Bcl-2, and Bcl-w, the rim around pockets 3 and 4 provides more specificity than affinity (see Figure 3.2 and SI figure 3.2). This is corroborated by experiments which demonstrate that mutating Noxa residue 18, which contacts the foregoing rim, from a lysine to a glutamate transforms Noxa from a non-binder to a weak binder to Bcl-xL and Bcl-w. It seems that this mutation alone is not enough to achieve detectable binding to Bcl-2 (see Table 3.1). Typically, position 18 is an acidic residue, which contacts R100/R107/R56 from the rim in Bcl-xL, Bcl-2, and Bcl-w. Only in Noxa is position 18 positively charged (see Figure 3.1E). Notably, Noxa is the only ligand that does not bind to these three proteins. In Mcl-1 and A1, the arginine has been mutated to N204 or E47, respectively.

For the Bcls in isolation, the calculated RMSD values seemed to be slightly higher than the complexed molecules, hinting at the stabilizing effect the peptides exert when bound (SI figure 3.1E). This has been observed previously for Bcl-xL (Guo et al., 2015). Compared to the Bcls, mouse Mcl-1 and A1 seem to be more stable in isolation, which agrees with the observation that they experience very little backbone conformational changes when binding different BH3 peptides (Day et al., 2008; Smits et al., 2008; Day et al., 2005), contrasting with Bcl-xL's notable structural plasticity (Lee et al., 2009; Moldoveanu et al., 2014). The ligands unfolded when not bound, in agreement with circular dichroism data (Chen et al., 2005).

3.3.2. Comparison to other computational and structural studies

It has previously been observed that helix stability is a factor contributing to affinity (Modi et al., 2012). Based on present simulations and energy correlation analysis, it may be added that C-terminal helicity contributes to binding by stabilizing the D17 – R (from the NWGR motif) and position 19 – pocket 4 interactions. Correspondingly, lower helix stability would facilitate the loss of these intermolecular interactions and would decrease binding affinity. Similarly, N-terminal stability of the peptide helix would help maintain peptide – receptor interactions in this region and the key hydrophobic residue – pocket 1 interaction. From the forgoing analysis of critical interactions, it is to be expected that Bad mutations S16G and V20N should enhance binding to Mcl-1 and A1, as would mutating residues H233 (Mcl-1) and K77 (A1) to acidic amino acids. Further, it is to be expected that mutations in the key acidic residues in the three Bcls (E129/E136/E85) should weaken or completely abolish binding to most of the BH3 domains reviewed here. Moreover, one might anticipate that mutating R100/R107/R56 in the Bcls to acidic amino acids would weaken binding to the peptides with an acidic residue in position 18 and strengthen binding to Noxa, which has a lysine in this position. Lastly, the E47K or E47R mutations in A1 should decrease affinity for Noxa and enhance binding to most of the remaining peptides.

A detailed analysis of the specificity determinants and energetic contributions for the groove – BH3 peptide interaction has been presented, discussing energies in relative, rather than absolute, terms, so as to make conclusions insensitive to the choice of MM-PBSA parameters. An important conclusion to be drawn from this work is that the highly conserved pockets provide affinity, but little to no specificity. Aiello and Caffrey (2012) previously investigated the balance between functionally conserved (i.e. binding the same ligand) and divergent interfaces in structural terms. Their analysis found that optimized hydrogen bonding networks in the rim regions of the binding pocket are important in specific interfaces, whereas functionally conserved interfaces tend to draw a larger portion of their total affinity from the central hub region. Their conclusions are consistent with the energetic analysis reported here.

3.3.3. Energetic mapping

The wealth and fine-grained nature of the energy data presented in this study allows one to explore the connection between conservation of sequence and of binding energetics. All investigated complexes have a similar fold and binding mode. Hence, observed correlations directly relate sequence to energy. In order to quantify these relations, an “energetic fingerprint” for each complex was constructed (see the Experimental procedures section). These energetic fingerprints were then correlated among simulations, grouped either by common ligand (SI figure 3.4A) or by common receptor (SI figure 3.4B). These similarity maps of energies were then compared to maps of sequence identity (SI figure 3.4, green). Thus, a (semi)quantitative approach that reveals to what degree similarity of sequence results in similarity of binding energetics is obtained (SI figure 3.4).

Careful inspection of the plots reveals that there are cases with a strong link between sequence and energy similarity (e.g. ligands Bak, Bim, Bad, Puma, Bmf and Noxa, receptors Bcl-xL, Bcl-w). However, in several cases, such a direct link is less apparent (e.g. ligands Bax, Bik, Bid and Hrk, receptor Mcl-1). The absence of a strong correlation in some cases allows one to rationalize the efficiency of gene duplication as a means by which specific pathways emerge. Although greater divergence in sequence is usually accompanied by greater divergence in the interaction energy patterns, in some cases even slight changes in sequence can lead to large changes in interaction patterns. From an evolutionary perspective, this discontinuity could rapidly alter the specificity or promiscuity of an interface. This would indicate that energetic recognition patterns are the most adaptable feature in a hierarchy of structure, sequence, and energy conservation. The groove – BH3 peptide example presented here is likely to be a manifestation of a more general pattern on the relationship between structure, sequence, and binding energetics. Indeed, instances where a pool of structurally similar small molecules/peptides/proteins bind a well defined region on a set of structurally similar protein partners are found in all domains of life and physiological pathways (Friedman and Hughes, 2001). The present work provides an attractive framework for investigating in a similar manner other physiologically and therapeutically relevant systems, e.g. the bZIP transcription factors (Nair and Burley, 2003) and EGF receptors (Arkhipov et al., 2014), which have been implicated in malignant cellular proliferation; histidine kinase – response regulator protein interactions, central to signal transduction in bacterial cells (Casino et al., 2009); Toll-like receptors (Berglund et al., 2015) and MHC proteins (Patronov et al., 2012; Ivanov et al., 2013), both of which regulate immunity; and the E2 – E3 enzyme interaction, part of the ubiquitination pathway (Kar et al., 2012).

3.4. Experimental procedures

BH3 peptides from human Bim, Bad, Bid, Puma, Bik, Hrk, Noxa, and three Noxa mutants, as well as mouse Bmf were modeled bound to human Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 and A1 (see Table 3.1 and Figure 3.1). Additionally, BH3 peptides from human Bax and Bak were modeled with human Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1. The following template structures were used: 2XA0 (Ku et al., 2011), 4CIM (Lee et al., 2014), 3PL7 (Czabotar et al., 2011), 2ROC (Day et al., 2008), and 2VOF (Smits et al., 2008). Terminal BH3 residues, missing from the templates, were modeled in using MODELLER 9.14 (Webb and Sali, 2014). Any mutations in the template antiapoptotic proteins were reverted back to wild type; BH3 sequences were modeled onto the BH3 template using in-house code. Briefly, the positions of backbone atoms were kept fixed, as were side chains in residues identical between model and template. Side chains for non-identical residues were re-packed (Bougouffa and Warwicker, 2008) using an adaptation (Cole and Warwicker, 2002) of a self-consistent mean-field method for rotamer selection from a rotamer library (Koehl and Delarue, 1994).

The resulting complexes were solvated with TIP3P water (Jorgensen et al., 1983) using the *tleap* module of Amber14 (Case et al., 2005) with a minimum wall distance of 12 Å. NaCl was added to neutralize system charge, to a concentration of 0.15 M. After 1,000 steps of minimization, the systems were gradually heated from 0 to 300 K over a period of 150 ps, applying weak restraints to protein and peptide heavy atoms. A 150 ps density-equilibration with restraints was followed by 2 ns of unrestrained constant pressure equilibration at 300 K. The protonation state of the solute and ionic strength and temperature of the system were set to match the conditions under which the pIC_{50} values were obtained. 100 ns of production dynamics were then carried out in triplicate at a pressure of 1 bar and a temperature of 300 K, maintained with the Berendsen barostat and Langevin thermostat. Bonds to hydrogen were constrained using the SHAKE algorithm (Ciccotti and Ryckaert, 1986), thus allowing for a 2 fs time step. An 8.0 Å cutoff was used for Lennard-Jones interactions, and long-range electrostatics were computed with the Particle mesh Ewald scheme (Darden et al., 1993). All simulations were carried out using the ff14SB force field (Maier et al., 2015); trajectories were processed with cpptraj *V14.25* (Roe and Cheatham, 2013). An identical protocol was utilized to simulate the individual components of the complexes.

For each complex simulation, the enthalpy of interaction between the antiapoptotic protein and the bound BH3 helix was computed with the Amber14 MMPBSA.py script (Miller et al., 2012) using both the “one-trajectory” and “three-trajectory” approach. MM-PBSA calculations were performed using Bondi radii (Bondi, 1964) and default settings for the nonpolar decomposition scheme, surface tension, cavity offset, and external and internal dielectric constants. The setting for the ionic strength was adjusted to match the one used during IC_{50} measurements in the reference dataset (0.15 M). Per-residue energy decompositions were also performed, adding 1-4 energy terms to internal energy terms. For each 100 ns MD run, energy calculations were performed on the latter 60 ns of dynamics. Snapshots for PBSA calculations were taken every 6 frames (60 ps apart), producing 1,000 frames per trajectory.

As interest lay primarily in relative rather than absolute binding energies (Homeyer and Gohlke, 2012; Huber et al., 2013), entropy calculations were omitted. This decision is reinforced by published calorimetric data, which demonstrates that BH3 helix binding is an enthalpically driven process (Day et al., 2008). Finally, the means and variance of the per-residue ΔH values were computed for the 39-trajectory sets for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1, and the 24-trajectory set for mouse A1.

The absolute values of the computed energy terms are sensitive to the choice of atomic radii and nonpolar decomposition scheme in the MM-PBSA approach, whereas their relative values have been shown to be insensitive to these parameters (Kumari et al., 2014). The results reported herein support this conclusion and demonstrate that the difference in computed ΔH values for a trajectory using bondi and mbondi2 radii (Onufriev et al., 2004) is around 4 to 5 kcal/mol. The chosen scheme for computing ΔGnonpolar yielded ΔH values which are of similar magnitude to calorimetric data (Day et al., 2008) (~ -10 to -25 kcal/mol), whereas the alternative scheme, where ΔGnonpolar is linearly dependent upon solvent accessible surface area, significantly overestimated ΔH (~ -80 to -100 kcal/mol). This work, therefore, corroborates the benefit of decomposing ΔGnonpolar into a dispersive (attractive) and cavitation (repulsive) term (Tan et al., 2007).

For each complex, the per-residue interaction energies derived from the MM-PBSA calculations were represented as a ~ 150-dimensional vector. Analogously to ideas used to compare specificity patterns of proteases (Fuchs et al., 2013), the inner product of the respective vectors was calculated to quantify the similarity between different energy patterns. This measure is 1 if the patterns are identical, 0 if the patterns are orthogonal (i.e. no energy contributions are in common between paired patterns), and -1 if the patterns are inverted. All energies were compared and subsequently plotted in groups of common ligands (SI figure 3.4A) or common receptor (SI figure 3.4B).

3.5. References

Aiello, D. and Caffrey, D.R., 2012. Evolution of specific protein-protein interaction sites following gene duplication. *Journal of molecular biology*, 423(2), pp.257–72.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Yeh, L.-S. L., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research, 32,* D115–119.

Arkhipov, A. et al., 2014. Membrane interaction of bound ligands contributes to the negative binding cooperativity of the EGF receptor. *PLoS Computational Biology*, *10*(7), p.e1003742.

Berggård, T., Linse, S., and James, P., 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, *7*(16), pp.2833–42.

Berglund, N. A. et al., 2015. The role of protein-protein interactions in Toll-like receptor function. *Progress in Biophysics and Molecular Biology*, *119*(1), pp. 72–83.

Berman, H. et al., 2000. The Protein Data Bank. *Nucleic Acids Research*, *28*(1), pp.235–42.

Boersma, M. D. et al., 2008. Hydrophile scanning as a complement to alanine scanning for exploring and manipulating protein-protein recognition: application to the Bim BH3 domain. *Protein Sci.*, *17*, pp.1232–1240.

Bondi, A., 1964. van der Waals Volumes and Radii. *The Journal of Physical Chemistry*, *68*, pp.441–451.

Bougouffa, S., and Warwicker, J., 2008. Volume-based solvation models out-perform area-based models in combined studies of wild-type and mutated protein-protein interfaces. *BMC Bioinformatics*, *9*, pp.448.

Campbell, S. T. et al., 2015. Mapping the BH3 Binding Interface of Bcl-xL, Bcl-2, and Mcl-1 Using Split-Luciferase Reassembly. *Biochemistry*, *54*, pp.2632–2643.

Case, D. A. et al., 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, *26*, pp.1668–1688.

Casino, P., Rubio, V., and Marina, A., 2009. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell*, *139*(2), pp.325–36.

Chen, L. et al., 2005. Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Molecular Cell*, *17*, pp.393–403.

Cheng, E. H. et al., 2001. BCL-2, BCL-X L sequester BH3 domain-only molecules preventing BAX- and BAK-mediated mitochondrial apoptosis. *Molecular Cell*, *8*, pp.705–711.

Ciccotti, G., and Ryckaert, J. P., 1986. Molecular dynamics simulation of rigid molecules. *Computer Physics Reports*, *4*, pp.346–392.

Cole, C., and Warwicker, J. I. M., 2002. side chain conformational entropy at protein – protein interfaces. *Protein Science*, *11*, pp.2860–2870.

Cory, S., Huang, D. C. S., and Adams, J. M., 2003. The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene*, *22*, pp.8590–8607.

Czabotar, P. E. et al., 2011. Mutation to bax beyond the BH3 domain disrupts interactions with pro-survival proteins and promotes apoptosis. *Journal of Biological Chemistry*, *286*(9), pp.7123–7131.

Czabotar, P. E. et al., 2007. Structural insights into the degradation of Mcl-1 induced by BH3 domains. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), pp.6217–6222.

Czabotar, P. E. et al., 2014. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. *Nature Reviews. Molecular Cell Biology*, *15*(1), pp.49–63.

Darden, T., York, D., and Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(1993), pp.10089.

Day, C. L. et al., 2005. Solution Structure of Prosurvival Mcl-1 and Characterization of Its Binding by Proapoptotic BH3-only Ligands. *Journal of Biological Chemistry*, *280*(6), pp.4738–4744.

Day, C. L. et al., 2008. Structure of the BH3 Domains from the p53-Inducible BH3-Only Proteins Noxa and Puma in Complex with Mcl-1. *Journal of Molecular Biology*, *380*(5), pp.958–971.

DeBartolo, J. et al., 2012. Predictive Bcl-2 family binding models rooted in experiment or structure. *Journal of Molecular Biology*, *422*(1), pp.124–144.

Fletcher, J. I. et al., 2008. Apoptosis is triggered when prosurvival Bcl-2 proteins cannot restrain Bax. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, pp.18081–18087.

Friedman, R., and Hughes, A. L., 2001. Gene duplication and the structure of eukaryotic genomes. Genome Research, 11, pp.373–381.

Fuchs, J. E. et al., 2013. Substrate-Driven Mapping of the Degradome by Comparison of Sequence Logos. PLoS Computational Biology, 9, p.e1003353.

Gromiha, M. M. et al., 2011. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome Science, 9,* p.S13

Gromiha, M. M., Yokota, K., and Fukui, K., 2009. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Molecular Biosystems*, *5*, pp.1779–1786.

Guo, Z. et al., 2015. Target the more druggable protein states in a highly dynamic protein-protein interaction system. *Journal of Chemical Information and Modeling*. 56 (1), pp 35–45

Hansen, N., and van Gunsteren, W. F., 2014. Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, *10*(7), pp.2632–2647.

Happo, L., Strasser, A., and Cory, S., 2012. BH3-only proteins in apoptosis at a glance. *Journal of Cell Science*, *125*, pp.1081–1087.

Herman, M. D. et al., 2008. Completing the family portrait of the anti-apoptotic Bcl-2 proteins: Crystal structure of human Bfl-1 in complex with Bim. *FEBS Letters*, *582*, pp.3590–3594.

Homeyer, N., and Gohlke, H., 2012. Free Energy Calculations by the Molecular Mechanics

Poisson−Boltzmann Surface Area Method. *Molecular Informatics*, *31*, pp.114–122.

Horovitz, A. et al., 1990. Strength and co-operativity of contributions to surface salt bridges to protein stability. *Journal of Molecular Biology*, *216*, pp.1031–1044.

Huber, R. G. et al., 2013. Entropy from state probabilities: hydration entropy of cations. *The Journal of Physical Chemistry. B*, *117*(1), pp.6466–72.

Ivanov, S., Dimitrov, I., and Doytchinova, I., 2013. Quantitative Prediction of Peptide Binding to HLA-DP1 Protein, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(3), pp.811–815.

Jenson, J.M. et al., 2017. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1. *eLife*, 6, pp.1–23.

Jorgensen, W. L. et al., 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*, pp.926.

Kar, G. et al., 2012. Human proteome-scale structural modeling of E2-E3 interactions exploiting interface motifs. *Journal of Proteome Research*, *11*(2), pp.1196–207.

Kim J. S. et al., 2015. Conversion of cell-survival activity of Akt into apoptotic death of cancer cells by two mutations on the BIM BH3 domain. *Cell Death Dis*.,p.e1804.

Koehl, P., and Delarue, M., 1994. Application of a Self-consistent Mean Field Theory to Predict Protein side chains Conformation and Estimate Their Conformational Entropy. *Journal of Molecular Biology*, *239*(2), pp.249–275.

Ku, B. et al., 2011. Evidence that inhibition of BAX activation by BCL-2 involves its tight and preferential interaction with the BH3 domain of BAX. *Cell Research*, *21*(4), pp.627–41.

Kumari, R., Kumar, R., and Lynn, A., 2014. g-mmpbsa - a GROMACS tool for high-throughput MM-PBSA calculations. *Journal of Chemical Information and Modeling*, *54*, 1951–1962.

Kvansakul, M. et al., 2008. Vaccinia virus anti-apoptotic F1L is a novel Bcl-2-like domain-swapped dimer that binds a highly selective subset of BH3-containing death ligands. *Cell Death and Differentiation*, *15*, pp.1564–1571.

Lee, E. F. et al., 2008. A novel BH3 ligand that selectively targets Mcl-1 reveals that apoptosis can proceed without Mcl-1 degradation. *Journal of Cell Biology*, *180*(2), pp.341–355.

Lee, E. F. et al., 2009. Conformational changes in Bcl-2 pro-survival proteins determine their capacity to bind ligands. *Journal of Biological Chemistry*, *284*(44), pp.30508–30517.

Lee, E. F. et al., 2014. The functional differences between pro-survival and pro-apoptotic B cell lymphoma 2 (Bcl-2) proteins depend on structural differences in their Bcl-2 homology 3 (BH3) domains. *The Journal of Biological Chemistry*, *289*(52), pp.3601–17.

Ma, B. et al., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(10),pp. 5772–7.

Maier, J. A. et al., 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone

Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11 (8), pp.3696–3713.

Manning, G. et al., 2002. The Protein Kinase Complement of the Human Genome. *Science (New York, N.Y.)*, *298*(5600), pp.1912–34.

Miller, B. R. et al., 2012. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.*, *8*(9), pp.3314–3321.

Modi, V., Lama, D., and Sankararamakrishnan, R., 2012. Relationship between helix stability and binding affinities: molecular dynamics simulations of Bfl-1/A1- binding pro-apoptotic BH3 peptide helices in explicit solvent. *Journal of Biomolecular Structure and Dynamics*, *31*(1), pp.65-77.

Moldoveanu, T. et al., 2014. Many players in BCL-2 family affairs. *Trends in Biochemical Sciences*, *39*(3), pp.101–111.

Nair, S. K., and Burley, S. K., 2003. X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, *112*(2), pp.193–205.

Nguyen, M. N., Tan, K. P., and Madhusudhan, M. S., 2011. CLICK - topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Research*, *39*(May), p.W24–W28.

Onufriev, A., Bashford, D., and Case, D. A., 2004. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Structure, Function and Genetics*, *55*, pp.383–394.

Patronov, A. et al., 2012. Peptide binding to HLA-DP proteins at pH 5.0 and pH 7.0: a quantitative molecular docking study. *BMC Structural Biology*, *12*(1), pp.20.

Petros, A. M., Olejniczak, E. T., and Fesik, S. W., 2004. Structural biology of the Bcl-2 family of proteins. *Biochimica et Biophysica Acta*, *1644*, pp.83–94.

Rajan S. et al., 2015. BH3 induced conformational changes in Bcl-Xl revealed by crystal structure and comparative analysis. *Proteins*. 83(7), pp.1262-1272.

Robin A. Y. et al., 2015. Crystal structure of Bax bound to the BH3 peptide of Bim identifies important contacts for interaction. *Cell Death Dis*. 6(7), p.e1809.

Roe, D. R., and Cheatham, T. E., 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, *9*(7), pp.3084–3095.

Sedlak, T. W. et al., 1995. Multiple Bcl-2 family members demonstrate selective dimerizations with Bax. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(17), pp.7834–8.

Shamas-Din, A. et al., 2011. BH3-only proteins: Orchestrators of apoptosis. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1813*(4), pp.508–520.

Shoemaker, B. A., and Panchenko, A. R., 2007. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology*, *3*(3), p.e42.

Smits, C. et al., 2008. Structural Plasticity Underpins Promiscuous Binding of the Prosurvival Protein

A1. *Structure*, *16*, pp.818–829.

Stewart, M. L. et al., 2010. The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer. *Nature Chemical Biology*, *6*(8), pp.595–601.

Suzuki, M., Youle, R. J., and Tjandra, N., 2000. Structure of Bax: coregulation of dimer formation and intracellular localization. *Cell*, *103*(4), pp.645–54.

Tan, C., Tan, Y. H., and Luo, R., 2007. Implicit nonpolar solvent models. *Journal of Physical Chemistry B*, *111*(2), pp.12263–12274.

van Wijk, S. J. et al., 2009. A comprehensive framework of E2-RING E3 interactions of the human ubiquitin-proteasome system. *Molecular Systems Biology*, *5*(295), pp.295.

Warwicker, J., and Watson, H. C., 1982. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *Journal of Molecular Biology*, *157*, pp.671–679.

Webb, B., and Sali, A., 2014. *Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics* (Vol. Chapter 5, Unit 5.6).

Willis, S. N. et al., 2005. Proapoptotic Bak is sequestered by Mcl-1 and Bcl-xL, but not Bcl-2, until displaced by BH3-only proteins. *Genes and Development*, *19*, pp.1294–1305.

Yin, X. M., Oltvai, Z. N., and Korsmeyer, S. J., 1994. BH1 and BH2 domains of Bcl-2 are required for inhibition of apoptosis and heterodimerization with Bax. *Nature*, *369*, pp.321–323.

Chapter 4 - Energetic fingerprinting of ligands binding to paralogous proteins: the case of the apoptotic pathway

PAPER 3

Stefan M. Ivanov[1,2], Roland G. Huber[2], Irfan Alibay[3], Jim Warwicker[1], Peter J. Bond[2,4*]

[1] Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

[2] Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Matrix 07-01, 30 Biopolis Street, Singapore 138671, Singapore

[3] Division of Pharmacy and Optometry, School of Health Sciences, The University of Manchester, Oxford Road, Manchester, M13 9PT

[4] Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore

*corresponding author: **peterjb@bii.a-star.edu.sg**

SMI compiled the datasets; SMI and RGH performed the calculations; all authors analyzed the data and wrote the manuscript.

Manuscript in preparation

Abstract

Networks of biological molecules are key to Life, transmitting signals and passing small molecule metabolites through pathways that generate the energy for cellular existence. Genetic duplication processes give rise to sets of regulatory proteins that have evolved from a common architecture. A better understanding of the determinants of specificity at interfaces, common in between functionally related proteins, is crucial to computer-aided drug design and delivering new and improved pharmacotherapeutic agents.

To that end, a comprehensive dataset on drug and drug-like binders of the Bcl-xL and Bcl-2 antiapoptotic proteins was assembled and used to derive a two-dimensional quantitative structure-activity relationship (2D QSAR) model, predicting ligand specificity for the two homologous proteins, as might be seen in a typical drug design campaign. The strengths and weaknesses of high-throughput 2D QSAR are then compared and contrasted to those of high-level theory, thermodynamic integration calculations, totaling 1.65 μs of simulation, performed on 14 complexes of Bcl-xL-specific, Bcl-2-specific, and potent dual binders, bound to the Bcl-xL and Bcl-2 proteins. It is shown that free energy calculations provide a layer of essential information, which traditional QSAR can not capture, and that proteins energetically distinguish between specific and unspecific binders. Moreover, it is shown that protein energetic responses to different ligands, expressed as per-residue energy values, can be used to fingerprint the protein – ligand interaction, extending the framework of four-dimensional molecular dynamics/quantitative structure-activity relationships (4D-MD/QSAR). Finally, directions for future work in 4D-MD/QSAR, further extending the present framework, are laid out, with the aim of facilitating future drug design campaigns.

4.1. Introduction

Association between biopolymers is an essential part of physiological processes in all domains of Life. Deviations from the genetically predetermined innate interactome are often deleterious, manifesting themselves in a broad spectrum of conditions, ranging from minor symptoms to debilitating syndromes (Rodriguez-Soca et al., 2010; Gonzalez and Kann, 2012). For example, overexpression of the antiapoptotic proteins Bcl-xL and Bcl-2 overwhelms the cell's proapototic defences, facilitating malignant proliferation (Park et al. 2013). Conversely, mutations that abrogate binding between pro- and antiapoptotic proteins shift the cellular balance toward premature cell death, and give rise to or are associated with degenerative diseases (Bouillet et al., 2001; Akhtar et al., 2004). Point mutations in leucine zipper transcription factors can lead to altered dimerization and DNA binding, resulting in a great number of documented malignancies (Rodriguez-Martinez et al., 2017). Emulating regulatory protein – protein interactions with small molecule or (stapled) peptide analogs – the so-called Bcl-inhibitors – has been successfully used to rectify imbalances in the apoptosis-regulation pathway, and has produced numerous promising clinical candidates (Lessene et al., 2008). Encouragingly, one Bcl-2-inhibitor – Venetoclax (ABT-199) – has already obtained FDA-approval for cancer treatment (Souers et al., 2013). Targeting other regulatory systems governed by inter- or intrafamily protein – protein recognition and association, such as the ubiquitination pathway, two-component signal transduction, and G protein receptor regulation is also expected to provide future drug candidates for a wide range of conditions (Cohen and Tcherpakov, 2010; Gotoh et al., 2010; Siryk-Bathgate et al. 2013). Thus, unraveling the intricacies of protein – protein and protein – ligand association - the origins of specificity and affinity - is a long-standing goal in the broader scientific field, not only from a theoretical, but also a practical, standpoint.

As of yet, there is no single overarching model of protein – protein and protein – ligand association, or intermolecular association in general, that can accommodate permanent and transient, strong and weak, specific and unspecific protein – protein/ligand interactions in a single theoretical framework, account for all aspects known to influence binding, and reliably predict the outcome of protein mutations, chemical alterations in ligands, and changes in solvent composition. Rather, weak, unspecific interactions tend to be the subject of colloid science (Curtis and Lue, 2006; Velev et al. 1998), whereas specific interactions tend to be examined at an atomistic-level of detail through the lens of bioinformatics (Gromiha et al., 2009; Gromiha et al., 2011; Aiello and Caffrey, 2012) or force field-based approaches (Ivanov et al.. 2016; Cuthbertson et al., 2006). Where structures or homology models of the drug targets of interest are not available, quantitative structure-activity relationship (QSAR) studies are often employed to rationalize observed patterns in ligand binding – the chemical moieties that confer affinity and specificity are identified and characterized (Hopfinger et al., 1997; Vicini et al., 2002; Winkler, 2002; Duchowicz et al., 2006; Patronov et al. 2012; Ivanov et al., 2013; Patronov and Doytchinova, 2013; Varnes et al., 2014; Yousefinejad and Hemmateenejad, 2015).

Intrafamily protein – protein interactions in key regulatory pathways, such as the Bcl-2 proteins, which regulate apoptosis (Moldoveanu et al., 2014), present a formidable challenge to drug design and cancer therapy. The high similarity between family members, which have arisen through gene duplication and subsequent divergence, makes off-target effects almost unavoidable and limits the therapeutic efficacy of pharmacological agents. For example, the Bcl-2 protein has been targeted to treat non-Hodgkin lymphoma and chronic lymphocytic leukaemia (Ng and Davids, 2014), but undesired binding to the highly similar Bcl-xL protein causes dose-limiting thrombocytopenia (Gandhi et al., 2011). To date, all known peptide or small molecule Bcl-2 binders also possess at least some residual affinity for Bcl-xL (Okamoto et al., 2012; Wendt et al. 2006; Bruncko et al., 2007; Porter et al., 2009; Tao et al., 2014). Thus, decoupling Bcl-2 binding from Bcl-xL binding has proven impossible in drug design campaigns. One solution is to design selective ligands for the desired target. Current experience with Bcl-2 inhibitors in humans demonstrates that a selectivity of several thousand-fold for the designated target is sufficient to make off-target binding clinically insignificant (Souers et al., 2013; Rudin et al., 2012; Sleebs et al. 2011). The lower bound for selectivity, however, has not been definitively established and may vary in different signaling pathways.

Medicinal chemists typically attempt to enhance binding affinity or achieve selectivity by optimizing polar contacts and hydrogen bonds or salt bridges between a ligand and receptor under the assumption that changes in ligands are small enough to not induce significant conformational change in the receptor. When this is the case, it is straightforward to visualize and comprehend the origins of any enhancements to affinity afforded by introducing hydrogen bond donors and acceptors or by enhanced van der Waals contacts. Even under these very favorable circumstances, however, it is not easy to anticipate every resulting change in binding energetics, e.g. the free energy cost of reorganizing the receptor from the unbound to the bound ensemble or the change in ligand conformational entropy upon binding (Rocklin et al., 2013a). Moreover, ligand – receptor interfaces often involve a great number of interdependent interactions. Optimizing affinity and/or specificity becomes a complex, multidimensional problem. Clearly, a way to reduce the dimensionality of this problem and better understand the behavior of different protein – ligand complexes is needed. To that end, a comprehensive dataset on binding affinity for Bcl-xL and Bcl-2 inhibitors was compiled from literature (Lee et al., 2009; Souers et al., 2013; Sleebs et al., 2011; Wendt et al., 2006; Bruncko et al., 2007; Perez et al., 2012; Porter et al., 2009; Zhou et al., 2012; Sleebs et al., 2013; Tao et al., 2014; Leverson et al., 2015; Yusuff et al., 2012; Brady et al., 2014; Hennessy, 2016; Lessene et al., 2013; Varnes et al., 2014; Bruncko et al., 2007; Feng et al., 2010; Aguilar et al., 2013; Park et al., 2008; Touré et al., 2013; Lessene et al., 2008; Vogler et al., 2009; Zhai et al., 2006). A two-dimensional quantitative structure-activity relationship (2D QSAR) analysis is presented and compared to an energetic analysis, obtained from molecular dynamics simulations, on two dual Bcl-2/Bcl-xL binders, two Bcl-xL-specific binders, and two Bcl-2-specific binders from a congeneric series of compounds (Figures 4.1 and 4.2). 2D QSAR results are compared and contrasted with thermodynamic integration calculations totaling 1.65 μs, performed on 12 modeled protein – ligand complexes and 2 template protein – ligand complexes, obtained from the Protein Data Bank (Berman et al., 2003) (Figure 4.1). It is demonstrated how the “computational microscope” of molecular dynamics simulations (Dror et al., 2012), coupled with free energy calculations, can be used to more fully characterize protein – ligand complexes, as, by the very nature of the techniques, it accounts for and provides details on the binding process, which traditional QSAR is incapable of capturing.

4.2. Methods

4.2.1. QSAR

Only compounds shown to bind to the hydrophobic groove of the antiapoptotic proteins were considered for future analysis; these were converted to SMILES format. For each compound, 165 molecular descriptors were calculated with RDKit (Gasteiger and Marsili, 1980; Balaban, 1982; Bertz, 1981; Bonchev and Trinajstić, 1977; Hall and Kier, 1991; Wildman and Crippen, 1999; Ertl et al., 2000; Labute, 2000; Nguyen et al., 2009).

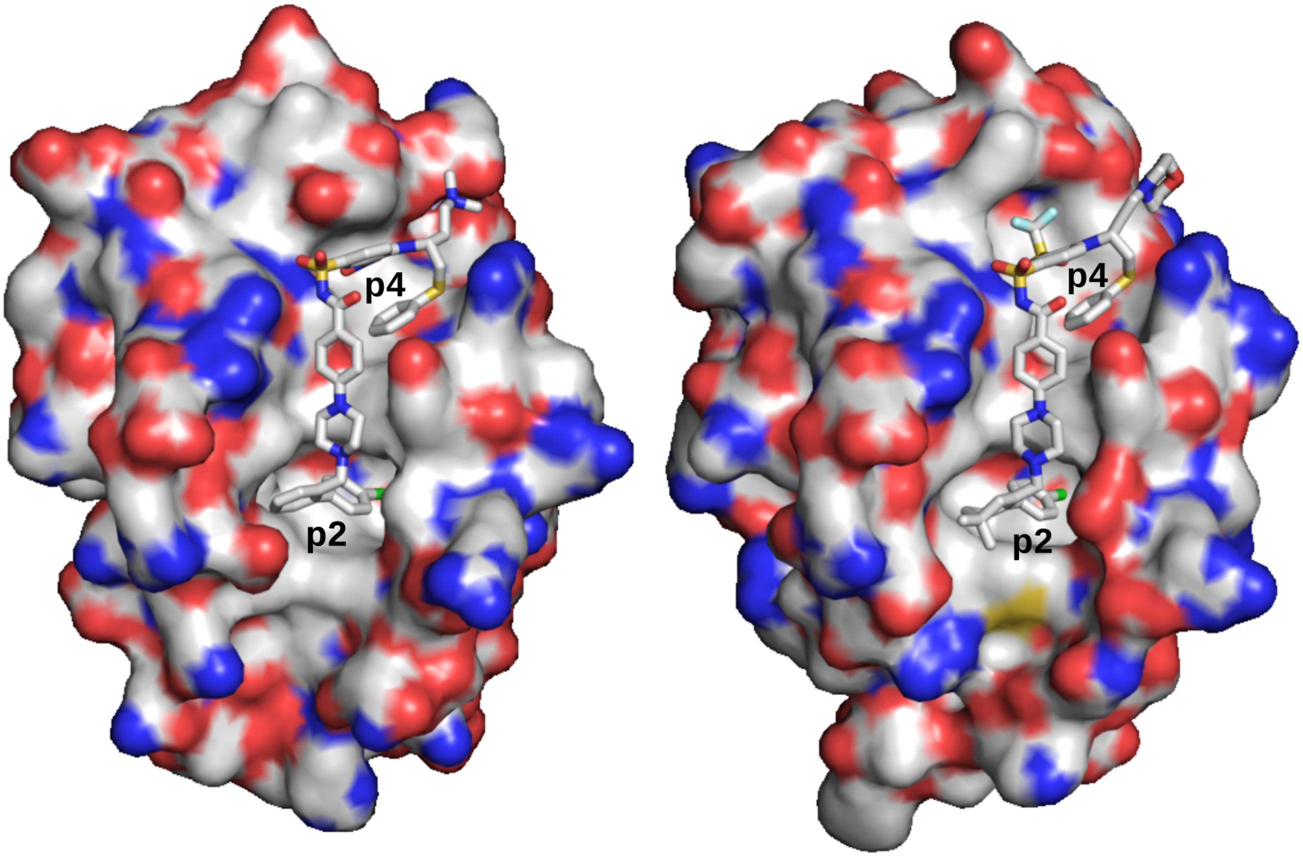
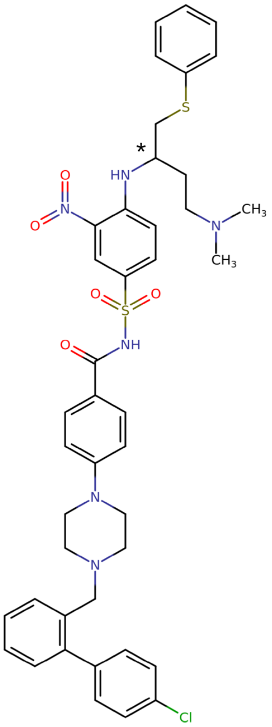


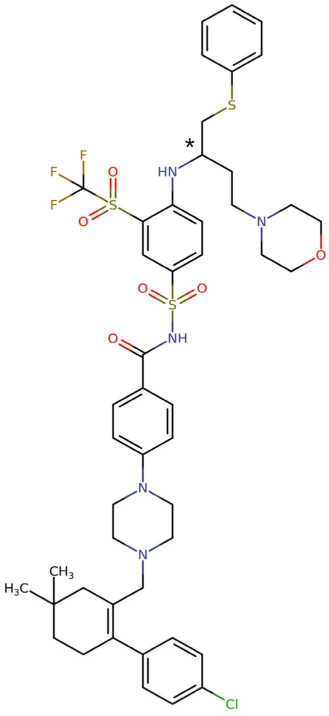
Figure 4.1. Crystal structures of Bcl-xL and Bcl-2 bound to small molecule ligands. Bcl-xL bound to ABT-737 (left, PDB ID 2YXJ (Lee et al., 2007)) and Bcl-2 bound to Navitoclax (ABT-263) (right, PDB ID 4LVT (Souers et a., 2013)). The proteins are in surface representation, the ligands are in stick representation with carbon atoms in white, oxygen in red, nitrogen in blue, sulfur in yellow, chlorine in green, and fluorine in gray. Deep hydrophobic pockets on the surface of the proteins, occupied by ligand moieties, are labeled with p2 and p4.

Moreover, descriptors were additionally normalized by molecular weight and heavy atom count; parameters, scaled to unit standard deviation, with zero or near zero variance, were removed from consideration. Partial least squares regression was then performed on the resulting set of descriptors, fitting the data to the logarithm of the affinity of a compound for Bcl-xL over Bcl-2 ($\log_{10}$(affinity for Bcl-xL/affinity for Bcl-2)) – a parameter henceforth referred to as specificity. Ligands with a specificity below -1 are defined as Bcl-xL selective, ligands with a specificity between -1 and

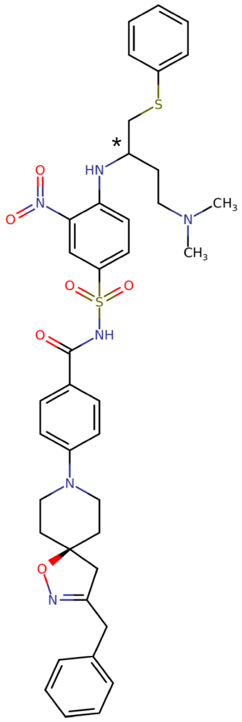
N3C (ABT-737)
0.5/~0.5



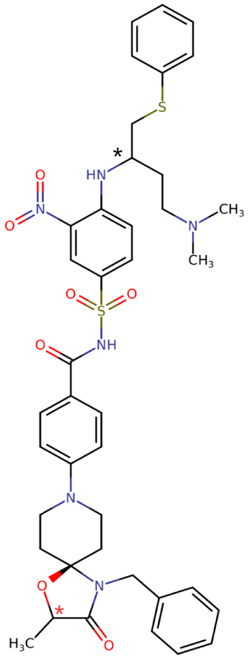
1XJ (ABT-263)
0.055/0.044



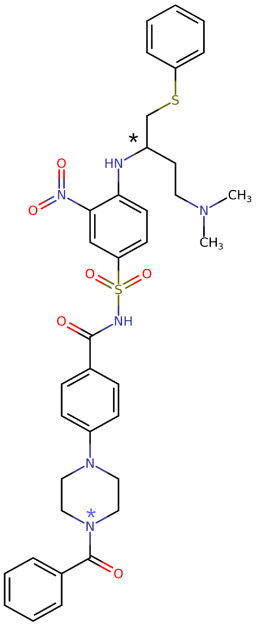
compound 3
1.7/6.5



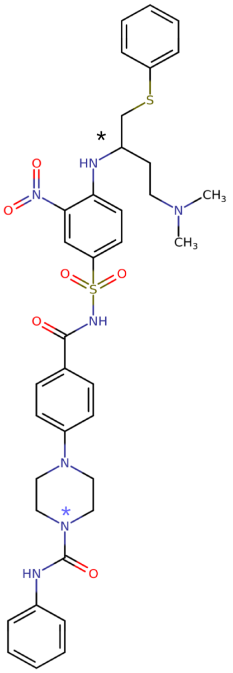
compound 4
3.7/9.6



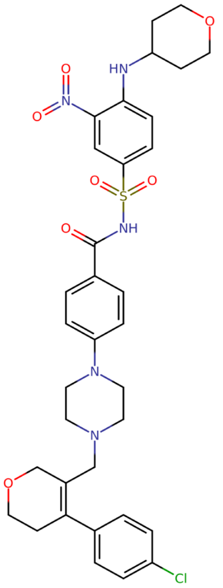
compound 5
< 1/61.9



compound 6
2.5/300



compound 7
5540/59



compound 8
48/< 0.01

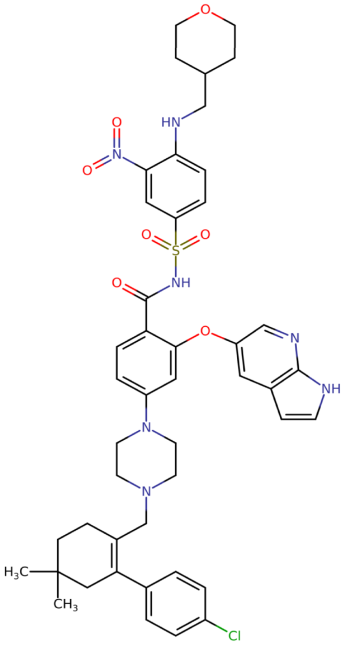


Figure 4.2. Structures of template and modeled compounds. Chemical formulae of the six model compounds (labeled 3 - 8) and the N3C (ABT-737) and 1XJ (Navitoclax, ABT-263) template compounds used in molecular dynamics simulations. Shown above every compound is its K_i value (in nM) for Bcl-xL, followed by the K_i value for Bcl-2 (in nM). Compounds 3, 4, N3C, and 1XJ are potent dual binders, compounds 5 and 6 are Bcl-xL selective, compounds 7 and 8 are Bcl-2 selective. The exact affinities of compounds 5 for Bcl-xL and N3C and 8 for Bcl-2 are above the measurement limit of the experimental procedures employed during measurements. Affinity data on compounds 3 – 6 is from Bruncko et al. (2007); 7, 8, 1XJ, and N3C is from Souers et al. (2013), IC_{50} values for N3C (3/6.1 nM) are also reported in Sleebs et al. (2011). Chiral carbon atoms marked with a black asterisk (*) were modeled in the R-configuration; such were the compounds used in the experimental affinity measurements. The chiral atom in compound 4, marked with a red asterisk (*), was modeled in the S-configuration; its configuration is not specified in the relevant publication. Thus, the S-configuration was chosen, as it shields the $-CH_3$ group from solvent to a greater degree than the R-configuration. The nitrogen atoms in compounds 5 and 6, marked with a blue asterisk (*), were modeled in both the protonated and unprotonated state. Henceforth, results reported for compounds 5 and 6 correspond to the state where these nitrogens are unprotonated (ligand net charge +1), except where explicitly stated otherwise. The canonical groove of the antiapoptotic proteins lies below the piperazinyl and sulfonamidophenyl rings. Compounds 3 – 6 had their terminal phenyl groups oriented in pocket 2 of the antiapoptotic proteins (see Figure 4.1), and N and O atoms pointing outward into the solvent, with the exception of the spiro nitrogen of compound 4, which was also buried in the groove. The configurations of the spiro moieties of compounds 3 and 4 were modeled to be identical to the ones reported in the experimental study (Bruncko et al., 2007).

and 1 - as dual binders, and ligands with a specificity greater than 1 - as Bcl-2 specific. The final dataset consisted of 57 Bcl-xL-selective compounds, 112 dual binders, and 19 Bcl-2-selective binders (Figure 4.3).

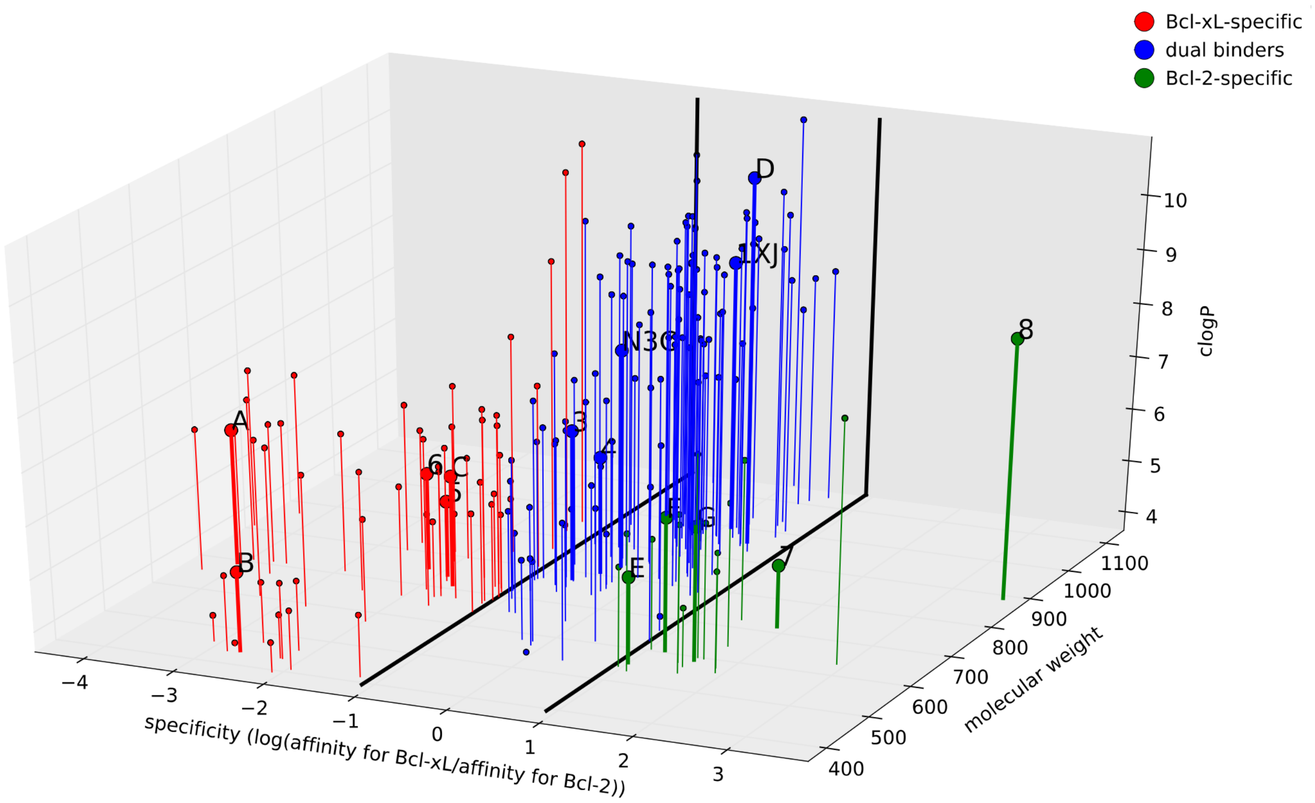


Figure 4.3. Compounds included in QSAR analysis. All compounds discussed in this work plotted by experimentally measured specificity, molecular weight and computed logP (clogP). For clarity, Bcl-xL-specific ligands, dual binders, and Bcl-2-specific compounds are colored differently and the specificity region between -1 and 1 is highlighted. Compounds A - G, N3C, 1XJ, and 3 - 8 are labeled, shown in bold, and discussed further in the text. N3C's specificity was computed from IC_{50} values reported in Sleebs et al. (2011). As the affinities of compounds 5 and 8 for Bcl-xL and Bcl-2 were beyond the measuring sensitivity of the experimental setup (K_i < 1 nM (Bruncko et al., 2007), K_i < 0.01 nM (Souers et al., 2013), respectively), provisional specificities for those compounds were computed using values of 1 and 0.01 nM, respectively (also see Figure 4.2).

The compounds were sorted by increasing specificity and split into a training and test set for external validation (Gramatica, 2007) in a 3:1 ratio – for every three compounds in the training set, one was placed in the test set, moving from Bcl-xL-selective to Bcl-2-selective ligands. The training dataset was used to derive a partial least squares regression model, from which the specificity of the compounds in the test set was predicted. No compound from the training set was present in the test set. To test the robustness and validity of the QSAR model, Monte Carlo cross-validation was performed (Mitchell, 2014; Gu and Lai, 1991), i.e. a random number of random compounds in the test set was swapped with compounds from the training set, after which the QSAR model was derived anew and used to make predictions on the new test set, repeating this procedure 20 000 times. Only swaps within classes were permitted, i.e. Bcl-xL selective compounds were exchanged only for Bcl-xL selective ligands, Bcl-2-selective compounds were exchanged only for Bcl-2-selective ligands, and dual binders were exchanged only for dual binders. Additionally, to remove any bias against the scarce Bcl-2 selective ligands, the dataset was “normalized,”, i.e. an analogous analysis was performed with only 19 compounds in each class – 14 in the training set and 5 in the test set. Partial least squares regression was performed with the R *pls* package (Mevik and Wehrens, 2007).

4.2.2. System setup for template structure equilibration simulations

Six of the compounds (see Figures 4.2 and 4.3) were modeled on the 2YXJ and 4LVT crystal structures (ABT-737 and ABT-263 bound to Bcl-xL and Bcl-2, respectively); the models were subjected to molecular dynamics simulation and free energy calculations (Brandsdal et al., 2003; Lyne et al., 2006; Steinbrecher et al., 2008; Mobley and Klimovich, 2012; Settimo et al., 2014; Homeyer et al., 2014; Christ and Fox, 2014; Hansen and van Gunsteren, 2014; Burusco et al., 2015; Aldeghi et al., 2016). Before simulations of the modeled complexes could be initiated, the template structures were simulated first; all simulations were performed with the Amber14 suite (Case et al., 2005). N-termini were capped with an acetyl group, C-termini were capped with a methylamino (-$NHCH_3$) group. Protein chains were protonated and solvated in a cubic box with TIP3P water (Jorgensen et al., 1983) with *tleap* with a minimal wall distance of 13 Å. 0.15 M NaCl was added to approximate a physiological salt concentration whilst ensuring charge neutrality; protonation states for the ligands were assigned using ChemAxon's Calculator Plugins (Dixon and Jurs, 1993; Csizmadia et al. 1997); parameters for the ABT-737 and ABT-263 compounds were obtained from the general Amber force field (GAFF 1.7) (Wang et al. 2004) with AM1-BCC charges (Jakalian et al., 2000) using *antechamber*.

4.2.3. Simulation protocol for template structure equilibration simulations

The solvated systems were subjected to 2000 steps of energy minimization with a harmonic restraint of 100 kcal/mol*Å^2 on all heavy atoms, followed by 2000 steps of minimization with restraints on all protein and ligand heavy atoms, followed by 2000 steps of minimization with restraints on protein heavy atoms only. The systems were heated from 100 to 300 K over a period of 1 ns at constant volume with 100 kcal/mol*Å^2 harmonic restraints on all heavy atoms, followed by constant pressure density equilibration of 1 ns, followed by cooling to 100 K at constant volume with harmonic restraints. The protein – ligand complexes were again subjected to 2000 steps of energy minimization with restraints of 100 kcal/mol*Å^2 on protein heavy atoms, followed by 2000 step minimization series with decreasing restraints – 50, 20, and 10 kcal/mol*Å^2 . The systems were then reheated from 100 to 300 K under constant volume conditions over a period of 1 ns, with a harmonic restraint of 20 kcal/mol*Å^2 on protein and ligand heavy atoms, followed by 1 ns of constant pressure density equilibration with restraints, followed by 1 ns of equilibration with restraints on protein heavy atoms, followed by 1 ns of equilibration with 20 kcal/mol*Å^2 restraints on Cα atoms only. The systems were then equilibrated for 1 ns without any restraints and simulated for 100 ns under constant pressure (1 bar) and temperature (300 K) conditions, maintained with the Berendsen barostat (Berendsen et al., 1984) and the Langevin thermostat (Adelman and Doll, 1974); collision frequencies were set to 2 ps^{-1} for both pressure and temperature coupling. Both systems were simulated in four independent replicas with the *ff14SB* force field (Maier et al., 2015). An 8.0 Å cutoff was used for van der Waals interactions; long-range electrostatics were computed with the particle-mesh Ewald scheme (Darden et al. 1993). Bonds to hydrogen were constrained using the SHAKE algorithm (Ciccotti and Ryckaert, 1986), allowing for a 2 fs time step.

4.2.4. System setup for thermodynamic integration simulations

Trajectories were processed with *cpptraj V14.25* (Roe and Cheatham, 2013) to perform rototranslational alignment and compute Cα and ligand heavy atom root-mean-square deviations (RMSDs). The instanteneous enthalpy of binding (ΔH) between protein and ligand was also monitored every picosecond; ΔH was computed with MMPBSA.py (Miller et al., 2012). Frames from the Bcl-xL – ABT-737 and Bcl-2 – ABT-263 trajectories were selected to serve as templates for modeling the 12 antiapoptotic protein – ligand complexes and the subsequent free energy calculations (Wang et al., 2015). The six compounds from Figure 4.2 were modeled bound to Bcl-xL and Bcl-2 using Schrödinger's Maestro suite (Anon, 2017). The guiding principle in modeling was to bury hydrophobic moieties in hydrophobic pockets; polar fragments were anticipated to be solvent exposed. The placement of the azaindoleoxy moiety of compound 8 was guided by the positioning of an analogous indoloxy moiety in the 4MAN crystal structure of Bcl-2 bound to a similar compound (Souers et al., 2013), i.e. nearly parallel to the nitroaryl fragment. Protonation states for the modeled ligands were assigned using ChemAxon's Calculator Plugins; parameters were obtained from the general Amber force field (GAFF 1.7) with AM1-BCC charges using *antechamber*. The complexes were solvated in a cubic box with TIP4P water (Jorgensen et al., 1983) and a minimal wall distance of 13 Å; NaCl was added to a concentration of 0.15 M.

4.2.5. Simulation protocol for thermodynamic integration simulations

One-step thermodynamic integration with softcore potentials and linear scaling (Hornak and Simmerling, 2004; Steinbrecher et al., 2011) was performed on the ABT-737 and ABT-263 compounds, bound to Bcl-xL and Bcl-2, transforming each of them into the six compounds in Figure 4.2 bound to the two proteins, as well as free in solution. Henceforth, this is termed the “forward” transformation, as opposed to the reverse process of transforming the model ligand into the template molecule, which was also performed. Moreover, the absolute free energies of binding of the template ligands (Mobley et al., 2007) were computed by decoupling them from the computational box (Klimovich et al., 2015) (for the reverse transformation) and by inserting them into the computational box (for the forward transformation).

The solvated model complexes were subjected to 2000 steps of energy minimization with a harmonic restraint of 20 kcal/mol*Å^2 on protein and ligand heavy atoms. The structures were heated from 100 to 300 K under constant volume conditions for 10 ps with 20 kcal/mol*Å^2 restraints on protein and ligand heavy atoms, followed by 10 ps of restrained density equilibration, and 2.5 ns of unrestrained constant pressure dynamics in each λ-window. The time step was set to 1 fs; bonds to hydrogen were not constrained during thermodynamic integration simulations. Transformations were carried out with 11 λ-windows for the forward (λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) and reverse (λ = 0.005, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.995) processes and default settings for the scalpha (0.5) and scbeta (12 Å^2) parameters, which control the softness of the potential. Energy minimization, heating, density equilibration, and production dynamics were all performed with potential energy functions, corresponding to the λ-value of every λ-window, thus avoiding any Hamiltonian lag (Kollman, 1993). Forward transformations were carried out with *pmemd*, whereas reverse transformations were performed with *sander*. All transformations were carried out with the *ff14SB* forcefield in three independent replicas under NPT conditions.

4.2.6. Free energy analysis

Autocorrelation times (τ) for the free energy calculations were computed from the derivative of the potential with respect to λ (δV/δλ). The pressure-volume component of the free energy was considered marginal and ignored, as is typically the case in biomolecular simulations (Steinbrecher et al., 2006; Gilson and Zhou, 2007; Steinbrecher et al., 2008; Kaus et al., 2013). Moreover, for the *sander* transformations, per-residue energy decomposition was enabled, adding 1–4 energy terms to internal energy terms. Energy values for the protein residues from the last 1 ns of simulation in each λ-window were saved, whereas water and ions were disregarded in subsequent analysis. Per-residue energy values were compared for the model and template ligands bound to Bcl-xL and Bcl-2 for the three replicas by computing Pearson correlation coefficients. When comparing ligands bound to the same protein, Pearson correlations from all protein residues were computed. When comparing ligands across proteins, values for residues 25 – 137 for Bcl-xL and 27 – 139 for Bcl-2 were used. Helix 1, which is not involved in binding, was discarded, i.e. comparisons were only made between the core fold residues among the two proteins, which have a 1:1 correspondence in sequence and structure. Correlation plots were generated with the *R corrplot* package (Murdoch and Chow, 1996; Friendly and Friendly 2002). Tanimoto similarities between the eight ligands from Figure 4.2, topological fingerprint similarities, and MACCS key similarities were computed with RDKit (Durant et al., 2002) and compared to per-residue energy patterns, obtained from free energy analysis.

4.3. Results

4.3.1. 2D QSAR

The initial QSAR model, derived from the entire training set, performed reasonably well in the external validation on the 47 compound test set, with an overall R^2 between predicted and measured specificity of 0.48 (Figure 4.4). Most dual binders were predicted accurately, as were most Bcl-xL selective compounds. Compounds A, B, C, and D, which belong to different chemical series, lie close to the x = y identity line, as do compounds 3, 4, N3C, and 1XJ, which belong to the arylsulphoneamidoaryl series, which dominates the set. Arylsulphoneamidoaryl compounds 5 and 6 have their specificity slightly shifted up towards the dual binders, likely because arylsulphoneamidoaryl dual binders dominate the set. Compounds E, F, and G – a set of tetrahydroisoquinoline derivatives – are well predicted, as similar compounds are present in the training set. Only the Bcl-2 selective arylsulphoneamidoaryls are mispredicted by more than 3 units of specificity, eroding overall R^2 and reducing the slope of the line of best fit for the predictions. This is because the entire dataset contains only two Bcl-2-selective arylsulphoneamidoaryls, none of which are in the training set. Randomizing the training and test sets 20 000 times produced similar results, with R^2 values varying around 0.5. Large decreases in predictive performance were mostly associated with ligand randomizations where training was performed on compounds from one chemical series and used to perform predictions on a very different series of compounds. Performing the analysis with reduced training and test sets improves the prediction accuracy for compound A, which belongs to a fairly sparsely represented chemical series. Again, compounds E, F, and G are accurately predicted, as compounds from this series are present in the training set. There is a slight downshift in predicted specificities for arylsulphoneamidoaryls, as evidenced by compounds 3-8, 1XJ, and N3C, likely because dual arylsulphoneamidoaryl binders do not dominate the “normalized” data set. Moreover, fluctuations in R^2 become more pronounced, as one or two severely mispredicted compounds influence R^2 much more dramatically with the reduced test set. Again, only large, Bcl-2-specific arylsulphoneamidoaryls, which were maintained in the test set throughout all randomizations, were severely mispredicted.

4.3.2. Template structure equilibration simulations

100-ns molecular dynamics (MD) equilibration simulations were performed in quadruplicate on the template compounds bound to Bcl-xL and Bcl-2, in order to relieve any strain potentially present in the complexes and to asses the stability of the structures (Figure 4.5). The proteins were stable over the course of the 100 ns trajectories with Cα RMSDs fluctuating around a steady value below 1.5 Å.

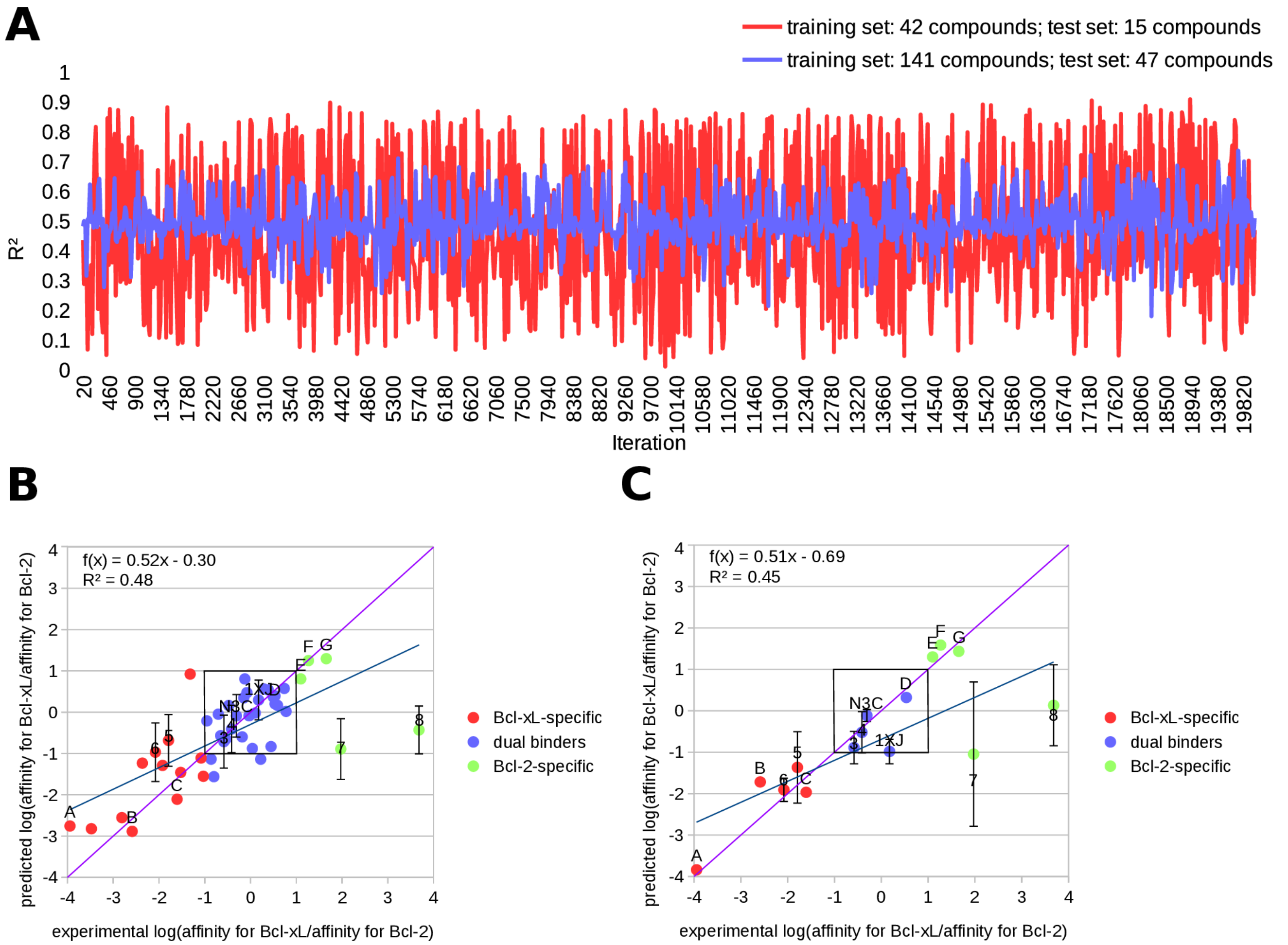


Figure 4.4. QSAR results. (A) R^2 between predicted and experimental specificity of the compounds in the external validation test sets plotted against the iteration number for the 20 000 iterations. N3C's specificity was computed from IC_{50} values reported in Sleebs et al. (2011), specificities for compounds 5 and 8 were computed using values of 1 (Bruncko et al., 2007) and 0.01 (Souers et al., 2013) nM, respectively (see Figure 4.3). **(B)** Predicted specificity plotted against experimental values for the 47 compound test set. Values for compounds without marked standard deviations are from the first QSAR

model, values for compounds 3 – 8, N3C, and 1XJ are averages over the 20 000 iterations; these do not differ significantly from values from the first model; also shown is the standard deviation of predicted specificities for these compounds over the 20 000 iterations. Ligands 3 – 8, N3C, and 1XJ were maintained in the test set throughout all iterations. The x = y line is shown in purple, also shown is the line of best fit from the initial model; the corresponding line equation and R^2 are in the top left corner. The region between -1 and 1 predicted *vs* experimental specificity is highlighted with a square box. **(C)** The same plot as in **(B)** prepared for the 15 compound test set.

The ABT-263 ligand remained highly stable throughout all four replicas, whereas ABT-737 appears to have a greater mobility in the hydrophobic groove of the Bcl-xL protein. Proximity to the crystal structure conformations of the ligands was prioritized, followed by low Cα RMSD values for the proteins. MM-PBSA energies were also monitored to further asses the energetics of binding; these indicated stable binding throughout the trajectories. Suitable snapshots from the simulations (see Figure 4.5) were chosen to serve as templates for modeling the different protein – ligand complexes. These models were subjected to subsequent thermodynamic integration simulations.

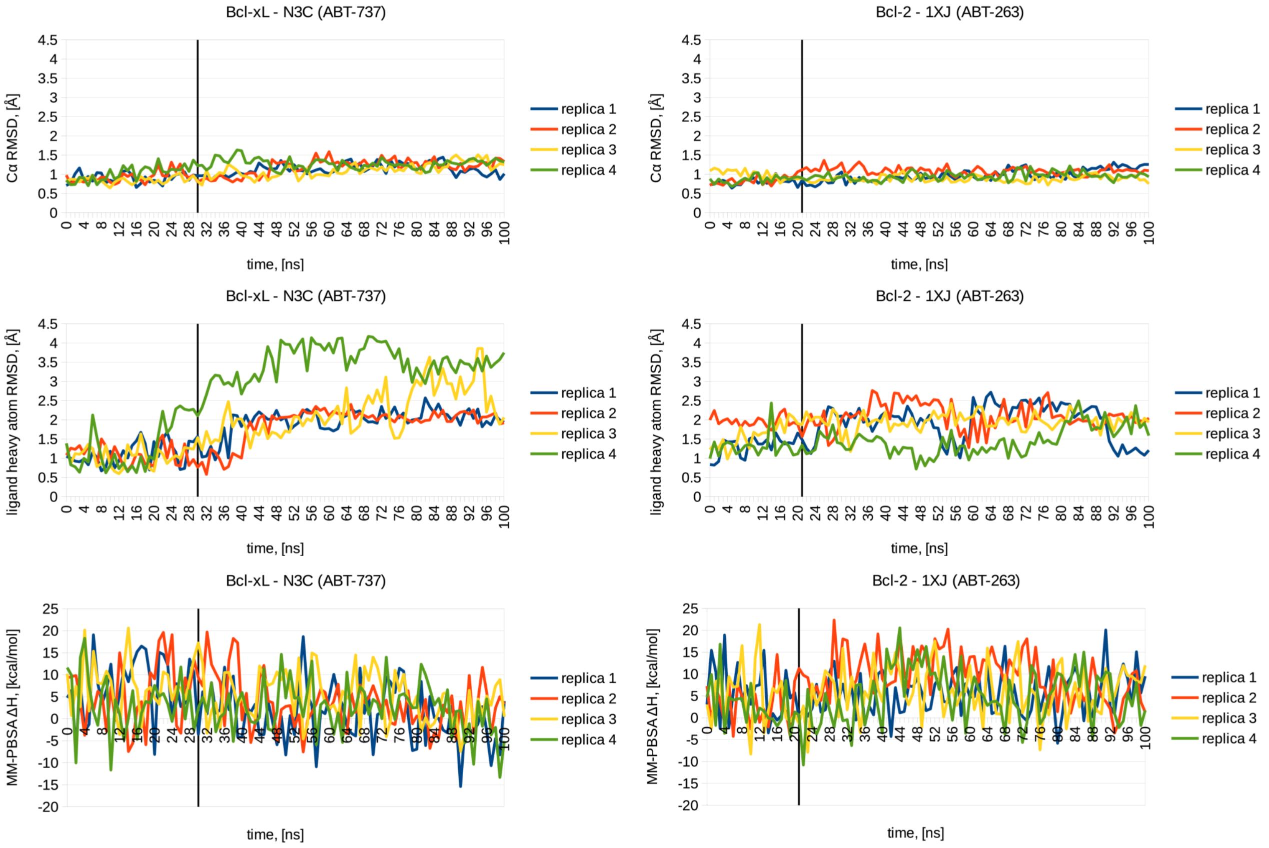


Figure 4.5. Results from template equilibration simulations. Cα RMSDs (top), ligand heavy atom RMSDs (middle), and computed MM-PBSA enthalpies of interaction (bottom) for the four 100-ns replicas of the Bcl-xL – ABT-737 (left) and Bcl-2 – ABT-263 (right) template structures. The vertical black lines indicate the time points chosen to be templates for the free energy calculations. Snapshots at t = 30 ns and t = 21 ns from replicas 2 and 1 were chosen for the Bcl-xL – ABT-737 and Bcl-2 – ABT-263 complexes, respectively.

4.3.3. Thermodynamic integration simulations

Thermodynamic integration calculations (Steinbrecher et al., 2011; Chodera et al. 2011) were performed on the ABT-737 and ABT-263 compounds, bound to Bcl-xL and Bcl-2, transforming each of them into the six compounds in Figure 4.2 bound to the two proteins, in triplicate. The reverse calculations were also performed, again in triplicate. Thus, relative binding affinities (ΔΔGs) were calculated. Moreover, the template ligands were also decoupled/inserted from/into the binding groove and free in solution to compute their absolute binding affinities (ΔGs). Autocorrelation times (τ) varied from several picoseconds to several hundred picoseconds among the different λ-windows. Thus, performing sufficiently long simulations (> 50τ) to obtain a sufficient number of uncorrelated samples in all λ-windows was unfeasible. Therefore, (Δ)ΔGs and convergence plots from δV/δλ values taken every picosecond are presented; these differ marginally from results sampled 10 ps apart. The means of the free energies for the transformations and their standard deviations are shown in Figures 4.6 and 4.7; the corresponding convergence and Cα RMSD plots are presented in Supplementary information figures 4.1 – 4.4 (see Appendix). The proteins were stable throughout the 2.5 ns of dynamics in each λ-window, with RMSDs typically remaining below 1 Å. For most of the ligands, the results from the forward and backward transformations are in good agreement. In most replicas, for the relative transformations, the ΔΔG is converging to a value near the experimental ΔΔG over the course of the simulations, in certain cases within “chemical accuracy,” i.e. within 1 kcal/mol. Only for the transformations where the ligands are inserted into or decoupled from the computational box are computed ΔG values off from experimental ones by a great margin, in certain cases by more than 10 kcal/mol. Moreover, with the exception of ligand 8 in isolation, these are the only simulations with large differences in computed ΔGs among replicas (standard deviations > 5 kcal/mol). The standard deviations of computed (Δ)ΔG values among replicas are severalfold larger than the (Δ)ΔG errors within replicas. The standard deviations, therefore, are the values being presented (Figures 4.6 and 4.7) – most are around 1 kcal/mol.

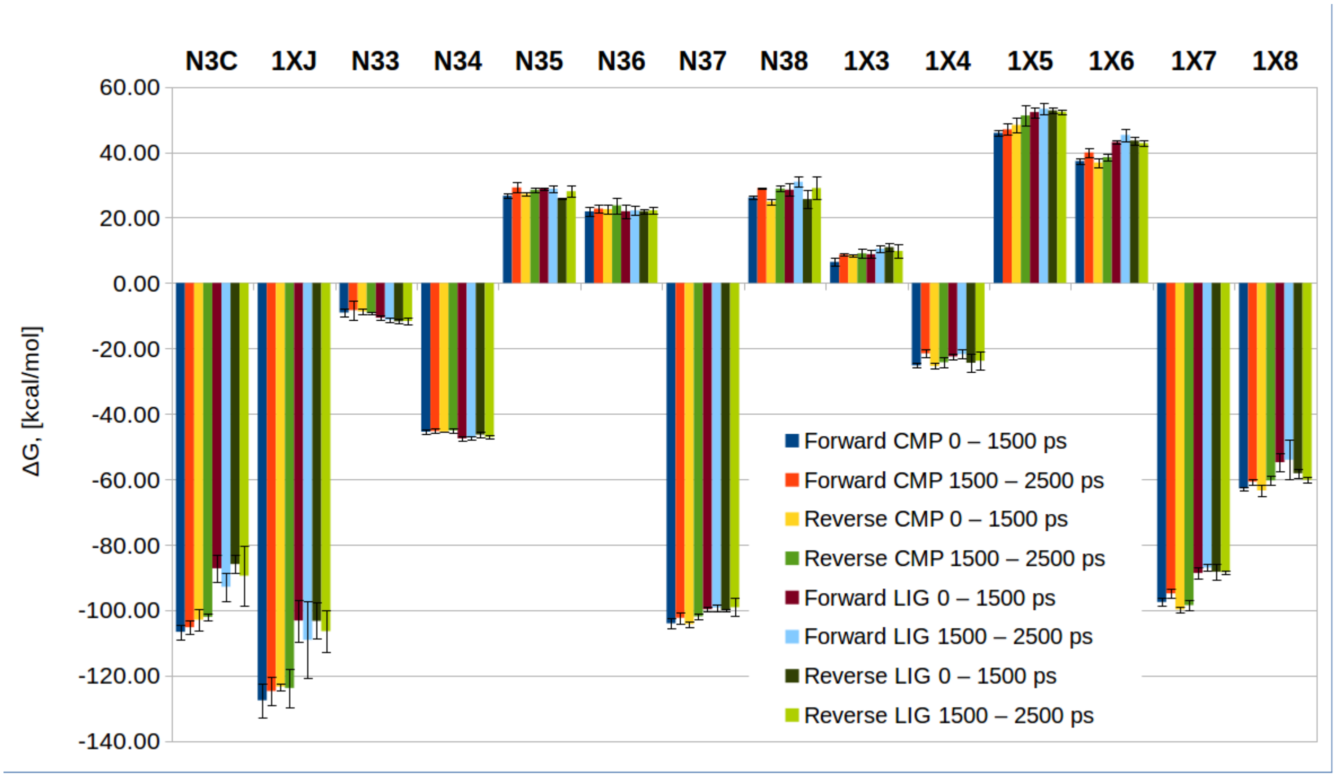


Figure 4.6. Calculated ΔGs. Computed ΔGs from the 11 λ-window alchemical transformations, averaged over the three replicas; standard deviations are also shown. Ligand transformations in complex with Bcl-xL or Bcl-2 are labeled "CMP," transformations where the ligand is free in solution are labeled "LIG." ΔGs were evaluated from 0 – 1500 ps and from 1500 to 2500 ps of production dynamics. For ease of comparison, *pmemd* results (labeled "Forward") are multiplied by -1. Compounds 3 – 8 are prefixed with "N3" and "1X" to designate transformations from/to the N3C and 1XJ templates, respectively.

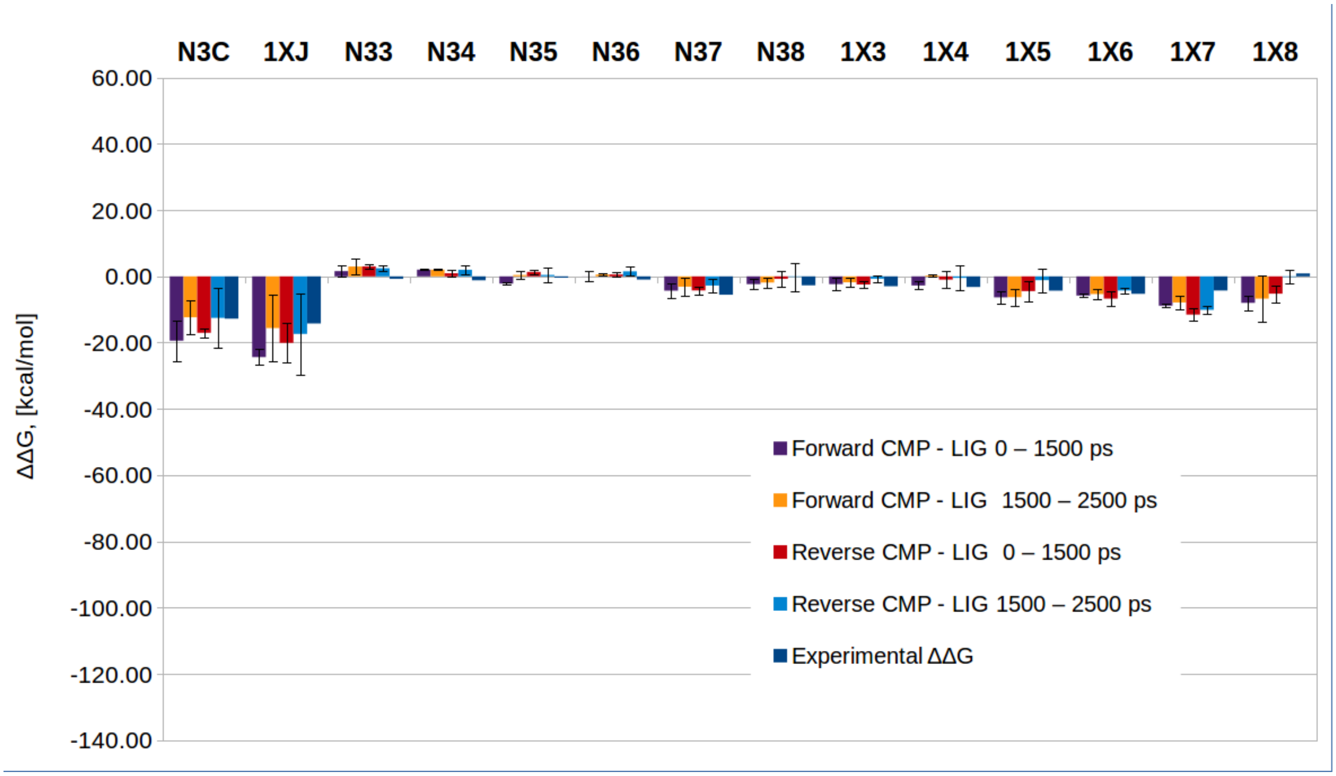


Figure 4.7. Calculated ΔΔGs. Computed ΔΔGs from the 11 λ-window alchemical transformations, averaged over the three replicas; standard deviations are also shown. Differences between ligand transformations in complex with Bcl-xL or Bcl-2 (labeled "CMP" in Figure 4.6) and transformations where the ligand is free in solution (labeled "LIG" in Figure 4.6) are labeled as "CMP - LIG," also shown is the experimentally measured free energy difference (last bar for every compound). ΔΔGs were evaluated from 0 – 1500 ps and from 1500 to 2500 ps of production dynamics; computed values are directly comparable to experimental data. For ease of comparison, *pmemd* results (labeled "Forward") are multiplied by -1. Compounds 3 – 8 are prefixed with "N3" and "1X" to designate transformations from/to the N3C and 1XJ templates, respectively.

4.3.4. Energetic fingerprinting

Examining the correlation plots of Bcl-xL bound to the different compounds (Figure 4.8A) reveals that the strong dual binders – compounds 3 and 4 – are energetically highly similar to each other and to the Bcl-xL-selective compounds – 5 and 6 – which are also potent Bcl-xL binders. Conversely, the dual and Bcl-xL selective ligands are energetically distinct from the Bcl-2-specific compounds, which are highly similar to each other. Furthermore, compound 7, which is the weakest Bcl-xL binder in the entire set, has the least energetic similarity to any of the other remaining compounds in two of the three replicas. This appears to be the general pattern for the Bcl-2 complexes as well (Figure 4.8B), with results being highly consistent among the three replicas. Moreover, comparison of energetic patterns in between proteins reveals that the dual binders elicit similar energetic responses in the two proteins (see the highlighted diagonals in Figure 4.9A). Conversely, the Bcl-xL specific binders appear to evoke less similar energetic responses, whereas the Bcl-2 selective compounds, like the dual binders, elicit similar responses in Bcl-xL and Bcl-2. Comparison of energetic similarities to Tanimoto, topological or MACCS key similarities – widely used measures of compound similarity in drug design – reveals another interesting pattern – dual binders (compounds 3 and 4) are structurally and energetically similar, Bcl-xL-selective binders (5 and 6) are structurally similar, but energetically distinct, whereas Bcl-2-selective ligands (7 and 8) are structurally different, but energetically similar (compare the three squares lying on the highlighted diagonals). Finally, Bcl-xL selective compounds are energetically least similar to Bcl-2 selective compounds than to any other compound in the dataset (see the off-diagonal squares in Figure 4.9A).

4.4. Discussion

4.4.1. Template structure equilibration simulations

The antiapoptotic proteins remained stable throughout the 100 ns trajectories, in agreement with extensive previous simulations (Ivanov et al., 2016). The ABT-263 template compound remained stably bound to the Bcl-2 protein, whereas the highly similar ABT-737 template displayed greater mobility, particularly the fragment that fits into pocket 2 (labeled “p2” in Figure 4.1) of the protein. This appears to be due to the topology of the Bcl-xL groove around the biphenyl fragment of ABT-737, which is shallower than that of Bcl-2, in part due to the A104D and S122R amino acid substitutions (nomenclature is Bcl-xL→Bcl-2; numbering corresponds to the canonical Bcl-xL sequnce). Pocket 2 accommodates the biphenyl fragment of ABT-737 and its corresponding moieties in other ligands. Loss

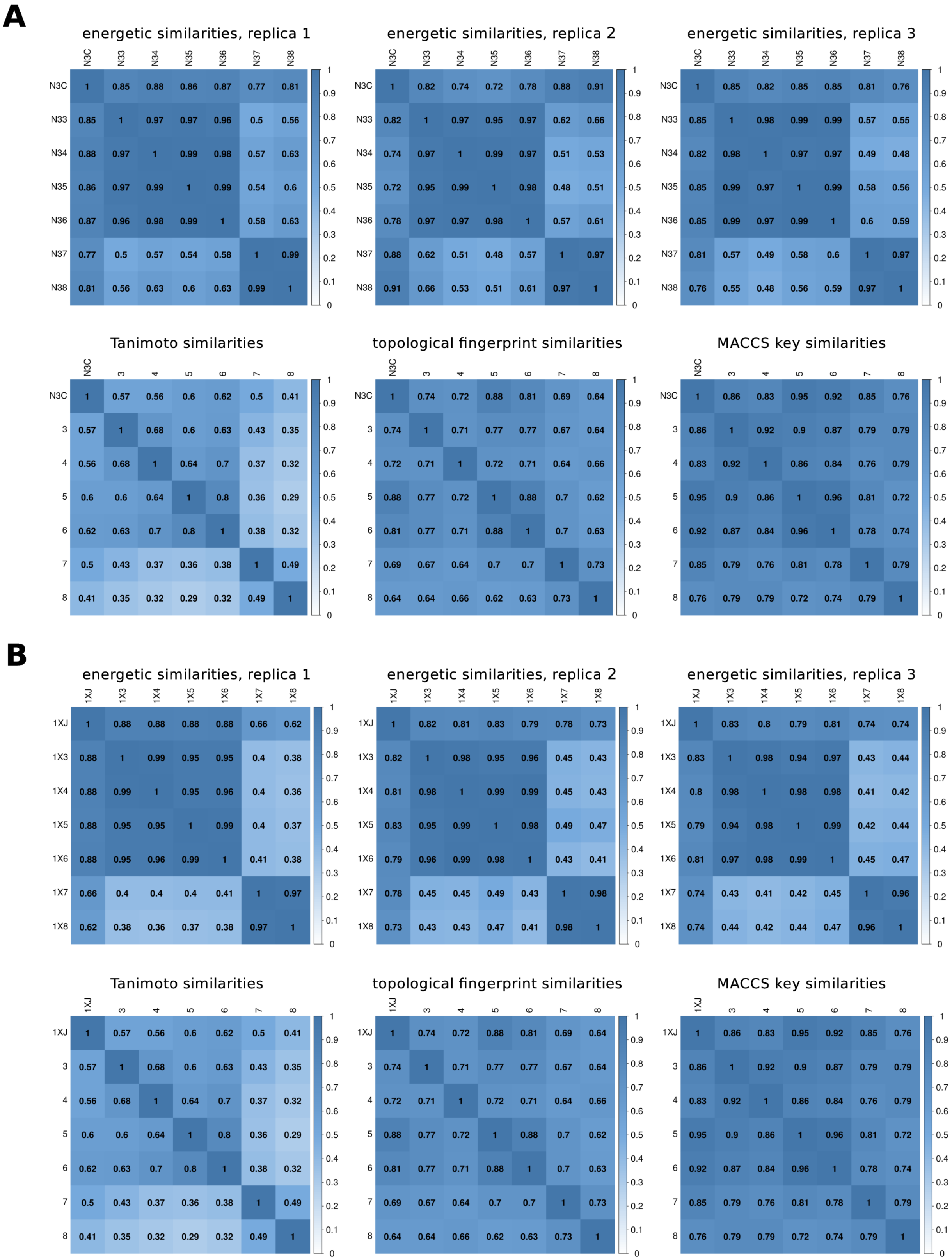

A
energetic similarities, replica 1
N3C N33 N34 N35 N36 N37 N38
N3C 1 0.85 0.88 0.86 0.87 0.77 0.81
N33 0.85 1 0.97 0.97 0.96 0.5 0.56
N34 0.88 0.97 1 0.99 0.98 0.57 0.63
N35 0.86 0.97 0.99 1 0.99 0.54 0.6
N36 0.87 0.96 0.98 0.99 1 0.58 0.63
N37 0.77 0.5 0.57 0.54 0.58 1 0.99
N38 0.81 0.56 0.63 0.6 0.63 0.99 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
energetic similarities, replica 2
N3C N33 N34 N35 N36 N37 N38
N3C 1 0.82 0.74 0.72 0.78 0.88 0.91
N33 0.82 1 0.97 0.95 0.97 0.62 0.66
N34 0.74 0.97 1 0.99 0.97 0.51 0.53
N35 0.72 0.95 0.99 1 0.98 0.48 0.51
N36 0.78 0.97 0.97 0.98 1 0.57 0.61
N37 0.88 0.62 0.51 0.48 0.57 1 0.97
N38 0.91 0.66 0.53 0.51 0.61 0.97 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
energetic similarities, replica 3
N3C N33 N34 N35 N36 N37 N38
N3C 1 0.85 0.82 0.85 0.85 0.81 0.76
N33 0.85 1 0.98 0.99 0.99 0.57 0.55
N34 0.82 0.98 1 0.97 0.97 0.49 0.48
N35 0.85 0.99 0.97 1 0.99 0.58 0.56
N36 0.85 0.99 0.97 0.99 1 0.6 0.59
N37 0.81 0.57 0.49 0.58 0.6 1 0.97
N38 0.76 0.55 0.48 0.56 0.59 0.97 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
Tanimoto similarities
N3C 3 4 5 6 7 8
N3C 1 0.57 0.56 0.6 0.62 0.5 0.41
3 0.57 1 0.68 0.6 0.63 0.43 0.35
4 0.56 0.68 1 0.64 0.7 0.37 0.32
5 0.6 0.6 0.64 1 0.8 0.36 0.29
6 0.62 0.63 0.7 0.8 1 0.38 0.32
7 0.5 0.43 0.37 0.36 0.38 1 0.49
8 0.41 0.35 0.32 0.29 0.32 0.49 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
topological fingerprint similarities
N3C 3 4 5 6 7 8
N3C 1 0.74 0.72 0.88 0.81 0.69 0.64
3 0.74 1 0.71 0.77 0.77 0.67 0.64
4 0.72 0.71 1 0.72 0.71 0.64 0.66
5 0.88 0.77 0.72 1 0.88 0.7 0.62
6 0.81 0.77 0.71 0.88 1 0.7 0.63
7 0.69 0.67 0.64 0.7 0.7 1 0.73
8 0.64 0.64 0.66 0.62 0.63 0.73 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
MACCS key similarities
N3C 3 4 5 6 7 8
N3C 1 0.86 0.83 0.95 0.92 0.85 0.76
3 0.86 1 0.92 0.9 0.87 0.79 0.79
4 0.83 0.92 1 0.86 0.84 0.76 0.79
5 0.95 0.9 0.86 1 0.96 0.81 0.72
6 0.92 0.87 0.84 0.96 1 0.78 0.74
7 0.85 0.79 0.76 0.81 0.78 1 0.79
8 0.76 0.79 0.79 0.72 0.74 0.79 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
B
energetic similarities, replica 1
1XJ 1X3 1X4 1X5 1X6 1X7 1X8
1XJ 1 0.88 0.88 0.88 0.88 0.66 0.62
1X3 0.88 1 0.99 0.95 0.95 0.4 0.38
1X4 0.88 0.99 1 0.95 0.96 0.4 0.36
1X5 0.88 0.95 0.95 1 0.99 0.4 0.37
1X6 0.88 0.95 0.96 0.99 1 0.41 0.38
1X7 0.66 0.4 0.4 0.4 0.41 1 0.97
1X8 0.62 0.38 0.36 0.37 0.38 0.97 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
energetic similarities, replica 2
1XJ 1X3 1X4 1X5 1X6 1X7 1X8
1XJ 1 0.82 0.81 0.83 0.79 0.78 0.73
1X3 0.82 1 0.98 0.95 0.96 0.45 0.43
1X4 0.81 0.98 1 0.99 0.99 0.45 0.43
1X5 0.83 0.95 0.99 1 0.98 0.49 0.47
1X6 0.79 0.96 0.99 0.98 1 0.43 0.41
1X7 0.78 0.45 0.45 0.49 0.43 1 0.98
1X8 0.73 0.43 0.43 0.47 0.41 0.98 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
energetic similarities, replica 3
1XJ 1X3 1X4 1X5 1X6 1X7 1X8
1XJ 1 0.83 0.8 0.79 0.81 0.74 0.74
1X3 0.83 1 0.98 0.94 0.97 0.43 0.44
1X4 0.8 0.98 1 0.98 0.98 0.41 0.42
1X5 0.79 0.94 0.98 1 0.99 0.42 0.44
1X6 0.81 0.97 0.98 0.99 1 0.45 0.47
1X7 0.74 0.43 0.41 0.42 0.45 1 0.96
1X8 0.74 0.44 0.42 0.44 0.47 0.96 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
Tanimoto similarities
1XJ 3 4 5 6 7 8
1XJ 1 0.57 0.56 0.6 0.62 0.5 0.41
3 0.57 1 0.68 0.6 0.63 0.43 0.35
4 0.56 0.68 1 0.64 0.7 0.37 0.32
5 0.6 0.6 0.64 1 0.8 0.36 0.29
6 0.62 0.63 0.7 0.8 1 0.38 0.32
7 0.5 0.43 0.37 0.36 0.38 1 0.49
8 0.41 0.35 0.32 0.29 0.32 0.49 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
topological fingerprint similarities
1XJ 3 4 5 6 7 8
1XJ 1 0.74 0.72 0.88 0.81 0.69 0.64
3 0.74 1 0.71 0.77 0.77 0.67 0.64
4 0.72 0.71 1 0.72 0.71 0.64 0.66
5 0.88 0.77 0.72 1 0.88 0.7 0.62
6 0.81 0.77 0.71 0.88 1 0.7 0.63
7 0.69 0.67 0.64 0.7 0.7 1 0.73
8 0.64 0.64 0.66 0.62 0.63 0.73 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
MACCS key similarities
1XJ 3 4 5 6 7 8
1XJ 1 0.86 0.83 0.95 0.92 0.85 0.76
3 0.86 1 0.92 0.9 0.87 0.79 0.79
4 0.83 0.92 1 0.86 0.84 0.76 0.79
5 0.95 0.9 0.86 1 0.96 0.81 0.72
6 0.92 0.87 0.84 0.96 1 0.78 0.74
7 0.85 0.79 0.76 0.81 0.78 1 0.79
8 0.76 0.79 0.79 0.72 0.74 0.79 1
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0

Figure 4.8. Pearson correlations between per-residue energy values. (A) Pearson correlation coefficients between the per-residue energy values of Bcl-xL for the three replicas of the *sander* transformations (top row), and ligand Tanimoto, topological fingerprint, and MACCS key similarities (bottom row). Compounds 3 - 8 are prefixed with "N3" to designate transformations to the N3C template for the alchemical transformations; correlations have been calculated from all protein residues using energy values from the latter 1 ns of production dynamics in every λ-window. **(B)** Pearson correlation coefficients between the per-residue energy values of Bcl-2 for the three replicas of the *sander* transformations (top row), and ligand Tanimoto, topological fingerprint, and MACCS key similarities (bottom row). Compounds 3 - 8 are prefixed with "1X" to designate transformations to the 1XJ template for the alchemical transformations; correlations have been calculated from all protein residues using energy values from the latter 1 ns of production dynamics in every λ-window.

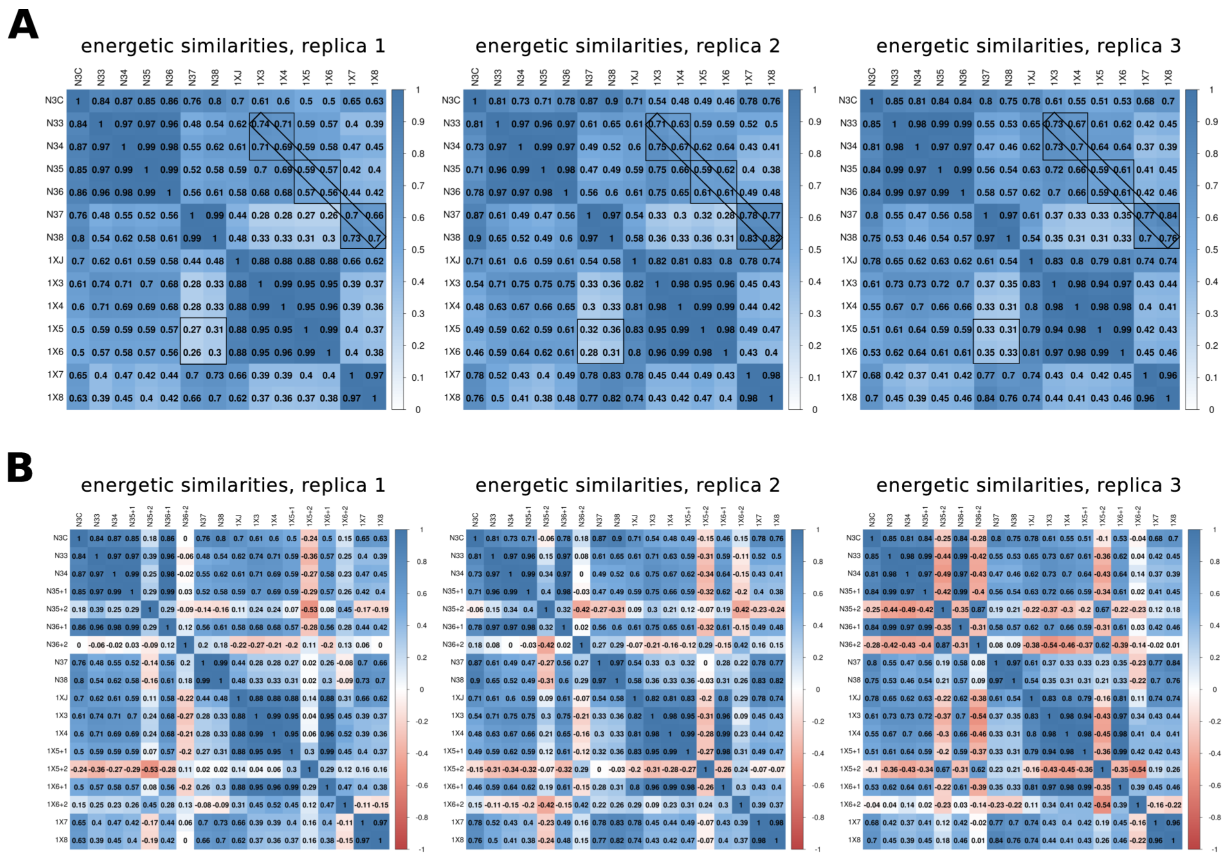


Figure 4.9. Pearson correlations between per-residue energy values. (A) Pearson correlation coefficients between the per-residue energy values of Bcl-xL and Bcl-2 for the three replicas of the *sander* transformations. Compounds 3 – 8 are prefixed with "N3" or "1X" to designate transformations

to the N3C and 1XJ templates, respectively; correlations have been calculated from Bcl-xL residues 25 – 137 and Bcl-2 residues 27 – 139 using energy values from the latter 1 ns of production dynamics in every λ-window. The energetic responses of Bcl-xL and Bcl-2 when bound to the same ligand are compared (highlighted diagonals), as are the energetic responses between Bcl-xL and Bcl-2 bound to compounds 3 and 4, 5 and 6, 7 and 8 (squares lying on the diagonals), and compounds 5 and 6, and 7 and 8 (off-diagonal squares). **(B)** Pearson correlation coefficients between the per-residue energy values of Bcl-xL and Bcl-2 for the three replicas of the *sander* transformations. Compounds 3 – 8 are prefixed with "N3" or "1X" to designate transformations to the N3C and 1XJ templates, respectively; correlations have been calculated from Bcl-xL residues 25 – 137 and Bcl-2 residues 27 – 139 using energy values from the latter 1 ns of production dynamics in every λ-window. Apart from the +1 protonation state, reported heretofore, compounds 5 and 6 are also examined in the +2 protonation state, where the piperazynyl nitrogen (see Figure 4.2) is protonated. The +1 and +2 states are designated with a corresponding suffix in this panel.

of this interaction, in turn, facilitates the loss of the intramolecular π-stacking interaction between the thiophenyl and nitrophenyl moieties in pocket 4 ("p4" in Figure 4.1) – a characteristic feature of the thiophenyl-bearing sulphoneamidoaryl class of Bcl-inhibitors, observed in multiple X-ray and NMR structures (Souers et al., 2013; Lee et al. 2007; Oltersdorf et al. 2005; Bruncko et al. 2007; Touré et al., 2013; Tanaka et al. 2013). For compounds of similar affinity to Bcl-xL and Bcl-2, these effects are likely to result in greater k_{off} rates for Bcl-xL than Bcl-2.

4.4.2. Thermodynamic integration simulations

Presented herein are thermodynamic integration calculations on large, complex, drug and drug-like molecules. It is reassuring that most of the relative transformations approach a value within 2 kcal/mol from the experimental ΔΔG. As can be expected, only the absolute ΔG calculations are off by a great margin, sometimes more than 10 kcal/mol. The reasons for this are twofold. First, inserting/deleting the ligand into/from the system is a much greater transformation than transforming the template compounds into the models or vice versa. Thus, it requires more sampling than the relative transformations. Second, the alchemical transformations described here involve charged species. This introduces an error of a complex nature, involving periodicity-induced net-charge interactions, periodicity-induced net-charge undersolvation, discrete solvent effects, and residual integrated potential effects (Rocklin et al., 2013b). Moreover, this error is likely to be compounded by the circumstance that the template ligands N3C and 1XJ, which have a net charge of +2, are transformed into the model ligands, which all have a net charge of +1 or vice versa. The error is likely to largely cancel in the relative transformations, and judging by the convergence plots (see SI figures 4.1 and 4.3) it does, but not in the absolute free energy calculations. Thus, it is shown that large and complex drug molecules can be reliably parameterized with a general force field, such as GAFF, and a rapid method for charge estimation, such as AM1-BCC. Absolute binding energy calculations of charged species, however, necessitate further theoretical and methodological developments (Liu et al., 2013; Boyce et al., 2009; Rocklin et al., 2013b).

The foregoing error also influences the per-residue values obtained from energy decomposition. To assess its influence, simulations on compounds 5 and 6 were also performed in the +2 protonation state (see Figure 4.3). This state is less likely, because the piperazynyl nitrogen is bound to electron withdrawing groups. Moreover, protonation of the nitrogen atom stabilized it in a tetrahedral geometry, which facilitated inversions in the terminal moieties where the phenyl groups pointed outward into the solvent and the polar atoms pointed inward into the hydrophobic groove. Such an arrangement seems unlikely. Conversely, the +1 protonation state stabilized the ligands in the groove. In the case of Bcl-xL, coupled with the relative shallowness of the groove around pocket 2, it also increased the propensity of the terminal moieties of ligands 5 and 6 to adopt an extended conformation along the surface of the groove. Such an arrangement has been observed crystallographically in similar compounds bound to Bcl-xL (Bruncko et al., 2007). Nevertheless, simulating compounds 5 and 6 in the +2 state reveals an interesting aspect of free energy calculations. It appears that the charging/decharging error strongly influences the per-residue values, perhaps more so than the behavior of the ligand itself. The +2 are the only simulations where the ligand net charge does not change during the alchemical transformation (Figure 4.9B). The per-residue values are quite different from rest of the simulations, where one of the ligand end states has a charge of +2 or 0, the other +1 or 0 (0 when the ligand is decoupled from the system (λ = 1) or inserted into the system (λ = 0)); this is the only instance where anticorrelations appear. Since the error resulting from the change in charge is present in all other transformations, it is to be expected that the energetic patterns reported herein are *bona fide*; it is extremely unlikely that they have arisen purely by chance, especially in three independent replicas.

4.4.3. QSAR and energetic fingerprinting

QSAR models, derived from the arylsulfonamidearyl series of compounds, perform well when predicting properties of other arylsulfonamidearyls. Similarly, training on the tetrahydroisoquinolines produces predictive models for other tetrahydroisoquinolines. However, training on one series of compounds to predict the properties of another leads to poor results. Moreover, Bcl-2-specific arylsulfonamidearyls, of which there are only two in the entire dataset, are poorly predicted by the present models. This is a well known shortcoming of traditional QSAR models – they are much better at interpolation than extrapolation (Biniashvili et al., 2012; Roy et al., 2015; Doweyko, 2008) – and even the most sophisticated QSAR models perform poorly on challenging targets (Fourches et al., 2013). Moreover, predictive performance is sensitive to the specific choice of test and training compounds, very much so in the case of small datasets. Quite often, QSAR models suffer from "activity cliffs" (Cruz-Monteagudo et al., 2014) where a small change in a compound, such as introduction or removal of a hydroxyl or a methyl group, translates into a large change in activity. Most models, based on 2D descriptors, or 2D measures of similarity, such as Tanimoto coefficients or MACCS keys, would experience little change upon such an alteration in a large molecule. It appears that two-dimensional models and descriptors simply do not contain in themselves all the information relevant to ligand activity. Similar findings have been reported previously (Oprea, 2002; Stefaniak 2015). In hindsight, this is not surprising – the descriptors used to construct low-dimensionl QSAR models are, by the very nature of the technique, invariant with respect to the macromolecule – a major determinant of the binding process; they are also invariant with respect to time. Moreover, most 0-, 1-, and 2-dimensional descriptors are not even capable of discriminating between enantiomers. In stark contrast, free energy calculations and energy decomposition should, at least in principle, provide a detailed understanding of the binding process. Distinguishing enantiomers, for example, becomes trivial by the very nature of the technique – where one enantiomer makes favorable interactions, and the other does not, there will exist a difference in potentials between the two systems, a difference that will manifest itself in a computationally detectable $\Delta\Delta G$. Moreover, the difference in residue - ligand contacts should, again, at least in principle, be evident on a per-residue level; finer-grained decomposition is also possible (Archontis et al., 1998).

It is not surprising that QSAR models, based on ligand properties computed from molecular dynamics trajectories, are much better at predicting ligand biological activity and resolving activity cliffs than 2D- and 3D-based models (Ash and Fourches, 2017). What is particularly noteworthy about the study by Ash and Fourches on a set of kinase inhibitors (2017), is that for some of the descriptors, it is not the means of the descriptors that correlate with biological activity, but their standard deviations over the course of the simulations, with more active ligands displaying greater variance. The authors hypothesize that this is a result of the ligands adapting to the dynamic kinase pocket, and demonstrate how 4D QSAR models, the fourth dimension being time, add an extra layer of information content over 3D models. Perhaps most important of all, they show that 4D models are capable of yielding information and features that no lower-level model is capable of providing. Perhaps, this is also not surprising – biological activity, be it K_i, or K_d, or IC_{50}, or EC_{50}, or another such value, measured in a laboratory, is, of course, a Boltzmann-weighted average of many conformations; molecules are never static.

While characterizing a ligand's dynamic behavior in complex with a biological macromolecule of interest is certainly a prerequisite to fully understanding binding, it is only one side of the medal. Clearly, the other side is characterizing the macromolecule's behavior with a particular binding partner. In the case of Bcl-2-selective arylsulfonamidearyls, for instance, traditional QSAR models perform poorly when predicting their specificity, as most presently available arylsulfonamidearyls are dual binders or Bcl-xL selective. Molecular dynamics and free energy calculations, however, demonstrate that the Bcl-xL and Bcl-2 proteins differentiate between these ligands from dual and Bcl-xL selective binders, as evident from the current analysis, which can be viewed as an assessment of a protein's energetic response to a particular binding partner.

Including water molecules and ions in the per-residue analysis in such calculations would enable detecting bridging (Ahmed et al., 2011) or trapped (Stegmann et al., 2009) waters/ions, which are often important, but neglected, factors in the binding process. This would mitigate the need for GIST- (Nguyen et al., 2012) or 3D-RISM-type calculations (Imai et al., 2003), and reduce the human workload, for example in a lead-optimization campaign, at no additional computational cost.

4.5. Outlook

A means of characterizing protein – ligand/peptide/protein interactions is presented and used to demonstrate how macromolecules energetically distinguish between different binding partners. This work paves the way for developing and validating more predictive descriptors, based on per residue energies. Such energy-based four dimensional descriptors (Pan et al., 2003; da Rocha Pita et al., 2012) would constitute a composite of multiple lower-level descriptors, encompassing in themselves aspects of the binding process such as steric and atom-type propensities for given ligands at given receptor locations, e.g. polar groups with positive partial charges in the vicinity of a key hot spot or specificity determinant (Ivanov et al., 2016; Ivanov et al., 2017). Thus, it is to be expected that with the increase in computing power, higher-dimensional descriptors will become more prevalent, at the expense of lower-level descriptors.

4.6. References

Adelman, S.A. and Doll, J.D., 1974. Generalized Langevin equation approach for atom/solid-surface scattering: Collinear atom/harmonic chain model. *The Journal of Chemical Physics*, 61(10), pp.4242–4245.

Aguilar, A. et al., 2013. A Potent and Highly Efficacious Bcl-2/Bcl-xL Inhibitor. *Journal of medicinal chemistry*, 56(7), pp.3048–3067.

Ahmed, M.H. et al., 2011. Bound water at protein-protein interfaces: Partners, roles and hydrophobic bubbles as a conserved motif. *PLoS ONE*, 6(9), p.e24712.

Aiello, D. and Caffrey, D.R., 2012. Evolution of specific protein-protein interaction sites following gene duplication. *Journal of molecular biology*, 423(2), pp.257–72.

Akhtar, R.S., Ness, J.M. and Roth, K.A., 2004. Bcl-2 family regulation of neuronal development and neurodegeneration. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1644(2), pp.189–203.

Aldeghi, M. et al., 2016. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.*, 7(1), pp.207–218.

Anon, Schrödinger Release 2017-2: Maestro, Schrödinger, LLC, New York, NY, 2017.

Archontis, G. et al., 1998. Specific amino acid recognition by aspartyl-tRNA synthetase studied by free energy simulations. *Journal of Molecular Biology*, 275(5), pp.823–846.

Ash, J. and Fourches, D., 2017. Characterizing the Chemical Space of ERK2 Kinase Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *Journal of Chemical Information and Modeling*, 57(6), pp.1286–1299.

Balaban, A.T., 1982. Highly Discriminating Distance-Based Topological index. *Chemical Physics Letters*, 89(5), pp.399–404.

Berendsen, H.J.C. et al., 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(1984), pp.3684–3690.

Bertz, S., 1981. The first general index of molecular complexity. *Journal of the American Chemical Society*, 83(7), pp.3599–3601.

Berman, H., Henrick, K. and Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. *Nature structural biology*, 10(12), p.980.

Biniashvili, T., Schreiber, E. and Kliger, Y., 2012. Improving classical substructure-based virtual screening to handle extrapolation challenges. *Journal of Chemical Information and Modeling*, 52(3), pp.678–685.

Bonchev, D. and Trinajstić, N., 1977. Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics*, 67(10), pp.4517–4533.

Bouillet, P. et al., 2001. Degenerative Disorders Caused by Bcl-2 Deficiency Prevented by Loss of Its

BH3-Only Antagonist Bim. *Developmental Cell*, 1(5), pp.645–653.

Boyce, S.E. et al., 2009. Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. *Journal of Molecular Biology*, 394(4), pp.747–763.

Brady, R.M. et al., 2014. De-novo designed library of benzoylureas as inhibitors of BCL-XL: synthesis, structural and biochemical characterization. *Journal of Medicinal Chemistry*, 57, pp.1323–43.

Brandsdal, B.O. et al., 2003. Free Energy Calculations and Ligand Binding. *Advances in Protein Chemistry*, 66, pp.123–158.

Bruncko, M. et al., 2007. Studies leading to potent, dual inhibitors of Bcl-2 and Bcl-xL. *Journal of Medicinal Chemistry*, 50(4), pp.641–662.

Burusco, K.K. et al., 2015. Free Energy Calculations using a Swarm-Enhanced Sampling Molecular Dynamics Approach. *ChemPhysChem*, 16(15), pp.3233–3241.

Case, D.A. et al., 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), pp.1668–1688.

Chodera, J.D. et al., 2011. Alchemical free energy methods for drug discovery: Progress and challenges. *Current Opinion in Structural Biology*, 21(2), pp.150–160.

Christ, C.D. and Fox, T., 2014. Accuracy assessment and automation of free energy calculations for drug design. *Journal of Chemical Information and Modeling*, 54(1), pp.108–120.

Ciccotti, G. and Ryckaert, J.P., 1986. Molecular dynamics simulation of rigid molecules. *Computer Physics Reports*, 4(6), pp.346–392.

Cohen, P. and Tcherpakov, M., 2010. Will the ubiquitin system furnish as many drug targets as protein kinases? *Cell*, 143(5), pp.686–693.

Cruz-Monteagudo, M. et al., 2014. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, 19(8), pp.1069–1080.

Csizmadia, F. et al., 1997. Prediction of distribution coefficient from structure. 1. Estimation method. *Journal of Pharmaceutical Sciences*, 86(7), pp.865–871.

Curtis, R.A. and Lue, L., 2006. A molecular approach to bioseparations: Protein–protein and protein–salt interactions. *Chemical Engineering Science*, 61(3), pp.907–923.

Cuthbertson, J.M., Bond, P.J. and Sansom, M.S.P., 2006. Transmembrane helix-helix interactions: Comparative simulations of the glycophorin A dimer. *Biochemistry*, 45(Md), pp.14298–14310.

Gilson, M.K. and Zhou, H.-X., 2007. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36, pp.21–42.

Darden, T., York, D. and Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(1993), p.10089.

Dixon, S.L. and Jurs, P.C., 1993. Estimation of pKa for organic oxyacids using calculated atomic charges. *Journal of Computational Chemistry*, 14(12), pp.1460–1467.

Doweyko, A.M., 2008. QSAR: Dead or alive? *Journal of Computer-Aided Molecular Design*, 22(2), pp.81–89.

Dror, R.O. et al., 2012. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41, pp.429–52.

Duchowicz, P.R., Castro, E.A. and Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Communications in Mathematical and in Computer Chemistry*, 55, pp.179–192.

Durant, J.L. et al., 2002. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Model*. 42(6):1273–80.

Ertl, P., Rohde, B. and Selzer, P., 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20), pp.3714–3717.

Feng, Y. et al., 2010. Design, synthesis, and interaction study of quinazoline-2(1H)-thione derivatives as novel potential Bcl-xL inhibitors. *Journal of medicinal chemistry*, 53, pp.3465–3479.

Fourches, D. et al., 2013. Predicting Binding Affinity of CSAR Ligands Using Both Structure- Based and Ligand-Based Approaches. *Journal of Chemical Information and Modeling*, 53(8), pp.1915–1922.

Friendly, M. and Friendly, M., 2002. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 34(9), pp.1447–9.

Gandhi, L. et al., 2011. Phase I study of navitoclax (ABT-263), a novel bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors. *Journal of Clinical Oncology*, 29(7), pp.909–916.

Gasteiger, J. and Marsili, M., 1980. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron*, 36(22), pp.3219–3228.

Gonzalez, M.W. and Kann, M.G., 2012. Chapter 4: Protein Interactions and Disease. *PLoS Computational Biology*, 8(12), p.e1002819.

Gotoh, Y. et al., 2010. Two-component signal transduction as potential drug targets in pathogenic bacteria. *Current opinion in microbiology*, 13(2), pp.232–9.

Gramatica, P., 2007. Principles of QSAR models validation: Internal and external. *QSAR and Combinatorial Science*, 26(5), pp.694–701.

Gromiha, M.M. et al., 2011. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome science*, 9(Suppl 1), p.S13.

Gromiha, M.M., Yokota, K. and Fukui, K., 2009. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Molecular bioSystems*, 5(12), pp.1779–1786.

Gu, M; Lai, T.L., 1991. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis.

Annals of Statistics, 19(3), pp.1403–1433.

Hall, L.H. and Kier, L.B., 1991. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Reviews in Computational Chemistry*, 2, pp.367–422.

Hansen, N. and van Gunsteren, W.F., 2014. Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, 10(7), pp.2632–2647.

Hennessy, E.J., 2016. Selective inhibitors of Bcl-2 and Bcl-xL: Balancing antitumor activity with on-target toxicity. *Bioorganic and Medicinal Chemistry Letters*, 26(9), pp.2105–2114.

Homeyer, N. et al., 2014. Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context. *Journal of Chemical Theory and Computation*, 10(8), pp.1–31.

Hopfinger, A.J. et al., 1997. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society*, 7863(5), pp.10509–10524.

Hornak, V. and Simmerling, C., 2004. Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. *Journal of Molecular Graphics and Modelling*, 22(5), pp.405–413.

Imai, T., Kovalenko, A. and Hirata, F., 2004. Solvation thermodynamics of protein studied by the 3D-RISM theory. *Chemical Physics Letters*, 395(1–3), pp.1–6.

Ivanov, S., Dimitrov, I. and Doytchinova, I., 2013. Quantitative prediction of peptide binding to HLA-DP1 protein. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3), pp.811–815.

Ivanov, S.M. et al., 2016. Energetics and Dynamics Across the Bcl-2-Regulated Apoptotic Pathway Reveal Distinct Evolutionary Determinants of Specificity and Affinity. *Structure*, 24(11), pp.2024–2033.

Ivanov, S.M. et al., 2017. Protein – protein interactions in paralogues: electrostatics modulates specificity on a conserved steric scaffold. *PLoS ONE*, 12(10), pp.e0185928

Jakalian, A. et al., 2000. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry*, 21(2), pp.132–146.

Jorgensen, W.L. et al., 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(1983), p.926.

Kaus, J.W. et al., 2013. Improving the efficiency of free energy calculations in the Amber molecular dynamics package. *Journal of Chemical Theory and Computation*, 9(9), pp.1–8.

Klimovich, P. V., Shirts, M.R. and Mobley, D.L., 2015. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5), pp.397–411.

Kollman, P., 1993. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chemical Reviews*, 93(7), pp.2395–2417.

Labute, P., 2000. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 18(4–5), pp.464–477.

Lee, E.F. et al., 2009. Conformational changes in Bcl-2 pro-survival proteins determine their capacity to bind ligands. *Journal of Biological Chemistry*, 284(44), pp.30508–30517.

Lee, E.F. et al., 2007. Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. *Cell death and differentiation*, 14(9), pp.1711–1713.

Lessene, G. et al., 2013. Structure-guided design of a selective BCL-XL inhibitor. *Nat Chem Biol*, 9(6), pp.390–397.

Lessene, G., Czabotar, P.E. and Colman, P.M., 2008. BCL-2 family antagonists for cancer therapy. *Nat Rev Drug Discov*, 7(December), pp.989–1000.

Leverson, J.D. et al., 2015. Exploiting selective BCL-2 family inhibitors to dissect cell survival dependencies and define improved strategies for cancer therapy. *Science translational medicine*, 7(279), p.279ra40.

Liu, S. et al., 2013. Lead optimization mapper: Automating free energy calculations for lead optimization. *Journal of Computer-Aided Molecular Design*, 27(9), pp.755–770.

Lyne, P.D., Lamb, M.L. and Saeh, J.C., 2006. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *Journal of Medicinal Chemistry*, 49(16), pp.4805–4808.

Maier, J.A. et al., 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), pp.3696–3713.

Mevik, B.-H. and Wehrens, R., 2007. The pls Package: Principle Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2), pp.1–24.

Miller, B.R. et al., 2012. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.*, 8(9), p.pp 3314–3321.

Mitchell, J.B.O., 2014. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5), pp.468–481.

Mobley, D.L. et al., 2007. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *Journal of Molecular Biology*, 371(4), pp.1118–1134.

Mobley, D.L. and Klimovich, P. V., 2012. Perspective: Alchemical free energy calculations for drug discovery. *Journal of Chemical Physics*, 137(23), p.230901.

Moldoveanu, T. et al., 2014. Many players in BCL-2 family affairs. *Trends in Biochemical Sciences*, 39(3), pp.101–111.

Murdoch, D.J. and Chow, E.D., 1996. A Graphical Display of Large Correlation Matrices. *The American Statistician*, 50(2), pp.178–180.

Ng, S.Y. and Davids, M.S., 2014. Selective Bcl-2 Inhibition to Treat Chronic Lymphocytic Leukemia and Non-Hodgkin Lymphoma. *Clinical Advances in Hematology and Oncology*, 12(4), pp.224–229.

Nguyen, C.N., Young, T.K. and Gilson, M.K., 2012. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *Journal of Chemical Physics*, 137(14), pp.1–17.

Nguyen, K.T. et al., 2009. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem*, 4(11), pp.1803–1805.

Okamoto, T. et al., 2012. Sheeppox Virus SPPV14 Encodes a Bcl-2-Like Cell Death Inhibitor That Counters a Distinct Set of Mammalian Proapoptotic Proteins. *Journal of virology*, 86(21), pp.11501–11.

Oltersdorf, T. et al., 2005. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, 435(June), pp.677–81.

Oprea, T.I., 2002. On the information content of 2D and 3D descriptors for QSAR. *Journal of the Brazilian Chemical Society*, 13(6), pp.811–815.

Pan, D., Tseng, Y. and Hopfinger, A.J., 2003. Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *Journal of Chemical Information and Computer Sciences*, 43(5), pp.1591–1607.

Park, C.M. et al., 2008. Discovery of an orally bioavailable small molecule inhibitor of prosurvival B-cell lymphoma 2 proteins. *Journal of Medicinal Chemistry*, 51(21), pp.6902–6915.

Park, D. et al., 2013. Novel small-molecule inhibitors of Bcl-XL to treat lung cancer. *Cancer Research*, 73(17), pp.5485–5496.

Patronov, A. et al., 2012. Peptide binding to HLA-DP proteins at pH 5.0 and pH 7.0: a quantitative molecular docking study. *BMC structural biology*, 12(1), p.20.

Patronov, A. and Doytchinova, I., 2013. T-cell epitope vaccine design by immunoinformatics. *Open biology*, 3(1), p.120139.

Perez, H.L. et al., 2012. Identification of a phenylacylsulfonamide series of dual Bcl-2/Bcl-xL antagonists. *Bioorganic and Medicinal Chemistry Letters*, 22(12), pp.3946–3950.

Porter, J. et al., 2009. Tetrahydroisoquinoline amide substituted phenyl pyrazoles as selective Bcl-2 inhibitors. *Bioorganic and Medicinal Chemistry Letters*, 19(1), pp.230–233.

da Rocha Pita, S.S. et al., 2012. Receptor-Dependent 4D-QSAR Analysis of Peptidemimetic Inhibitors of Trypanosoma cruzi Trypanothione Reductase with Receptor-Based Alignment. *Chemical Biology and Drug Design*, 79(5), pp.740–748.

Rocklin, G.J., et al., 2013a. Blind prediction of charged ligand binding affinities in a model binding site. *Journal of Molecular Biology*, 425(22), pp.4569–4583.

Rocklin, G.J., et al., 2013b. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *Journal of Chemical Physics*, 139(18), p.184103.

Rodriguez-Martinez, J.A. et al., 2017. Combinatorial bZIP dimers define complex DNA-binding specificity landscapes. *In review*, pp.1–29.

Rodriguez-Soca, Y. et al., 2010. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *Journal of proteome research*, 9(2), pp.1182–90.

Roe, D.R. and Cheatham, T.E., 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, 9(7), pp.3084–3095.

Roy, K., Kar, S. and Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 145, pp.22–29.

Rudin, C.M. et al., 2012. Phase II study of single-agent navitoclax (ABT-263) and biomarker correlates in patients with relapsed small cell lung cancer. *Clinical Cancer Research*, 18(5), pp.3163–3169.

Settimo, L., Bellman, K. and Knegtel, R.M.A., 2014. Comparison of the accuracy of experimental and predicted pKa values of basic and acidic compounds. *Pharmaceutical Research*, 31(4), pp.1082–1095.

Siryk-Bathgate, A., Dabul, S. and Lymperopoulos, A., 2013. Current and future G protein-coupled receptor signaling targets for heart failure therapy. *Drug design, development and therapy*, 7, pp.1209–1222.

Sleebs, B.E. et al., 2013. Discovery of potent and selective benzothiazole hydrazone inhibitors of Bcl-XL. *Journal of Medicinal Chemistry*, 56(13), pp.5514–5540.

Sleebs, B.E. et al., 2011. Quinazoline sulfonamides as dual binders of the proteins B-cell lymphoma 2 and B-cell lymphoma extra long with potent proapoptotic cell-based activity. *Journal of Medicinal Chemistry*, 54(6), pp.1914–1926.

Souers, A.J. et al., 2013. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, 19(2), pp.202–8.

Stefaniak, F., 2015. Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors. *Frontiers in Environmental Science*, 3(77), pp.1–7.

Stegmann, C.M. et al., 2009. The thermodynamic influence of trapped water molecules on a protein-ligand interaction. *Angewandte Chemie - International Edition*, 48(28), pp.5207–5210.

Steinbrecher, T. et al., 2008. Bornyl (3,4,5-trihydroxy)-cinnamate - An optimized human neutrophil elastase inhibitor designed by free energy calculations. *Bioorganic and Medicinal Chemistry*, 16(5), pp.2385–2390.

Steinbrecher, T., Case, D.A. and Labahn, A., 2006. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. *Journal of Medicinal Chemistry*, 49(6), pp.1837–1844.

Steinbrecher, T., Joung, I. and Case, D.A., 2011. Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. *Journal of Computational Chemistry*, 32(15), pp.3253–3263.

Tanaka, Y., Aikawa, K. and Nishida, G., 2013. Discovery of potent Mcl-1/Bcl-xL dual inhibitors by using a hybridization strategy based on structural analysis of target proteins. *Journal of Medicinal Chemistry*, 56(23), pp.9635–9645.

Tao, Z. et al., 2014. Discovery of a Potent and Selective Bcl-xL Inhibitor with in Vivo Activity. *ACS Med Chem Lett*, 5(10), pp.1088–1093.

Touré, B.B. et al., 2013. The Role of the acidity of N-heteroaryl sulfonamides as inhibitors of Bcl-2 family protein-protein interactions. *ACS Medicinal Chemistry Letters*, 4(2), pp.186–190.

Varnes, J.G. et al., 2014. Towards the next generation of dual Bcl-2/Bcl-xL inhibitors. *Bioorganic and Medicinal Chemistry Letters*, 24(14), pp.3026–3033.

Velev, O.D., Kaler, E.W. and Lenhoff, A.M., 1998. Protein Interactions in Solution Characterized by Light and Neutron Scattering: Comparison of Lysozyme and Chymotrypsinogen. *Biophysical Journal*, 75(6), pp.2682–2697.

Vicini, P. et al., 2002. Hydrazones of 1,2-benzisothiazole hydrazides: synthesis, antimicrobial activity and QSAR investigations. *European Journal of Medicinal Chemistry*, 37(7), pp.553–564.

Vogler, M. et al., 2009. Bcl-2 inhibitors: small molecules with a big impact on cancer therapy. *Cell Death and Differentiation*, 16(3), pp.360–367.

Wang, J.M. et al., 2004. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9), pp.1157–1174.

Wang, L. et al., 2015. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7), pp.2695–2703.

Wendt, M.D. et al., 2006. Discovery and structure- activity relationship of antagonists of B-cell lymphoma 2 family proteins with chemopotentiation activity in vitro and in vivo. *J. Med. Chem*, 49(3), pp.1165–1181.

Winkler, D.A., 2002. The role of quantitative structure--activity relationships (QSAR) in biomolecular discovery. *Briefings in bioinformatics*, 3(1), pp.73–86.

Wildman, S. and Crippen, G., 1999. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Modeling*, 39(5), pp.868–873.

Yousefinejad, S. and Hemmateenejad, B., 2015. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149, pp.177–204.

Yusuff, N. et al., 2012. Lipophilic isosteres of a π-π Stacking interaction: New inhibitors of the Bcl-2-Bak protein-protein interaction. *ACS Medicinal Chemistry Letters*, 3(7), pp.579–583.

Zhai, D. et al., 2006. Comparison of chemical inhibitors of antiapoptotic Bcl-2-family proteins. *Cell Death and Differentiation*, 13(8), pp.1419–1421.

Zhou, H. et al., 2012. Design of Bcl-2 and Bcl-xL inhibitors with subnanomolar binding affinities based upon a new scaffold. *Journal of Medicinal Chemistry*, 55(10), pp.4664–4682.

Chapter 5 - Conclusions

The present work presents methods to characterize protein – ligand/peptide/protein interactions and discern the origins of affinity in such complexes at high- and low-levels of theory. Furthermore, the goal of elucidating the "specificity determinants within families of protein – protein interactions" is pursued throughout and achieved in different ways – via computationally inexpensive, high-throughput side chain replacement and buried surface and Debye-Hückel calculations, and the more computationally demanding MM-PBSA calculations. Importantly, it shows how results from previous bioinformatics, structural, and mutational studies on affinity and specificity in transient protein – protein interactions (Jones and Thornton 1996; Lo Conte et al., 1999; Guharoy and Chakrabarti, 2005; Gromiha et al., 2009; Gromiha et al., 2011; Kosloff et al., 2011; Aiello and Caffrey, 2012; Kim et al., 2015) are all manifestations of the core and rim model of protein – protein binding, formulated by Chakrabarti and Janin at the turn of the millennium (Chakrabarti and Janin, 2002). This model can, in turn, be viewed as an extension of the O-ring model, proposed in 1998 by Bogan and Thorn, who noticed from alanine-scanning data that hot spots are surrounded by energetically unimportant residues, shielding them from solvent (Bogan and Thorn, 1998). Shortly thereafter, Lo Conte et al. (1999) observed that around a third of interface atoms are fully buried, i.e. have zero solvent accessible surface area, a third are in contact with immobilized water, and the remainder are in contact with bulk solvent, and proposed dividing the interface into a core and rim. That and subsequent structural studies have shown that the core typically comprises around 55% of the interface in single-patch, transient protein – protein interactions, as opposed to the more hydrophobic permanent complexes, and have revealed the different residue propensities of the core and rim, leading to the present day understanding of interprotein binding (Chakrabarti and Janin, 2002; Ma et al., 2003; Guharoy and Chakrabarti, 2005; Keskin et al., 2008; Gromiha et al., 2009; Gromiha et al., 2011; Cukuroglu et al., 2014; Guo et al., 2014; David and Sternberg, 2015).

Presented herein is a large-scale, thorough sequence, structural, energetics, dynamics, and bioinformatics analysis on protein – protein binding in paralogous systems, which unifies previously reported aspects of protein – protein binding into the framework of core and rim interfaces, and rationalizes its origins. Using datasets orders of magnitude larger than previous reports (Lo Conte et al., 1999; Chakrabarti and Janin, 2002), it is confirmed that nonpolar surface constitutes the majority of the interface (see Figures 2.2 – 2.7). Further, it is shown that cores are highly conserved, whereas rims are variable (see SI figure 2.1), in agreement with previous reports (Guharoy and Chakrabarti, 2005). Through detailed energetics analysis, it is shown how the pattern in sequence and structure conservation translates to energetics and binding patterns – the low sequence entropy core regions in SI figure 2.1 correspond to regions that invariably make favorable contributions to binding in Figure 3.2 and SI figure 3.2. Conversely, the high-entropy rim makes an energetic contribution to binding, which is variable, rather than uniform – favorable for some protein – peptide pairs, unfavorable for others, thereby modulating binding specificity. Finally, it is shown how mutations in paralogous signaling pathways become interdependent in between binding partners, and how all of the above is ultimately driven by the hydrophobic effect (Chandler, 2005). Perhaps, the time has now come to rename the "core and rim model" to "core and rim mode" of protein – protein interactions.

Chapter 2 describes a computationally cheap method to identify specificity determinants, simply based on charge interactions and buried surface, whereas chapters 3 and 4 provide means to probe deeper into the binding process, quantifying the energetics of binding and the binding patterns. This work also opens up several avenues for future work. Chapter 2 lists several regulatory systems in metazoans, identified from literature, and provides an attractive method for rapidly identifying and manipulating the specificity determinants in these key signaling pathways. Several possible directions for future improvements to the workflow, most notably an explicit account of desolvation energies, are proposed.

In chapters 3 and 4, methods for detailed characterization of protein – ligand/peptide/protein interactions are presented – the former based on MM-PBSA, the latter on TI. They can be used to aid drug or protein design (Childers and Daggett, 2017), particularly the more reliable TI calculations, albeit at a (what is presently perceived to be) great computational cost. However, what is considered computationally intensive today will likely be considered trivial a decade from now. Indeed, the original implementation of 4D QSAR, as first proposed by Hopfinger and coworkers in 1997 (Hopfinger et al., 1997) was receptor-independent, i.e. entirely focused on the ligands. One can be absolutely confident that this was due to the severe computational limitations of the time, not under-appreciation of the importance of the macromolecule. Moreover, one can be certain that the simulation lengths and system sizes (~30 000 atoms), reported herein, are something researchers could only long for at the time. What may take longer than a decade is the derivation and validation of novel 3D and 4D descriptors for the framework proposed herein. Indeed, there exist thousands (Duchowicz et al., 2006), potentially tens of thousands of zero-to-three-dimensional descriptors, yet understanding of intermolecular association and predictive capabilities remain limited, at best. Clearly, all those thousands of descriptors do not tell the whole story and there are gaps in current models.

It is then worth considering what course of action is needed to further extend the present framework. The obvious next steps would be to achieve convergence in simulations such as the ones reported in chapter 4, analyze the time-dependence of computed per-residue energies, characterize the processes leading up to the converged state, and then describe the converged state itself. Moreover, for computational efficiency, and perhaps from a theoretical point of view, it would be worthwhile investigating whether cheaper techniques, such as MM-PBSA, are capable of providing reliable per-residue energies in protein – ligand simulations, at a fraction of the computational cost. MM-PBSA simulations sample only the end states, which are physical or “chemical,” unlike thermodynamic integration, which simulates intermediate, unphysical or “alchemical” states. In a remarkable turn of events in the history of science, alchemical transformations are once again considered to be “higher-level” or more scientifically sound than chemical ones (for good reasons, of course, which have been extensively reviewed elsewhere (Kollman, 1993; Brandsdal et al., 2003; Hansen and van Gunsteren, 2014) and will not be discussed here). The “philosopher's stone” of modern-day computational chemistry and drug design, then, are the force fields and descriptors that will make hyperpredictive 4D-MD/QSAR models (Ash and Fourches, 2017) possible. If computational chemistry is to truly deliver on its promise of new and better drugs, a qualitative improvement in understanding of protein – ligand/peptide/protein interactions needs to occur. What this implies is rigorously validated force fields and (potentially, new) descriptors, not the constant increase in computing power one takes for granted. Per-residue energies are only one descriptor out of many more possible. For example, in more recent, receptor-dependent work, involving molecular dynamics simulations (RD-4D-QSAR), the Hopfinger group has successfully used grid cell occupancy at the interface as a descriptor (Pan et al., 2003; da Rocha Pita et al., 2012). One might expect that an analogous energetics-based descriptor might prove even richer in information content. This work furthers understanding of intermolecular association and, hopefully, will help bring about a more comprehensive model of it. From a practical perspective, in the short to mid-term, this would entail bringing together strong, specific interactions with weak, nonspecific ones into a common framework.

5.1. References

Ash, J. and Fourches, D., 2017. Characterizing the Chemical Space of ERK2 Kinase Using Descriptors Computed from Molecular Dynamics Trajectories. *Journal of Chemical Information and Modeling*, 57(6), pp.1286–1299.

Bogan, A.A. & Thorn, K.S., 1998. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1), pp.1–9.

Brandsdal, B.O. et al., 2003. Free Energy Calculations and Ligand Binding. *Advances in Protein Chemistry*, 66, pp.123–158.

Chakrabarti, P. and Janin, J., 2002. Dissecting protein-protein recognition sites. *Proteins: Structure, Function and Genetics*, 47(3), pp.334–343.

Chandler, D., 2005. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature*, 437(7059), pp.640–647.

Childers, M.C. et al., 2017. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.*, 2(1), pp.9–33.

Cukuroglu, E. et al., 2014. Hot spots in protein-protein interfaces: Towards drug discovery. *Progress in Biophysics and Molecular Biology*, 116(2–3), pp.165–173.

David, A. and Sternberg, M.J.E., 2015. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *Journal of Molecular Biology*, 427(17), pp.2886–2898.

Duchowicz, P.R., Castro, E.A. and Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Communications in Mathematical and in Computer Chemistry*, 55, pp.179–192.

Gromiha, M.M. et al., 2011. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome science*, 9(Suppl 1), p.S13.

Gromiha, M.M., Yokota, K. and Fukui, K., 2009. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Molecular bioSystems*, 5(12), pp.1779–1786.

Guharoy, M. & Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15447–52.

Guo, W., Wisniewski, J.A. & Ji, H., 2014. Hot spot-based design of small-molecule inhibitors for protein-protein interactions. *Bioorganic and Medicinal Chemistry Letters*, 24(11), pp.2546–2554.

Hansen, N. and van Gunsteren, W.F., 2014. Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, 10(7), pp.2632–2647.

Hopfinger, A.J. et al., 1997. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society*, 7863(5), pp.10509–10524.

Jones, S. and Thornton, J.M., 1996. Principles of protein-protein interactions. *Proceedings of the national Academy of sciences*, 93(1), pp.13–20.

Keskin, O. et al., 2008. Principles of Protein-Protein Interactions: What are the Preferred Ways for Proteins to Interact? *Chemical reviews*, 108(4), pp.1225–44.

Kim, J.-S. et al., 2015. Conversion of cell-survival activity of Akt into apoptotic death of cancer cells by two mutations on the BIM BH3 domain. *Cell death and disease*, 6, p.e1804.

Kollman, P., 1993. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chemical Reviews*, 93(7), pp.2395–2417.

Kosloff, M. et al., 2011. Integrating energy calculations with functional assays to decipher the specificity of G protein-RGS protein interactions. *Nature structural & molecular biology*, 18(7), pp.846–53.

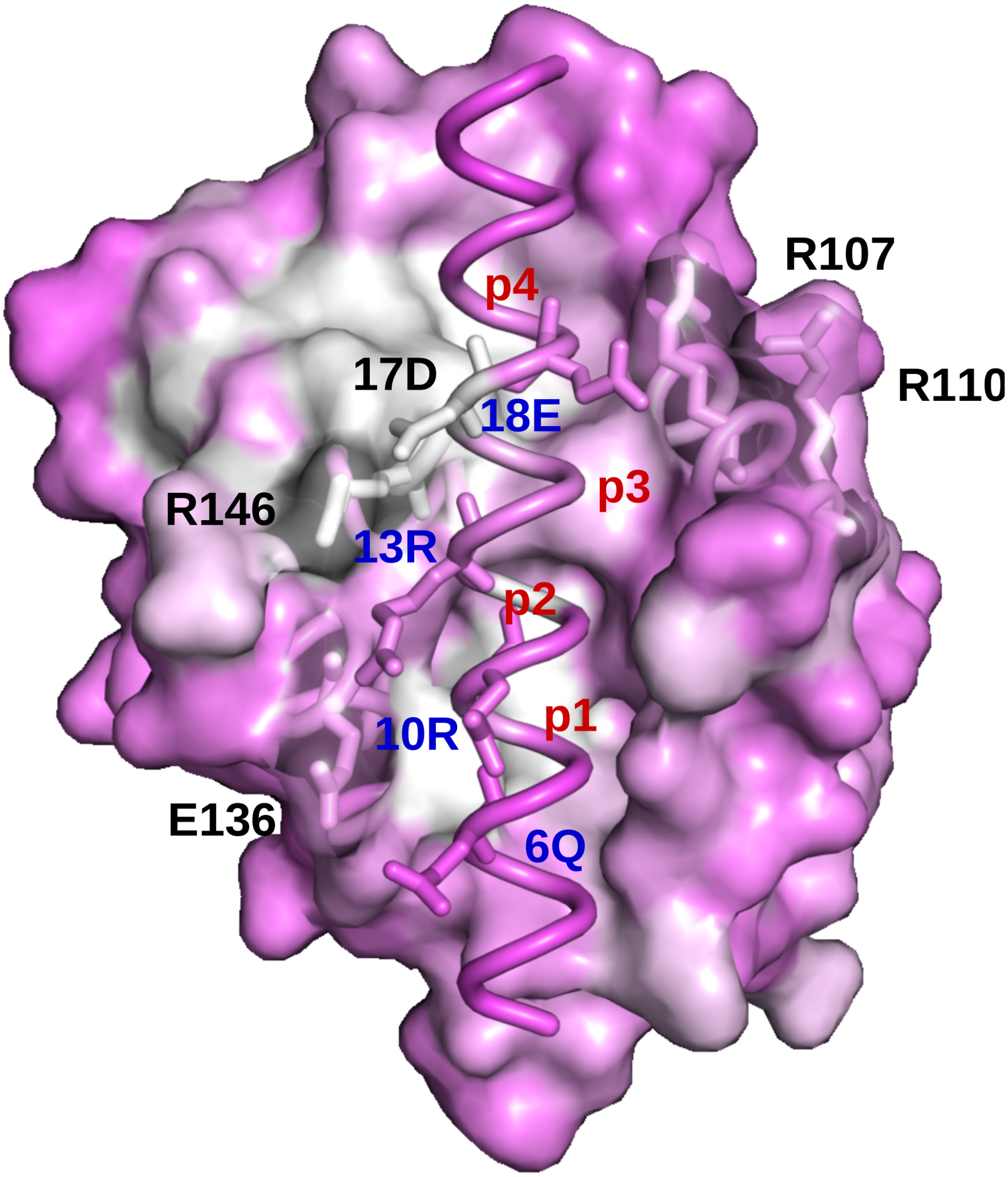
Lo Conte, L., Chothia, C. & Janin, J., 1999. The Atomic Structure of Protein-Protein Recognition Sites. *Journal of molecular biology*, 285(5), pp.2177–98.

Ma, B. et al., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp.5772–7.

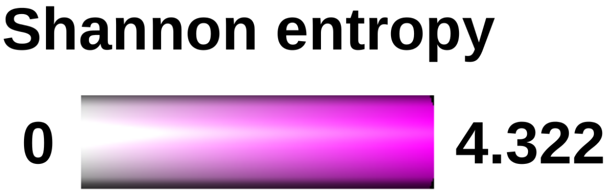
Pan, D., Tseng, Y. & Hopfinger, A.J., 2003. Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *Journal of Chemical Information and Computer Sciences*, 43(5), pp.1591–1607.

da Rocha Pita, S.S. et al., 2012. Receptor-Dependent 4D-QSAR Analysis of Peptidemimetic Inhibitors of Trypanosoma cruzi Trypanothione Reductase with Receptor-Based Alignment. *Chemical Biology and Drug Design*, 79(5), pp.740–748.

Appendix

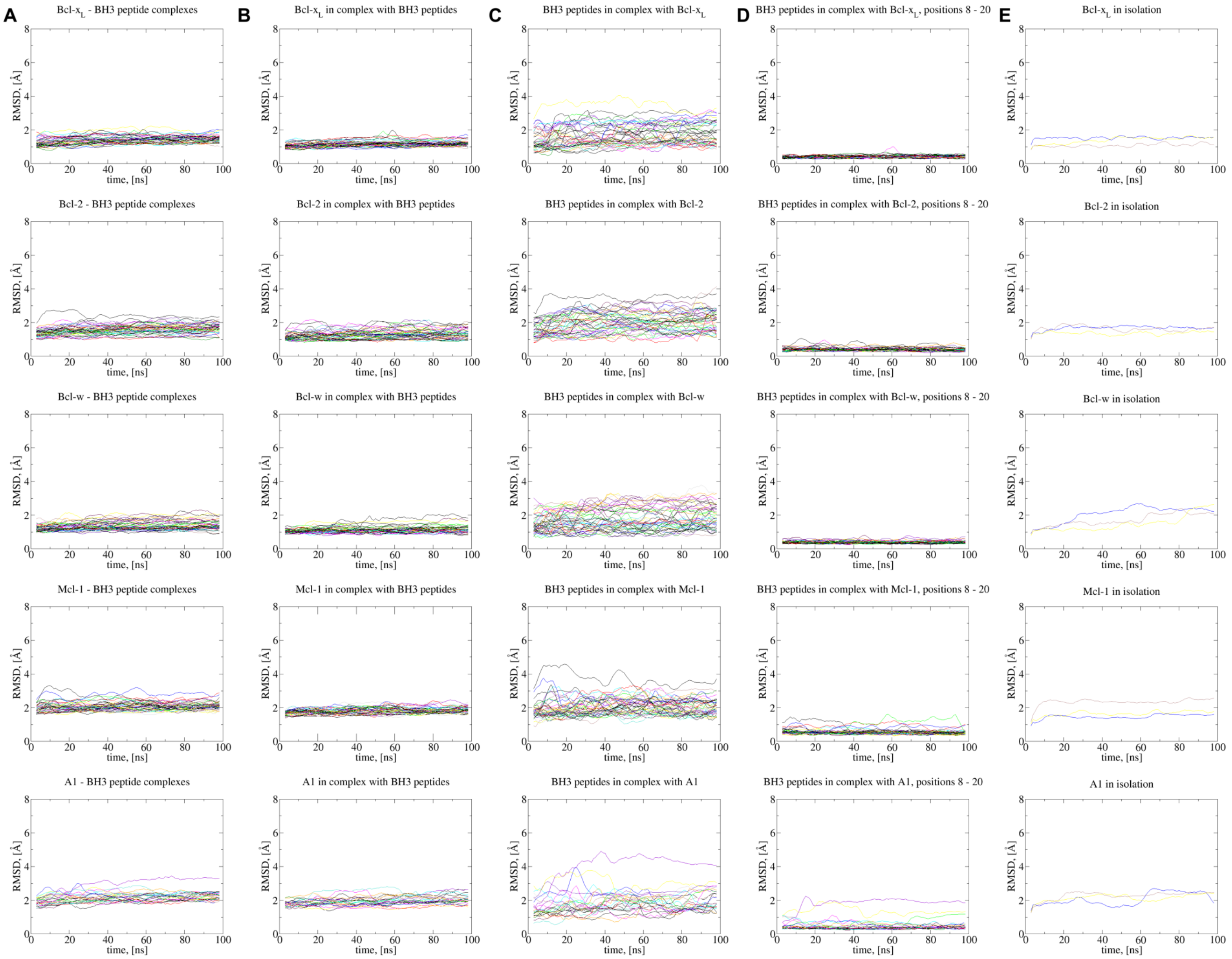


R107
p4
17D
R110
18E
p3
R146
13R
p2
p1
10R
E136
6Q



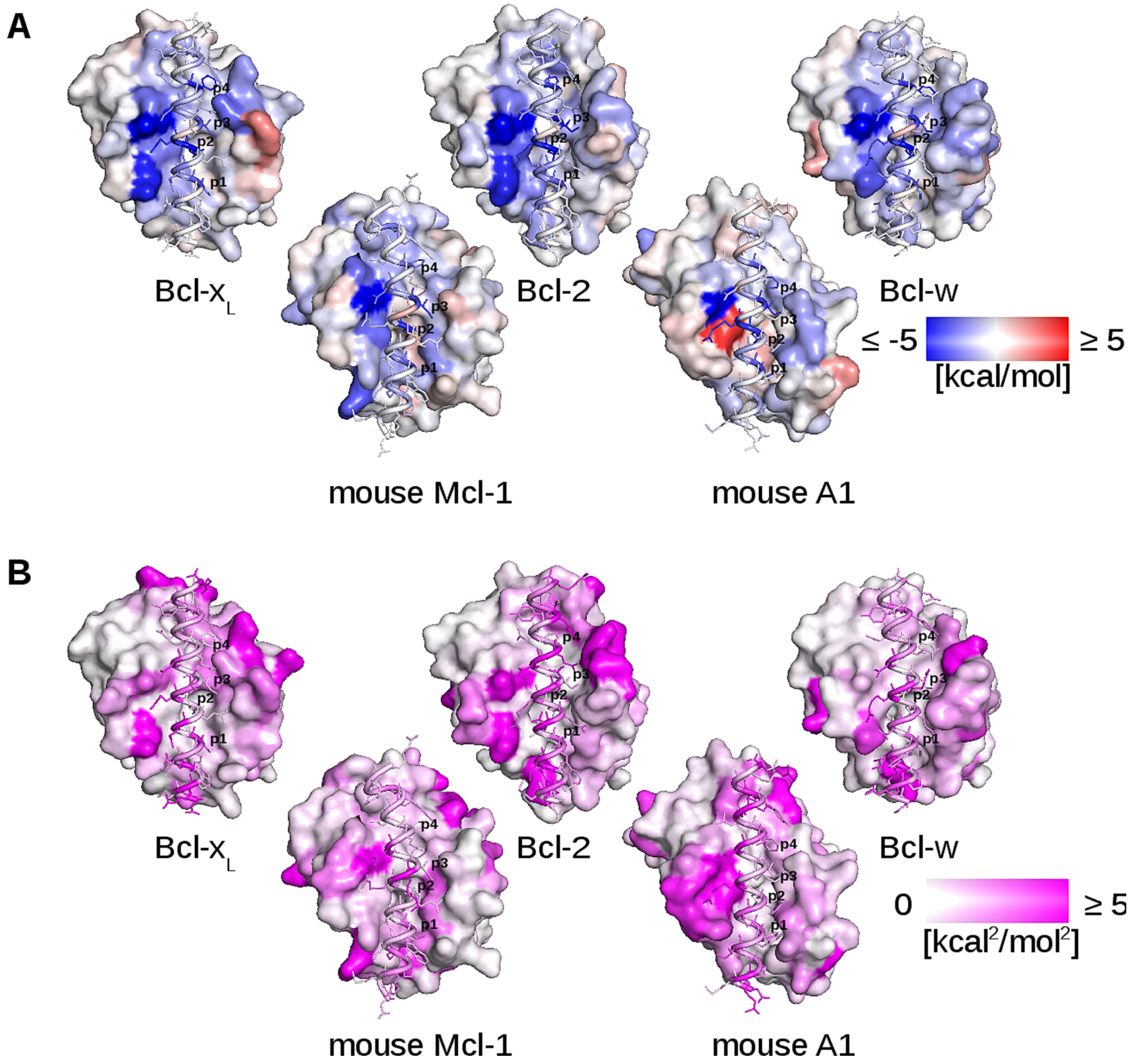
Shannon entropy
0
4.322

Supplementary information figure 2.1. The Bcl-2 – Bad complex colored by Shannon entropy. The Bad peptide is in cartoon representation with key residues labeled and in stick representation; the Bcl-2 protein is in surface representation, with key residues in stick and semitransparent surface representation. Also labeled are the positions of the four hydrophobic pockets (p1, p2, p3, p4) on the surface of the hydrophobic groove, the coloring scheme for the pockets and residues matches the coloring in the template structure and sequence alignment in Figure 2.1. Positions 8, 12, 15, and 19, colored red in the sequence alignment in Figure 2.1, are conserved hydrophobic residues, that fit into pockets 1, 2, 3, and 4, respectively.

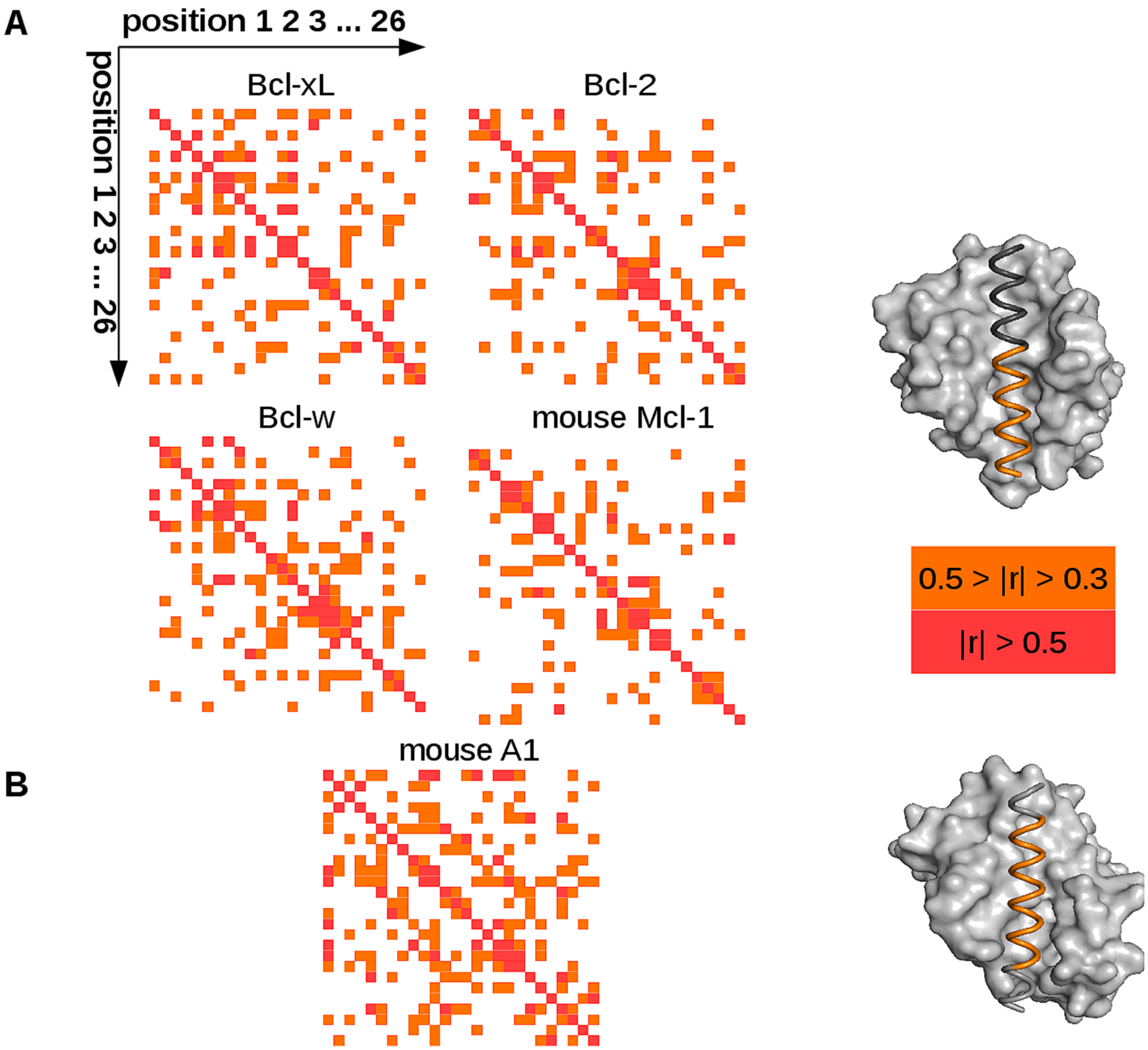


Supplementary information figure 3.1. Cα RMSD values during molecular dynamics simulations. **(A)** Cα RMSD values for the entire antiapoptotic protein – BH3 peptide complexes. Data shown are 5-ns running averages of the complex RMSD values for each trajectory of every complex, i.e. 39 trajectories for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 – BH3 peptide complexes, and 24 trajectories for the mouse A1 – BH3 peptide complexes. **(B)** Cα RMSD values for the antiapoptotic proteins. Data shown are 5-ns running averages of the receptor RMSD values for each trajectory of every complex, i.e. 39 trajectories for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 – BH3 peptide complexes, and 24 trajectories for the mouse A1 – BH3 peptide complexes. **(C)** Cα RMSD values for the BH3 peptides. Data shown are 5-ns running averages of the ligand RMSD values for each trajectory of every complex, i.e. 39 trajectories for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 – BH3 peptide complexes, and

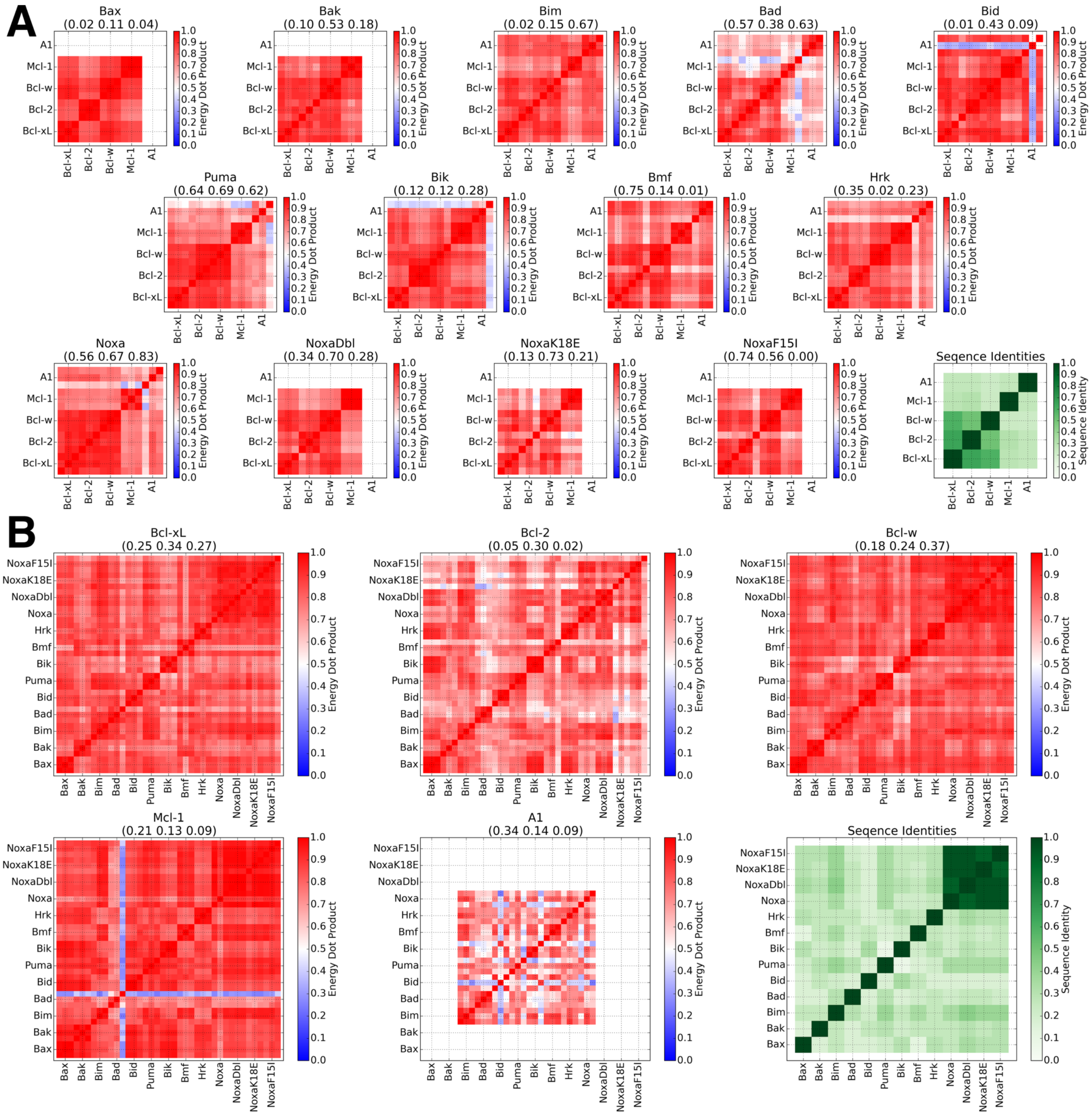
24 trajectories for the mouse A1 – BH3 peptide complexes. **(D)** Cα RMSD values for the core residues of the BH3 peptides (positions 8 - 20). Data shown are 5-ns running averages of the ligand core RMSD values for each trajectory of every complex, i.e. 39 trajectories for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1 – BH3 peptide complexes, and 24 trajectories for the mouse A1 – BH3 peptide complexes. **(E)** Cα RMSD values for the antiapoptotic proteins in isolation. Each protein was simulated in triplicate; data shown are 5-ns running averages.



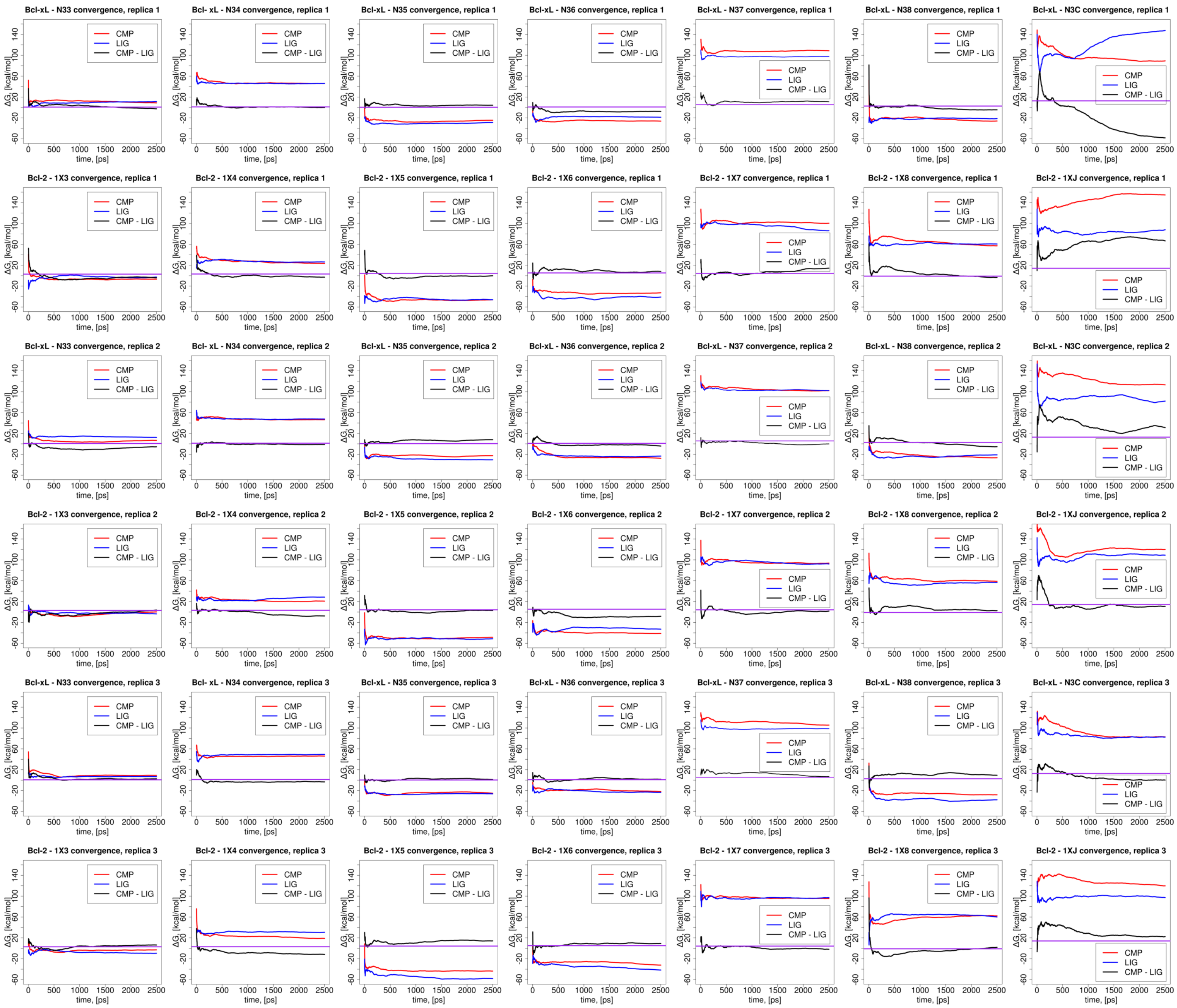
Supplementary information figure 3.2. Sources of affinity and specificity assessed via energetics analysis, based on trajectories of complex, receptor, and ligand. (A) Antiapoptotic protein – BH3 peptide complexes colored by average per-residue ΔH values. **(B)** Antiapoptotic protein – BH3 peptide complexes colored by the variance of per-residue ΔH values. Averages and variance were computed across 39-trajectory sets for Bcl-xL, Bcl-2, Bcl-w, and mouse Mcl-1, and across 24 trajectories for mouse A1. Ligand N-termini are at the bottom of the figures, C-termini are at the top. ΔH was calculated from complex, receptor, and ligand trajectories. See also Figure 3.2.



Supplementary information figure 3.3. Energy correlation analysis, based on trajectories of complex, receptor, and ligand. (A) Energy correlation analysis performed for the 26 ligand residues across 39-trajectory sets. BH3 ligands seem to display two regions of energetic correlation – an N-terminal one, spanning up to around position 15 (colored in orange in the structure to the right), and a C-terminal one (colored in gray). **(B)** Energy correlation analysis performed for the 26 ligand residues across the 24-trajectory set for A1. BH3 ligands seem to display an almost uninterrupted region of helix-like energetic correlation, spanning most of the peptide length (colored in orange in the structure to the right). ΔH was computed from complex, receptor, and ligand trajectories. See also Figure 3.4.



Supplementary information figure 3.4. Comparison of energy pattern similarity with sequence identity. (A) Similarity of complex energies grouped by ligand simulations as an inner product of energy pattern vectors. Sequence identities of the respective receptors are shown in the bottom right corner in green. Correlation coefficients for the patterns (separately for replicas 1, 2 and 3) are shown in the plot titles. **(B)** The same data is shown but grouped by receptor. The sequence identity of the ligands is shown in green. Note that NoxaDbl stands for the Noxa double mutant (F15I, K18E).



Bcl-xL - N33 convergence, replica 1

Bcl- xL - N34 convergence, replica 1

Bcl-xL - N35 convergence, replica 1

Bcl-xL - N36 convergence, replica 1

Bcl-xL - N37 convergence, replica 1

Bcl-xL - N38 convergence, replica 1

Bcl-xL - N3C convergence, replica 1

Bcl-2 - 1X3 convergence, replica 1

Bcl-2 - 1X4 convergence, replica 1

Bcl-2 - 1X5 convergence, replica 1

Bcl-2 - 1X6 convergence, replica 1

Bcl-2 - 1X7 convergence, replica 1

Bcl-2 - 1X8 convergence, replica 1

Bcl-2 - 1XJ convergence, replica 1

Bcl-xL - N33 convergence, replica 2

Bcl- xL - N34 convergence, replica 2

Bcl-xL - N35 convergence, replica 2

Bcl-xL - N36 convergence, replica 2

Bcl-xL - N37 convergence, replica 2

Bcl-xL - N38 convergence, replica 2

Bcl-xL - N3C convergence, replica 2

Bcl-2 - 1X3 convergence, replica 2

Bcl-2 - 1X4 convergence, replica 2

Bcl-2 - 1X5 convergence, replica 2

Bcl-2 - 1X6 convergence, replica 2

Bcl-2 - 1X7 convergence, replica 2

Bcl-2 - 1X8 convergence, replica 2

Bcl-2 - 1XJ convergence, replica 2

Bcl-xL - N33 convergence, replica 3

Bcl- xL - N34 convergence, replica 3

Bcl-xL - N35 convergence, replica 3

Bcl-xL - N36 convergence, replica 3

Bcl-xL - N37 convergence, replica 3

Bcl-xL - N38 convergence, replica 3

Bcl-xL - N3C convergence, replica 3

Bcl-2 - 1X3 convergence, replica 3

Bcl-2 - 1X4 convergence, replica 3

Bcl-2 - 1X5 convergence, replica 3

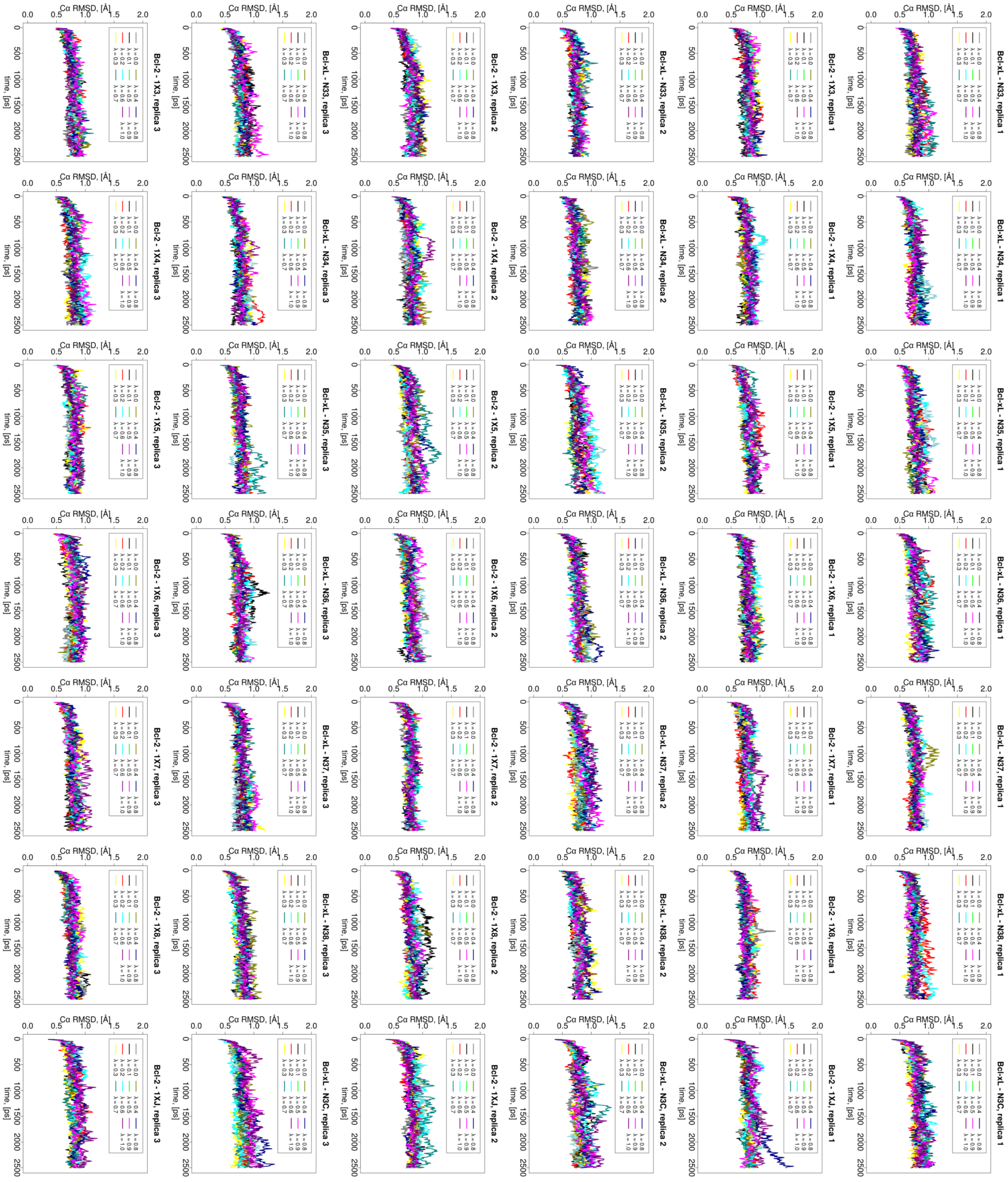
Bcl-2 - 1X6 convergence, replica 3

Bcl-2 - 1X7 convergence, replica 3

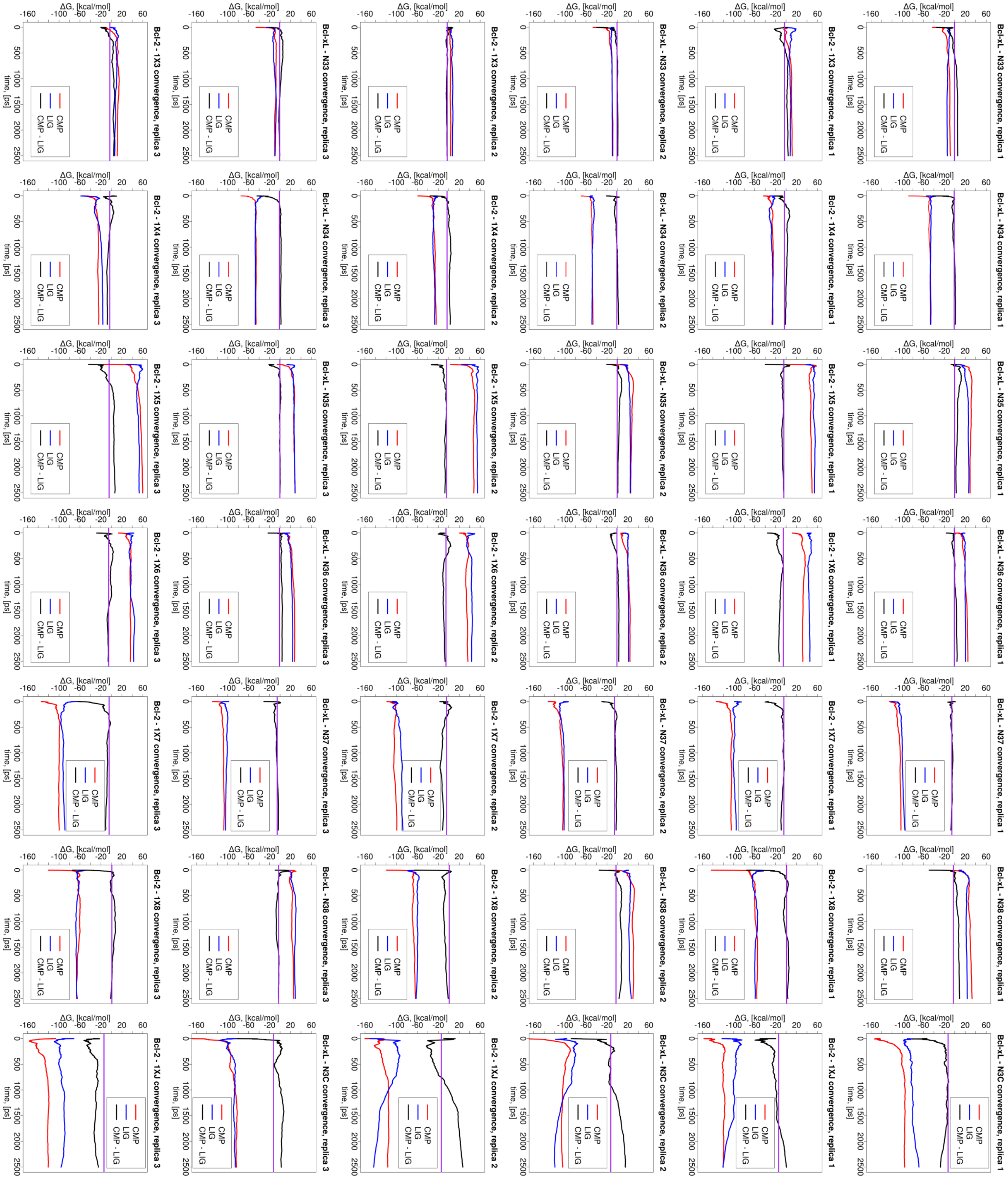
Bcl-2 - 1X8 convergence, replica 3

Bcl-2 - 1XJ convergence, replica 3

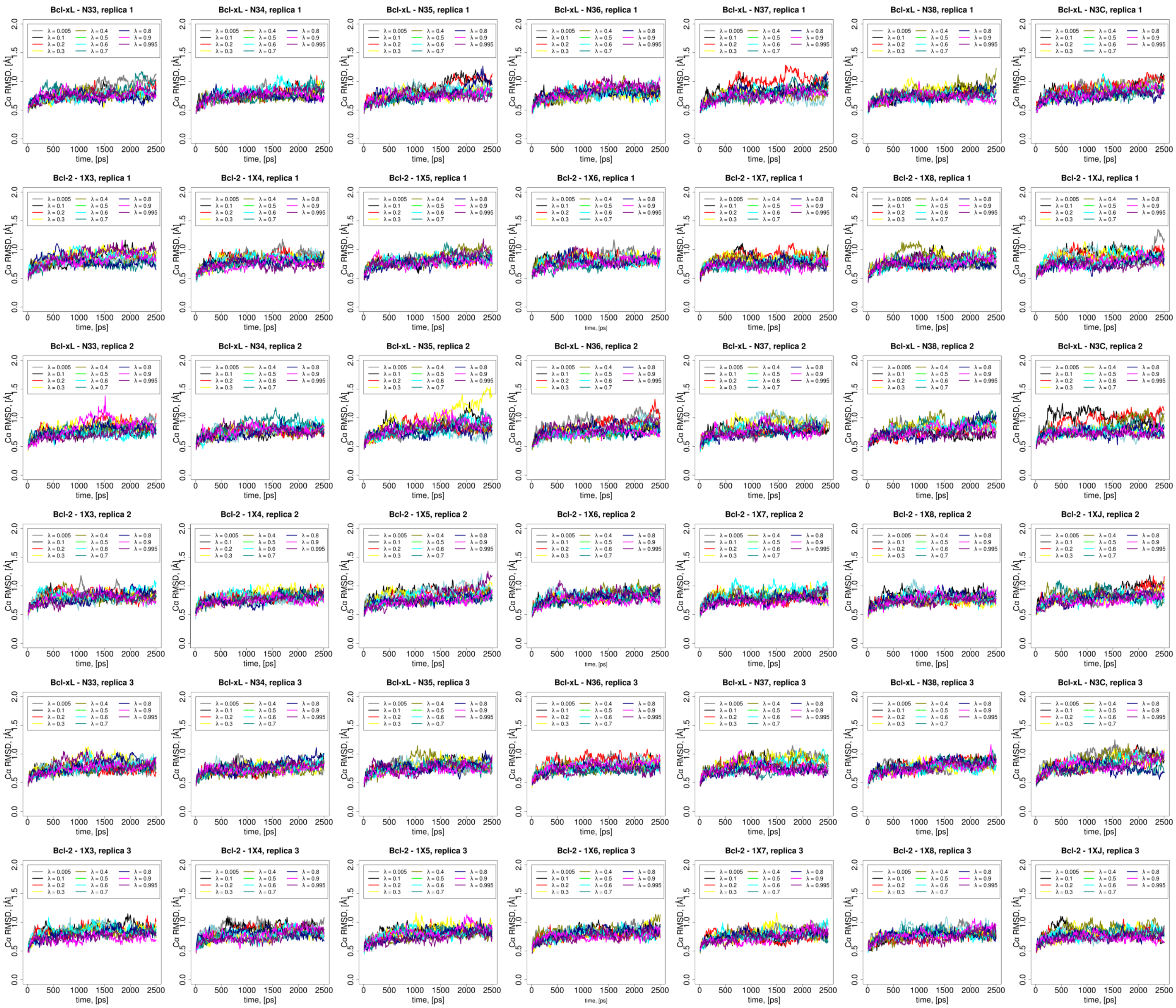
Supplementary information figure 4.1. Free energy convergence. Free energy convergence for the three *pmemd* replicas for the six model and N3C and 1XJ template compounds. “CMP” designates the ΔG of transforming the template ligand into the model ligand in complex with the protein, “LIG” designates the transformation in solution, “CMP – LIG” designates the calculated relative binding free energy (ΔΔG), obtained by subtracting the LIG ΔG from the CMP ΔG. For the template ligand simulations, “CMP” is the ΔG of decoupling the ligand from the hydrophobic groove of the protein, “LIG” is the ΔG of decoupling the ligand free in solution from the computational box, and “CMP – LIG” is the free energy of binding to the proteins. The purple horizontal lines indicate the experimental free energy values. As the affinities of compounds 5 and 8 for Bcl-xL and Bcl-2 were beyond the measuring capabilities of the experimental setup (K_i < 1 nM (Bruncko et al., 2007), K_i < 0.01 nM (Souers et al., 2013), respectively), provisional ΔΔGs for those compounds were computed using values of 1 and 0.01 nM, respectively. Compounds 3 – 8 are prefixed with “N3” and “1X” to designate transformations from the N3C and 1XJ templates, respectively; each λ-window was simulated for 2.5 ns of production dynamics.


Bcl-xL - N33, replica 1
Bcl-xL - N34, replica 1
Bcl-xL - N35, replica 1
Bcl-xL - N36, replica 1
Bcl-xL - N37, replica 1
Bcl-xL - N38, replica 1
Bcl-xL - N3C, replica 1
Bcl-2 - 1X3, replica 1
Bcl-2 - 1X4, replica 1
Bcl-2 - 1X5, replica 1
Bcl-2 - 1X6, replica 1
Bcl-2 - 1X7, replica 1
Bcl-2 - 1X8, replica 1
Bcl-2 - 1XJ, replica 1
Bcl-xL - N33, replica 2
Bcl-xL - N34, replica 2
Bcl-xL - N35, replica 2
Bcl-xL - N36, replica 2
Bcl-xL - N37, replica 2
Bcl-xL - N38, replica 2
Bcl-xL - N3C, replica 2
Bcl-2 - 1X3, replica 2
Bcl-2 - 1X4, replica 2
Bcl-2 - 1X5, replica 2
Bcl-2 - 1X6, replica 2
Bcl-2 - 1X7, replica 2
Bcl-2 - 1X8, replica 2
Bcl-2 - 1XJ, replica 2
Bcl-xL - N33, replica 3
Bcl-xL - N34, replica 3
Bcl-xL - N35, replica 3
Bcl-xL - N36, replica 3
Bcl-xL - N37, replica 3
Bcl-xL - N38, replica 3
Bcl-xL - N3C, replica 3
Bcl-2 - 1X3, replica 3
Bcl-2 - 1X4, replica 3
Bcl-2 - 1X5, replica 3
Bcl-2 - 1X6, replica 3
Bcl-2 - 1X7, replica 3
Bcl-2 - 1X8, replica 3
Bcl-2 - 1XJ, replica 3
Cα RMSD, [Å]
0.0 0.5 1.0 1.5 2.0
time, [ps]
0 500 1000 1500 2000 2500
λ = 0.0
λ = 0.1
λ = 0.2
λ = 0.3
λ = 0.4
λ = 0.5
λ = 0.6
λ = 0.7
λ = 0.8
λ = 0.9
λ = 1.0

Supplementary information figure 4.2. Cα root-mean-square-deviations. Cα RMSDs for each λ-window in all three replicas of the *pmemd* transformations. Compounds 3 – 8 are prefixed with “N3” and “1X” to designate transformations from the N3C and 1XJ templates, respectively; each λ-window was simulated for 2.5 ns of production dynamics.


Bcl-xL - N33 convergence, replica 1
Bcl-xL - N34 convergence, replica 1
Bcl-xL - N35 convergence, replica 1
Bcl-xL - N36 convergence, replica 1
Bcl-xL - N37 convergence, replica 1
Bcl-xL - N38 convergence, replica 1
Bcl-xL - N3C convergence, replica 1
Bcl-2 - 1X3 convergence, replica 1
Bcl-2 - 1X4 convergence, replica 1
Bcl-2 - 1X5 convergence, replica 1
Bcl-2 - 1X6 convergence, replica 1
Bcl-2 - 1X7 convergence, replica 1
Bcl-2 - 1X8 convergence, replica 1
Bcl-2 - 1XJ convergence, replica 1
Bcl-xL - N33 convergence, replica 2
Bcl-xL - N34 convergence, replica 2
Bcl-xL - N35 convergence, replica 2
Bcl-xL - N36 convergence, replica 2
Bcl-xL - N37 convergence, replica 2
Bcl-xL - N38 convergence, replica 2
Bcl-xL - N3C convergence, replica 2
Bcl-2 - 1X3 convergence, replica 2
Bcl-2 - 1X4 convergence, replica 2
Bcl-2 - 1X5 convergence, replica 2
Bcl-2 - 1X6 convergence, replica 2
Bcl-2 - 1X7 convergence, replica 2
Bcl-2 - 1X8 convergence, replica 2
Bcl-2 - 1XJ convergence, replica 2
Bcl-xL - N33 convergence, replica 3
Bcl-xL - N34 convergence, replica 3
Bcl-xL - N35 convergence, replica 3
Bcl-xL - N36 convergence, replica 3
Bcl-xL - N37 convergence, replica 3
Bcl-xL - N38 convergence, replica 3
Bcl-xL - N3C convergence, replica 3
Bcl-2 - 1X3 convergence, replica 3
Bcl-2 - 1X4 convergence, replica 3
Bcl-2 - 1X5 convergence, replica 3
Bcl-2 - 1X6 convergence, replica 3
Bcl-2 - 1X7 convergence, replica 3
Bcl-2 - 1X8 convergence, replica 3
Bcl-2 - 1XJ convergence, replica 3
ΔG, [kcal/mol]
-160 -100 -60 -20 20 60
time, [ps]
0 500 1000 1500 2000 2500
CMP
LIG
CMP - LIG

Supplementary information figure 4.3. Free energy convergence. Free energy convergence for the three *sander* replicas for the six model and N3C and 1XJ template compounds. “CMP” designates the ΔG of transforming the model ligand ligand into the template ligand in complex with the protein, “LIG” designates the transformation in solution, “CMP – LIG” designates the calculated relative binding free energy (ΔΔG), obtained by subtracting the LIG ΔG from the CMP ΔG. For the template ligand simulations, “CMP” is the ΔG of inserting the ligand into the hydrophobic groove of the protein, “LIG” is the ΔG of inserting the ligand free in solution, and “CMP – LIG” is the negative of the free energy of binding to the proteins. The purple horizontal lines indicate the experimental free energy values. As the affinities of compounds 5 and 8 for Bcl-xL and Bcl-2 were beyond the measuring capabilities of the experimental setup (K_i < 1 nM (Bruncko et al. 2007), K_i < 0.01 nM (Souers et al. 2013), respectively), provisional ΔΔGs for those compounds were computed using values of 1 and 0.01 nM, respectively. Compounds 3 – 8 are prefixed with “N3” and “1X” to designate transformations to the N3C and 1XJ templates, respectively; each λ-window was simulated for 2.5 ns of production dynamics.



Bcl-xL - N33, replica 1
Bcl-xL - N34, replica 1
Bcl-xL - N35, replica 1
Bcl-xL - N36, replica 1
Bcl-xL - N37, replica 1
Bcl-xL - N38, replica 1
Bcl-xL - N3C, replica 1

Bcl-2 - 1X3, replica 1
Bcl-2 - 1X4, replica 1
Bcl-2 - 1X5, replica 1
Bcl-2 - 1X6, replica 1
Bcl-2 - 1X7, replica 1
Bcl-2 - 1X8, replica 1
Bcl-2 - 1XJ, replica 1

Bcl-xL - N33, replica 2
Bcl-xL - N34, replica 2
Bcl-xL - N35, replica 2
Bcl-xL - N36, replica 2
Bcl-xL - N37, replica 2
Bcl-xL - N38, replica 2
Bcl-xL - N3C, replica 2

Bcl-2 - 1X3, replica 2
Bcl-2 - 1X4, replica 2
Bcl-2 - 1X5, replica 2
Bcl-2 - 1X6, replica 2
Bcl-2 - 1X7, replica 2
Bcl-2 - 1X8, replica 2
Bcl-2 - 1XJ, replica 2

Bcl-xL - N33, replica 3
Bcl-xL - N34, replica 3
Bcl-xL - N35, replica 3
Bcl-xL - N36, replica 3
Bcl-xL - N37, replica 3
Bcl-xL - N38, replica 3
Bcl-xL - N3C, replica 3

Bcl-2 - 1X3, replica 3
Bcl-2 - 1X4, replica 3
Bcl-2 - 1X5, replica 3
Bcl-2 - 1X6, replica 3
Bcl-2 - 1X7, replica 3
Bcl-2 - 1X8, replica 3
Bcl-2 - 1XJ, replica 3

Supplementary information figure 4.4. Cα root-mean-square-deviations. Cα RMSDs for each λ-window in all three replicas of the *sander* transformations. Compounds 3 – 8 are prefixed with "N3" and "1X" to designate transformations to the N3C and 1XJ templates, respectively; each λ-window was simulated for 2.5 ns of production dynamics.