

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Regulation by phosphorylation from an atomic perspective

**Permalink**

<https://escholarship.org/uc/item/3ch433zc>

**Author**

Narayanan, Arjun

**Publication Date**

2009

Peer reviewed|Thesis/dissertation

# Regulation by phosphorylation from an atomic perspective

by

Arjun Narayanan

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2008  
by  
Arjun Narayanan

# Acknowledgements

First the required content:

The text of Chapter 2 of this dissertation is a revised reprint of the material as it appears in PLoS Computational Biology [1], in which Eli S. Groban was a co-first author. The text of Chapter 3 of this dissertation is a revised reprint of the material as it appears in Journal of Molecular Biology [2], in which Benjamin D. Sellers was a co-author. Portions of the text of Chapters 1 and 5 of this dissertation are revised reprints of the material as it appears in Current Opinion in Structural Biology [3]. The co-author Matthew P. Jacobson in the above publications directed and supervised the research that forms the basis for the dissertation.

I would like to thank Matt Jacobson for being an excellent advisor who always looked out for my interests. I have learned a tremendous amount from him, not only about issues in computational biophysics, but also about the scientific process more generally. I would also like to thank Robert Fletcher and Ken Dill for providing very helpful advice in planning for my future as well as helpful and interesting scientific discussions.

Finally, I would like to thank my mother and father, who made sure I took my education seriously even when I wanted to drive garbage trucks or become a professional basketball player, and who supported me tremendously - instructionally, financially, and emotionally. I would like to thank my little sister for being awesome and fun and helping me see the lighter side of things. And last but not least, I would like to thank Elisabeth

Humphris for her support, her perspective, and the happiness and excitement that being with her brings me.

## Regulation by phosphorylation from an atomic perspective

Arjun Narayanan

### Abstract:

Phosphorylation is essential for mediating switch-like changes in protein function and cellular behavior. In eukaryotes, protein kinases catalyze phosphorylation of the hydroxyl groups of serine, threonine, and tyrosine, adding a -2 charge to otherwise neutral residues. This large electrostatic perturbation can alter protein structure, leading to switch-like changes in protein function. While mass spectrometry has identified many thousands of phosphorylation sites, few have been functionally characterized, and the structural effects of phosphorylation have been elucidated for even fewer. It is particularly difficult to obtain atomic-level resolution of structural changes due to phosphorylation. The use of atomic-level computational methods is an attractive candidate to augment low-resolution structural and functional studies in probing phosphorylation-induced conformational changes.

This thesis presents work using all-atom computational techniques to study phosphorylation-induced conformational changes. We studied phosphorylation-induced conformational changes in protein loops and helices for several diverse proteins. We present a case study examining the structural effects of activation loop phosphorylation in cyclin-dependent kinase 2 and the underlying energetic basis, highlighting the important balance between electrostatics and desolvation. From observations of phosphorylation-induced changes in the orientation of protein kinase lobes, we devised the beginnings of a method for the prediction of the orientation of protein domains, using antibodies as a

model system. This method involved extensive sampling of side-chain conformations at the interface of the antibody heavy and light chain variable domains, and was successful at reconstructing existing antibody crystal structures. In addition, this energy-based method for the prediction of variable domain orientation, when coupled with a filter based on sequence identity, was more successful than using sequence identity alone when predicting the variable domain orientation in antibody homology models. Finally, we used all-atom molecular dynamics simulations to prospectively study conformational changes in the Arp2/3 complex upon phosphorylation of the Arp2 subunit. We find that Arp2 phosphorylation induces changes in the interactions in the highly charged Arp2-Arp3-ARPC4 interface, leading to a breakage of contacts between Arp2 and Arp3. We posit a model by which phosphorylation relieves the auto-inhibitory interactions at this interface, leading to activating conformational changes that allow full activation upon binding of other co-activators.

# Table of Contents

<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1. Protein basics .....	1
1.2. Protein regulation by phosphorylation.....	2
1.3. Experimental methods for studying phosphorylation-induced structural changes.....	3
1.4. Computational methods for studying phosphorylation-induced structural changes.....	3
1.5. Preview .....	15
<b>Chapter 2 Conformational Changes in Protein Loops and Helices Induced by Post-Translational Phosphorylation .....</b>	<b>16</b>
2.1. Introduction.....	16
2.2. Results and Discussion .....	19
2.3. Conclusion .....	40
2.4. Methods.....	40
<b>Chapter 3 Energy-based analysis and prediction of the orientation between antibody light and heavy variable domains.....</b>	<b>47</b>
3.1. Introduction.....	48
3.2. Results.....	51
3.3. Discussion.....	67
3.4. Methods.....	69
<b>Chapter 4 Phosphorylation-induced conformational changes of the Arp2/3 complex</b>	<b>76</b>
4.1. Introduction.....	76
4.2. Results.....	79
4.3. Discussion.....	94
4.4. Methods.....	97
<b>Chapter 5 Future Directions .....</b>	<b>100</b>
5.1. Scoring and sampling.....	100
5.2. Phosphoregulation of ion channels .....	101
5.3. Lysine methylation.....	106
<b>Appendix A. Supplementary Tables and Figures for Chapter 2.....</b>	<b>131</b>
<b>Appendix B. Supplementary Tables and Figures for Chapter 3.....</b>	<b>134</b>
<b>Appendix C. Supplementary Figures for Chapter 4.....</b>	<b>145</b>



# List of Tables

Table 2.1 Results from the CDK2 Case Study .....	24
Table 2.2 Energy Differences (kcal/mol) between “Active” and “Inactive” Conformations of the Activation Loop .....	32
Table 2.3 Proteins Used in Our Test Set.....	34
Table 2.4 Reconstruction of Phosphorylated Loop Conformations.....	36
Table 2.5 Prediction of Phosphorylated Loop Conformations .....	37
Table 3.1 The differences in orientation between antibodies are more pronounced at lower sequence identities. ....	56
Table 3.2 The crystallographic orientation of light and heavy chain antibody variable domains is defined by the interface sequence and is at or near the global minimum of the energy landscape.....	61
Table 3.3 Predicted Orientations After Building Comparative Models From Multiple Templates.....	65

# List of Figures

Figure 2.1 Phosphorylation Can Perturb the Energy Landscape of a Protein to Cause Changes in Conformation and Dynamics .....	18
Figure 2.2 Example of the Hierarchical Loop Prediction, Applied to Reconstructing the Phosphorylated Activation Loop of CDK2/Cyclin A.....	23
Figure 2.3 Loop Reconstruction and Prediction in CDK2/Cyclin A .....	24
Figure 2.4 Detailed View of One Portion of the Structural Superposition between the Phosphorylated and Unphosphorylated Crystal Structures of CDK2/Cyclin A .....	25
Figure 2.5 CDK2 Case Study.....	27
Figure 2.6 Differences in Conformational Predictions for Phosphorylated CDK2 with and without Cyclin A Bound .....	29
Figure 2.7 Active and Inactive Conformations of the CDK2 Activation Loop.....	32
Figure 2.8 Helix Reconstruction and Prediction of FixJ.....	39
Figure 3.1 Distribution of differences in domain orientation .....	53
Figure 3.2 Mean difference in domain orientation is correlated with simple measures of difference in sequence, but there is large variability at a given identity or similarity. ....	55
Figure 3.3 Near-native orientations of the light and heavy chain variable domains represent global energy minima using a molecular mechanics scoring function. ....	59
Figure 3.4 Method for predicting $V_L:V_H$ orientation in antibody comparative modeling	65
Figure 4.1 Principal component 1 differentiates between unphosphorylated and phosphorylated simulations .....	80
Figure 4.2 First principal component characterized by large subunit rearrangements of Arp2, ARPC1, and ARPC3.....	81
Figure 4.3 Arp2 phosphorylation sites Thr237 and Thr238 .....	83
Figure 4.4 Interactions in the vicinity of Arp2 Thr237 in simulations .....	85
Figure 4.5 Changes in electrostatic interactions upon phosphorylation of Arp2 Thr237 or Thr238.....	86
Figure 4.6 Overall structural changes of mutant complexes.....	88

Figure 4.7 Changes in electrostatic interactions in the Arp2 Thr237Glu/Thr238Glu double mutant simulations .....	89
Figure 4.8 Changes in electrostatic interactions in the unphosphorylated ARPC4 Arg105Ala mutant simulations .....	91
Figure 4.9 Changes in electrostatic interactions in the Arp2 Thr237 phosphorylated ARPC4 Arg105Ala mutant simulations.....	92
Figure 4.10 Loss of Arp2-Arp3 contacts upon phosphorylation and mutation .....	93

# Chapter 1 Introduction

The text of this introduction was, in large part, taken from a review article published in *Current Opinion in Structural Biology* (A. Narayanan and M.P. Jacobson (2009) **19**: 156-63), and to a lesser extent, from work published in *PLoS Computational Biology* (E.S. Groban\*, A. Narayanan\*, M.P. Jacobson (2006) **2**: e32). Matt Jacobson directed and co-wrote both publications, while Eli Groban co-wrote the latter.

## 1.1. Protein basics

Proteins make up a major class of biological macromolecules that are involved in the execution of cellular functions. A protein is composed of a covalently-linked chain of amino acids that is synthesized from messenger ribonucleic acid molecules by the ribosome in a process known as translation. This process allows the incorporation of 20 amino acids unique in their chemical composition and physical properties. These 20 amino acids provide sufficient variety to allow for proteins to fold into specific 3-dimensional structures, perform enzymatic catalysis, and bind to small molecule metabolites and other proteins.

## 1.2. Protein regulation by phosphorylation

Protein function is regulated at several levels in order to control cellular growth, division, and response to external stimuli. One important mechanism of protein regulation is by post-translational chemical modification, essentially expanding the 20 amino-acid alphabet through, for example, phosphorylation and sulfonation of serine (Ser), threonine (Thr), and tyrosine (Tyr), and acetylation and methylation of lysine (Lys) and arginine (Arg). Since the addition and removal of a modifying group are enzyme-catalyzed, the cell can finely tune the magnitude and duration of the regulatory effect and can construct complex molecular switches with rapid temporal responses.

Phosphorylation is a ubiquitous mechanism of post-translational modification. It is estimated that the fraction of proteins that are phosphorylated *in vivo* may be up to 30% [4]. In eukaryotes, the phosphate group is added to the hydroxyl groups of Ser, Thr, and Tyr side chains by protein kinases and removed by phosphatases. Phosphorylation serves as an important mode of information transfer for numerous signaling pathways in the cell. The phosphate predominantly carries a -2 charge at physiologically relevant pH, and the resulting large electrostatic perturbation modulates the energy landscapes governing protein folding, protein-protein and protein-ligand interactions, catalytic activity, and conformational dynamics. Consequently, phosphorylation can serve as a packet of information effectively dictating whether or not two proteins should associate (e.g., p53-HDM2 [5], pKID/KIX [6,7]), or whether or not a protein should be catalytically active (many protein kinases, reviewed in [8,9,10]). The ability of phosphorylation to induce binary-like changes to protein behavior is a crucial element to the induction of switch-

like behavioral changes at the cellular level. Numerous pathologies result from aberrant regulation of these electrostatic switches, including several cancers [11,12].

### 1.3. Experimental methods for studying phosphorylation-induced structural changes

Mass-spectrometric studies have identified many thousands of sites of post-translational phosphorylation [4]. However, obtaining an atomic-level understanding of how phosphorylation alters protein structure and function is difficult experimentally. X-ray crystallography and NMR provide the highest resolution structural information but are also the most difficult and time-consuming to obtain. A particular challenge can be obtaining sufficient quantities of pure protein with a particular post-translational modification. Electron paramagnetic resonance, circular dichroism, Förster resonance energy transfer (FRET), and small-angle X-ray scattering all can provide low-resolution information about structural changes due to phosphorylation but do not provide atomic detail.

### 1.4. Computational methods for studying phosphorylation-induced structural changes

Given the wide gap between the number of known post-translational modifications and the number that have been structurally or functionally characterized, we believe that computational methods can play a valuable role in elucidating principles and consequences of post-translational phosphorylation. This requires the accurate scoring and adequate sampling of relevant conformations, which will each be briefly

discussed below. This is followed by a review of previous applications of computation to study phosphorylation-induced conformational changes.

#### ***1.4.1. Scoring phosphorylation-induced conformational changes***

Phosphorylation of Ser, Thr, and Tyr residues change the charge of these previously neutral amino acids to carry a -2 charge at physiological pH localized predominantly on the phosphate group. The interactions most affected by this alteration are the Coulombic electrostatic interactions and the interactions with solvent. For example, the -2 charged phosphorylated residues will exhibit both a much more favorable Coulombic interaction with positively charged Arg and Lys residues and a much more favorable solvation free energy with the high dielectric aqueous solvent than will their unphosphorylated counterparts. These increased propensities for interactions with oppositely charged species and with solvent molecules are often at odds with each other – the formation of strong Coulombic interactions with other residues necessitating the displacement of favorable interactions with solvent. Consequently, correctly balancing Coulombic and solvent interactions is imperative.

Most computational studies of protein conformation involve the use of molecular mechanics energy functions such as the OPLS, AMBER, and CHARMM force fields that amongst other approximations, involve the use of fixed point-charges to describe the charge on each atom in calculating Coulombic interaction energies. This description neglects the effects of electronic polarizability; as the presence of a -2 charge could significantly affect the electronic structure and the charge “on” each atom, this could lead to substantially incorrect estimates of the Coulombic interaction energies involving phosphorylated species.

Calculating free energies of interaction with solvent quantitatively is an ongoing problem in computational chemistry. Solvent can be represented either explicitly through fixed-charge water models such as TIP3P [13], or implicitly through the use of models based on continuum electrostatics [14,15,16]. While treating solvent explicitly is computationally more demanding due to the need to average over water conformations (and still suffers from the fixed-charge approximation), it captures first-shell water effects, such as the energetic minimum observed when a water molecule bridges between ionic species. Alternatively, implicit solvent models treat the solvent as a continuum dielectric environment, without consideration of the size, number, or orientation of individual water molecules. Consequently, first-shell water effects are not captured, but the dielectric response to the point-charges of the protein atoms are captured in an averaged way. Because they implicitly average over all solvent molecule degrees-of-freedom, they are much faster at computing the solvation free energy of protein conformations, though the calculated values are not considered as accurate as those achieved by using explicit solvent models with sufficient sampling.

Work from our lab [17,18] (as well as the work appearing in Chapter 1 of this thesis) and others [19] has outlined the accuracy of calculating energies of different species with models of phosphorylated amino acids. Mandell, et al. compared explicit solvent molecular mechanics with implicit solvent quantum mechanical and molecular mechanical calculations of methyl phosphate with propyl guanidinium and butyl ammonium, analogs of Arg and Lys side chains, respectively [17]. Relative to the explicit solvent molecular mechanics calculations, the implicit solvent molecular mechanics calculations consistently over-estimated the free energy minimum for direct



contact between the phosphate group and the Arg and Lys analogs, although in both cases bidentate hydrogen-bonding interactions with the guanidinium were substantially more favored. The quantum mechanical calculations with implicit solvent showed a substantial increase in the favorability of hydrogen bonds between the phosphate and Arg or Lys analogs relative to the fixed charge explicit solvent molecular mechanics calculations, and these discrepancies were greater for the -2 charged phosphate than the -1 charged phosphate, suggesting electronic polarizability does play a significant role in phosphate interactions. Nonetheless, Wong, et al. showed that both explicit solvent and implicit solvent models combined with a molecular mechanics energy function were able to provide good agreement with  $^3J_{\text{NH-CH}}$  coupling constants from NMR experiments for a model tetrapeptide, suggesting both may be useful for studying phosphorylation-induced conformational changes [18].

#### ***1.4.2. Sampling phosphorylation-induced conformational changes***

As with other aspects of protein function, molecular dynamics (MD) has been widely used to study effects of phosphorylation. MD involves the integration of Newton's equation of motion for a system of atoms. In addition to being relatively easy to perform, MD simulations have the advantage that they can be used in an unbiased manner, i.e., it is not necessary to make assumptions about the scale of conformational change and the relevant degrees-of-freedom. A drawback to conventional molecular dynamics simulations is that they can only realistically probe timescales of tens of nanoseconds, while the timescale of the conformational changes may be several orders of magnitude longer. The computational time required may range from several CPU-weeks to many CPU-years depending on the size of the system. The timescales accessible to

MD simulations, however, continue to grow through increases in software efficiency and in computational power.

A great deal of effort has been devoted to developing sampling methods that can be used with molecular mechanics force fields but are less subject to kinetic-trapping, i.e., insufficient sampling due to the short timescales accessible by traditional molecular dynamics. This issue can be severe when studying phosphorylated proteins due to very strong electrostatic interactions involving the phosphate group. Three methods that have been applied to studying protein phosphorylation, as discussed below, are targeted MD [20,21,22,23,24], accelerated MD [25,26], and torsion-angle based sampling methods (see Chapter 1). Targeted MD and related methods can be used to explore the free energy (and other properties) along pathways connecting 2 states, such as the phosphorylated and unphosphorylated conformations of a protein. Such methods obviously require knowledge of both states, a major limitation to prospectively making biologically-relevant predictions. Accelerated MD is one of a class of methods that improve sampling efficiency by modifying the energy function to effectively reduce the height of barriers separating low-energy states. A correct thermodynamic ensemble can be obtained by re-weighting the results. Along with conceptually related methods such as replica exchange, accelerated MD can dramatically increase the efficiency of generating a thermodynamic ensemble, and can be applied in an unbiased manner. Finally, torsion-angle based conformational search strategies represent an entirely different method of sampling than MD. Such methods randomly or systematically perturb a pre-determined set of torsion angles to sample conformations. These methods are much faster than MD simulations (typically days to weeks of CPU time), and less susceptible to kinetic traps

because torsion-angle moves can frequently cross large energy barriers. However, these methods generally require the use of implicit solvent, and require assumptions to be made about the torsion angles that are involved in the conformational change.

Molecular mechanics methods cannot be used to study catalysis, and combined quantum mechanics/molecular mechanics (QM/MM) methods have been used in a few cases to study how phosphorylation may affect catalysis [27,28]. These methods allow for a subset of the atoms to be treated quantum mechanically, while the rest of the system is treated classically using molecular mechanics energy functions. Single-point QM/MM energies can be calculated to examine changes in electronic structure due to phosphorylation, or MD simulations can be performed, though the timescales that can be reached—and the conformational changes that can be observed—are severely limited relative to purely classical MD.

#### ***1.4.3. Applications of computation to study phosphorylation-induced conformational changes***

Many protein kinases, responsible for the phosphorylation of specific protein substrates, themselves are regulated by phosphorylation as part of signal transduction and regulatory cascades. Over the past decade, x-ray crystallographic studies have contributed tremendously to our understanding of the functioning of protein kinases; protein kinase A (PKA), the cyclin-dependent kinase (CDK) family and particularly CDK2, and the Src-family kinases are especially well studied.

Phosphorylation sites known to regulate kinase activity can be found within the kinase domain itself, or in other portions of the protein. Within the kinase domain, a phosphorylation site (or sometimes more than one) is commonly located in the

“activation loop”, which forms part of the groove responsible for binding the peptide substrate and contacts structural elements containing residues important for catalysis [8,9,10]. In many cases, phosphorylation of the activation loop is a requirement for kinase activity. QM/MM studies combined with MD simulations showed that phosphorylation of Thr197 of the PKA catalytic subunit activation loop caused stabilization of the side-chain of the substrate serine in a conformation that is primed for the phosphotransfer reaction [28]. Activation loop phosphorylation has also been studied computationally for CDK2, a key player in cell-cycle regulation, using MD [29].

CDKs are also regulated by phosphorylation on the “glycine rich” loop, which provides part of the ATP binding site, forming interactions with the phosphate groups [30,31]. Puzzlingly, phosphorylation of Thr14 in both CDK2 and CDK5 is inhibitory [32] while phosphorylation of Tyr15 in CDK2 is inhibitory and in CDK5 is stimulatory [33]. Simulations of CDK2 suggested that Thr14 and Tyr15 phosphorylation inhibited activity by causing alterations in  $Mg^{2+}$  coordination and ATP alignment [34,35,36]. However, crystallographic studies have suggested that Tyr15 phosphorylation inhibits substrate peptide binding [37]. Simulations of CDK5 showed a similar inhibitory mechanism for Thr14 phosphorylation, but showed a shortened distance between the gamma phosphate and the P-site serine hydroxyl group upon Tyr15 phosphorylation along with stabilized coordination of  $Mg^{2+}$  [38]. Thus, the simulations suggest a mechanism by which Tyr15 phosphorylation may facilitate the phospho-transfer reaction.

Computational studies can also elucidate how the structural and dynamical effects of phosphorylation can propagate to portions of the protein outside the immediate vicinity of the phosphorylation site. Src-family kinases contain a kinase catalytic domain, a

phosphopeptide binding Src-homology 2 (SH2) domain, and a polyproline helix binding Src-homology 3 (SH3) domain, and can be phosphorylated on the activation loop, which stimulates kinase activity, as well as on the C-terminal tail, which inhibits kinase activity (see [8,9]). Conventional molecular dynamics simulations have been used to understand how the dynamics of the SH2-SH3 domain linker in the Src-family kinase Hck plays a key role in mediating C-terminal phosphorylation-mediated inhibition [39]. However, targeted MD simulations using the known inactive and active states of Src-family kinases have also been useful for studying dynamic properties of long-range conformational changes [20,22,23,39,40]. For example, Banavali and Roux [20] performed extensive simulations on Hck using RMSD-based biasing potentials to drive the kinase conformation towards the active conformation from the inactive conformation to observe the effects of activation loop phosphorylation on the kinase activation process. They showed that bidirectional allostery between the hinge connecting the two lobes and the phosphorylated region of the activation loop proceeds through sequential changes in local contacts between structural elements, including the DFG motif and the C-helix, as opposed to a single concerted structural transition.

Several recent computational studies have examined structural changes in peptides upon phosphorylation. Peptides are very amenable to computational studies, as much greater sampling can be performed due to the relatively small size of the system. In many cases, including the ones discussed here, the peptides are portions of a larger protein. Clearly, studying phosphorylation, either computationally or experimentally, of one small portion of a larger protein has limitations; it is not always clear that the conclusions from such studies are biologically relevant.

However, in certain cases such studies can provide valuable insights, both with respect to fundamental biophysical aspects of phosphorylation, and with respect to the structural biology. In particular, certain protein-protein interactions are mediated by specific peptide motifs located within “unstructured” portions of a well-folded protein, or within intrinsically unstructured proteins. The ability of these peptides to form interactions depends on the accessibility of the peptide segment and may also depend on the conformation, both of which may be regulated by phosphorylation.

Several computational studies of peptide fragments have suggested that phosphorylation induces increased local structural ordering of an otherwise disordered segment. For example, MD simulations suggested that phosphorylation of the kinase-inducible domain (KID) of CREB primes it for recognition by the KIX domain of CREB-binding protein by inducing formation of a local turn-conformation similar to that seen in the pKID/KIX NMR structure [41]. However, because thermodynamic analysis found that the binding of pKID to the KIX domain was mostly enthalpic [42], it seems that this induction of structure does not contribute significantly to the binding free energy [43]. Increased ordering upon phosphorylation was also observed in simulations of the N-terminal domain of the smooth muscle myosin regulatory light chain (RLC). Using MD simulations of a 25-residue segment of the RLC, Thomas and co-workers provided an all-atom structural model in accordance with EPR data that begins to explain how phosphorylation produces a 1000-fold increase in myosin activation [44]. Phosphorylation was observed to increase the helical ordering, and an MM/PBSA approach with quasi-harmonic approximation of the conformational entropy was used to estimate the enthalpic and entropic consequences of this transition [45].

Early studies indicated that phosphorylation could stabilize the N-termini of alpha-helices through favorable electrostatic interactions with the amides of the helix backbone [46]. An increased preference for alpha-helical conformations after phosphorylation was also observed in peptides in which the phosphorylated Ser or Thr immediately precedes a proline [25]. This peptide motif is important because many serine-threonine kinases, so-called proline-directed kinases, phosphorylate only a serine or threonine residue preceding a proline. After phosphorylation, these sites may be targeted for cis/trans isomerization of the proline residue, most notably by Pin1, resulting in a conformational switch [47]. Use of a novel MD method also allowed McCammon and colleagues to show that rates of cis/trans isomerization of proline residues are slowed by phosphorylation of a preceding serine [25].

Some proteins contain peptide regions that are hyperphosphorylated, i.e., phosphorylated on several residues close in sequence. For example, hyperphosphorylation of the microtubule-associated protein tau is implicated in Alzheimer's disease (reviewed in [48]). The structural effects of hyperphosphorylation are not well understood. One hypothesis is that the addition of several bivalent dianions could cause severe electrostatic repulsion and cause disorder in peptide structure. However, in a study of the phosphorylation of all eight serines of the (RS)<sub>8</sub> motif of the SR protein, McCammon and colleagues present a more complicated picture [26]. The simulations suggested that phosphorylation induces unfolding of the alpha-helix and causes the peptide to take on an extended structure or a compact "arginine-claw" structure. The arginine-claw is characterized by a single phosphate coordinated by 5–6 arginines while the other phosphates remain solvent-exposed. While experimental

evidence is required to validate the existence of the arginine-claw structure, these results generate testable hypotheses regarding the characteristics of (RS)<sub>8</sub> motif phosphorylation and its potential role as a nuclear import signal.

The important tumor suppressor protein p53 contrasts with the studies discussed above in that MD simulations showed that phosphorylation did not induce a structural transition to inhibit binding to MDM2, and may operate instead through direct electrostatic repulsion [49]. These findings have been partially supported through fluorescence anisotropy measurements of binding of unphosphorylated and phosphorylated p53 peptides to wild-type and mutant MDM2, where anionic-to-neutral mutations in the binding site resulted in a decrease in the inhibitory effect of phosphorylation [50].

Many membrane proteins are regulated through phosphorylation. In some cases, phosphorylation regulates membrane protein trafficking, while in others, phosphorylation directly affects membrane protein function. As with other classes of proteins, the function and mechanism of many phosphorylation sites in membrane proteins are unknown. High-resolution structural analysis of membrane proteins is notoriously difficult experimentally, and computation has proven useful to study membrane protein structure and dynamics [51].

A small protein lending itself well to all-atom computation, the 52-residue phospholamban binds and inhibits the SERCA calcium pump when unphosphorylated. Phosphorylation of the phospholamban N-terminal cytoplasmic domain on Ser16 (or Thr17) relieves this inhibition and activates calcium uptake. NMR and FRET studies provided different and potentially conflicting views of the mechanism of the



phosphoregulation. NMR studies showed disordering of part of the  $\alpha$ -helical cytoplasmic domain upon phosphorylation of Ser16, which was suggested to not form a stable salt-bridge [52]. By contrast, FRET studies observed a decrease in the distance between the ends of the helices suggesting a coil-to-helix transition, which was hypothesized to be mediated by salt-bridge interactions between pSer16 and Arg13 [53]. Replica-exchange molecular dynamics (REMD) of the isolated N-terminal domain recapitulated the phosphorylation-induced disordering of the alpha-helix observed by NMR. However, the simulations also demonstrated that phosphorylation results in an ensemble of conformational states showing a decreased end-to-end distance mediated in part by pSer16-Arg13 salt bridge, but also through the formation of other salt bridges in different conformational states as well [54]. These simulations seem to resolve the apparently conflicting NMR and FRET results, and highlight the use of computation to aid interpretation of experimental measurements of conformationally-averaged observables.

Several ion channels and transporters are known to be regulated by phosphorylation, and in some cases, misregulation of ion channels by phosphorylation can have pathological consequences [55,56]. Comparative modeling and MD studies of the SUR2B subunit of the Kir6.1/SUR2B ATP-sensitive potassium channel indicated that phosphorylation-induced conformational changes strengthen interactions between a nucleotide-binding domain and the associated transmembrane helix, potentially suggesting that phosphorylation is involved in coupling ATP hydrolysis to structural changes of the TM helices [57]. The modeling and simulations suggested that large motions of secondary structure elements are driven by the formation of a stable tyrosine-

arginine-phosphoserine triad. Experimental studies provided some evidence supporting this model; mutation of two residues from the triad (tyrosine to phenylalanine and arginine to alanine) severely impaired the ability of PKA to activate the channel. It remains to be seen whether these conformational changes would occur in the context of the full tetrameric channel, but the initial validation through experiment is encouraging.

## 1.5. Preview

The work presented here for my dissertation is arranged as follows: Chapter 2 presents work using molecular mechanics methods originally developed for comparative modeling to predict conformational changes in protein loops and helices induced by phosphorylation; Chapter 3, which is a digression inspired by phosphorylation-induced changes in the orientation of the N- and C-terminal lobes of protein kinases, presents the beginnings of a method for the prediction of the orientation of protein domains using antibodies as a model system; Chapter 4 presents currently unpublished work prospectively predicting the structural effects of phosphorylation on the Arp2/3 complex; and chapter 5 outlines future directions in the use of computation to study the structural and energetic effects of important post-translational modifications.

# Chapter 2 Conformational Changes in Protein Loops and Helices Induced by Post-Translational Phosphorylation

The remainder of this chapter presents a revised version of work published in PLoS Computational Biology (E.S. Groban\*, A. Narayanan\*, and M.P. Jacobson (2006) 2: e32 ). I was joint first author with Eli Groban, with whom I equally split the dataset selection, predictions of phosphorylated loop and helix structures, and the analysis of the CDK2 case study. The writing of the manuscript was split between Eli Groban, Matt Jacobson, who also directed this research, and myself.

## 2.1. Introduction

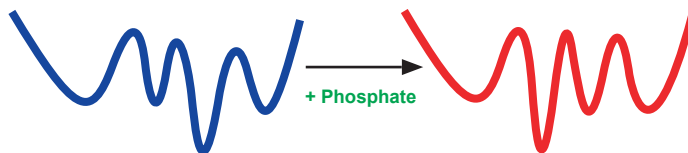
In this work, we examine how phosphorylation can modulate the conformation of a protein by changing the energy landscape. We limit our attention in this work to relatively small conformational changes involving a phosphorylated loop and its surroundings, and in one case, conformational changes in a helix and its surroundings. At physiological pH, the phosphate group predominantly carries a  $-2$  charge, whereas the

site of phosphorylation is neutral before modification (in the case of Ser, Thr, and Tyr). The conceptual foundation of our work is shown in Figure 2.1. We view the phosphate as a perturbation to the energy landscape of the protein. Our goal in this work is to predict the conformational changes caused by this perturbation. That is, given the structure (or an accurate model) of the unphosphorylated protein, can we predict the structure of the phosphorylated protein? The critical tools we need to understand and predict the structural effects of post-translational phosphorylation are as follows:

One tool that is needed is an energy function that captures the essential physics, especially for the modified residues. We use a molecular mechanics energy function consisting of the OPLS-AA force field and a Generalized Born implicit solvent model. The use of a physics-based energy function is important for two reasons. First, there are too few structures of phosphorylated proteins to make knowledge-based energy functions feasible. Second, conformational changes induced by phosphorylation are largely driven by the electrostatic perturbation induced by the phosphate group. We use the energy function not only to predict conformational changes induced by phosphorylation, but also to better understand the key physical effects underlying these conformational changes, especially in our case study of cyclin dependent kinase 2 (CDK2).

Another necessary tool is sampling algorithms capable of exploring critical degrees of freedom. After phosphorylating a protein *in silico*, we need to explore the new energy landscape and identify the new global energy minimum, if in fact it is significantly different than that of the unphosphorylated protein. In principle, molecular dynamics could be used for this purpose, and in some cases, this type of strategy has been successful, in studies by ourselves and others [35,40,58,59,60,61,62,63,64,65,66,67,68].

However, the timescales for converting from the non-phospho to the phosphorylated form are unknown, and in fact are not physically meaningful because the *in silico* phosphorylation is alchemical, i.e., we do not attempt to mimic a kinase actually performing the phosphorylation. For relatively large conformational changes, or those involving high-energy barriers, e.g., Pro peptide bond isomerization, the relevant timescale could be quite long—microseconds or more. Instead, we use a strategy in which we have adapted algorithms that we previously developed for homology modeling, i.e., sampling methods for side chains [69,70], loops [71,72,73], and helices [74]. The essence of the approach is to combine dihedral angle sampling methods, which enable large energy barriers to be surmounted, with direct minimization to enumerate many local minima. We discuss our strategy in detail in Methods.



**Figure 2.1 Phosphorylation Can Perturb the Energy Landscape of a Protein to Cause Changes in Conformation and Dynamics**

This figure is a visual representation of such changes. In this work, we predict the structural change of proteins due to phosphorylation by locating the new global energy minimum of the energy surface.

The key conclusion of this study is that our molecular mechanics–based methods appear to be capable of reproducing conformational changes induced by post-translational phosphorylation, with near-atomic resolution in most cases considered here, which are limited to relatively modest conformational changes and not, e.g., more drastic order–disorder transitions. This work thus represents a significant step toward a broadly applicable method for predicting structural effects of phosphorylation. Through a detailed case study of CDK2, we also illustrate how the computational methods can be

used to provide new understanding of how phosphorylation drives conformational change, why substituting Glu or Asp for a phosphorylated amino acid does not always mimic the effects of phosphorylation, and how a phosphatase can “capture” a phosphorylated amino acid.

## 2.2. Results and Discussion

We first present an in-depth study of activation loop phosphorylation in CDK2, both to introduce the computational methods and to highlight the ability of these methods to gain new insights into how phosphorylation drives conformational changes. We then present a broader survey of our ability to predict phosphorylation-driven conformational changes.

### 2.2.1. *Case Study: CDK2*

CDK2 is a member of the cyclin-dependent kinase protein family [75], whose members play a central role in cell cycle regulation. CDK2 is activated through binding of cyclin A and post-translational phosphorylation of a threonine residue on the activation loop, which lies close to the catalytic site. Compared to other phosphorylated proteins, there is a wealth of structural information for CDK2 in both its phosphorylated and unphosphorylated forms [9]. In this case study, we examine the ability of our method to: (1) reconstruct and predict the active conformation of the activation loop in CDK2 bound to cyclin A [76]; (2) reconstruct the cyclin A dependence of CDK2 activation [76,77]; (3) discriminate between two potential phosphate localization sites in the CDK2/kinase-associated phosphatase (KAP) complex [78]; and (4) examine the effects of substituting

Thr160 with a Glu, which results in only partial activation relative to phosphorylation of Thr160 [79].

A structural alignment and superposition of the active and inactive [80] forms of CDK2 bound to cyclin A reveals that the global C $\alpha$  root mean square deviation (RMSD) between the two structures is only 0.5 Å if residues 152–163 are removed from the calculation. The phosphorylatable threonine residue itself moves approximately 10 Å upon phosphorylation, as measured at the  $\gamma$  oxygen. In the unphosphorylated structure, Thr160 is well solvated and somewhat disordered (B-factors > 50). In the phosphorylated structure, pThr160 localizes to interact with a cluster of Arg residues (50, 126, and 150). We will refer to this and similar conformations as “active” conformations of the loop.

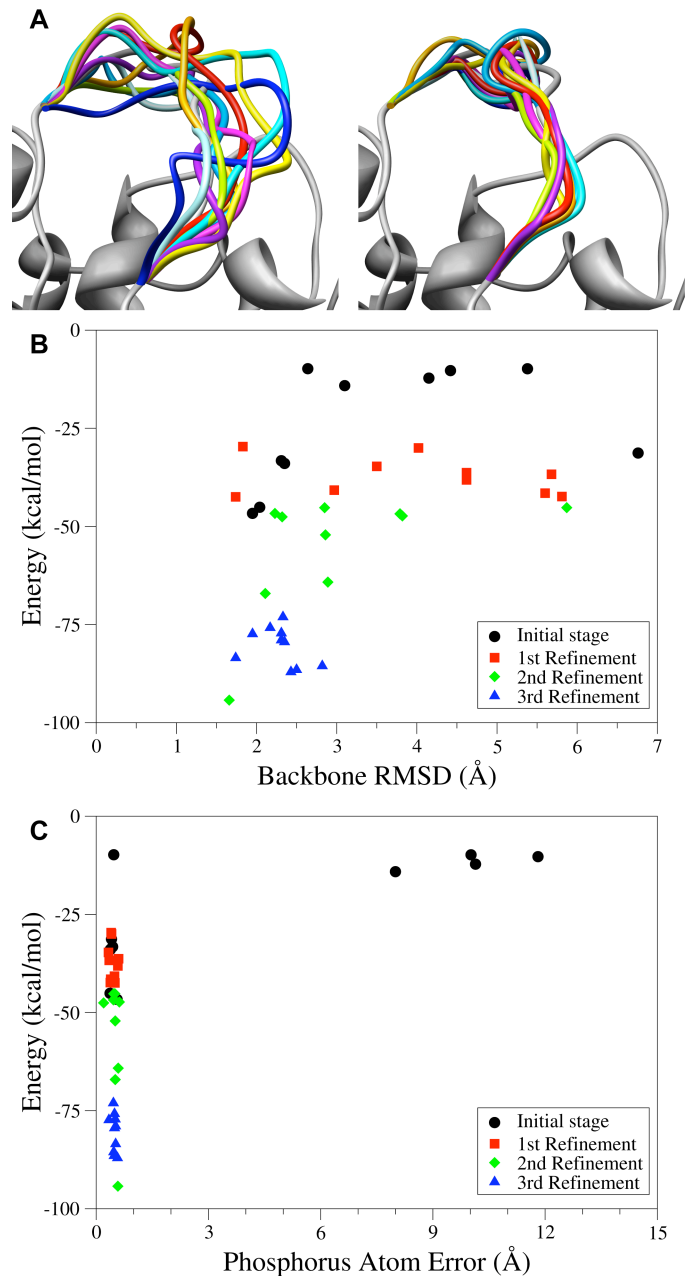
As described in Methods, we use a hierarchical loop prediction algorithm [72] based upon dihedral angle backbone sampling, rotamer-based side chain optimization, an all-atom force field and a Generalized Born solvation model [14,15] to predict the structural consequences of phosphorylation on loops and their surroundings. The hierarchical prediction algorithm allows us to quickly prune the conformational space and focus sampling on energetically favorable regions of conformation space (Figure 2.2). Unlike Monte Carlo and molecular dynamics sampling schemes, the algorithm itself has no knowledge of the starting conformation of the loop. The entire conformational space of the loop is sampled with varying sampling resolution in a hierarchical manner. This method has proven successful in recreating crystallographic conformations of unphosphorylated loops to loop lengths of 12 residues. In this paper, we expand on this loop prediction method to predict the structures of phosphorylated loops.

Before attempting to predict the phosphorylated structure from the unphosphorylated structure, we first ensure that we can predict the conformation of the phosphorylated loop with high accuracy when all other portions of the phosphorylated structure are taken from the crystal structure. We refer to this test as “loop reconstruction.” This tests the suitability of the energy function for identifying the correct conformation of the phosphorylated loop and the suitability of the sampling function for generating native-like structures. Much of our analysis will focus on the accuracy with which the phosphate group is positioned, which is measured by the distance of the phosphorus atom to its crystallographic position. We also provide other measures of accuracy for the overall loop, including a backbone RMSD measure (calculated based on the N, C $\alpha$ , and C atoms), and the RMSD calculated over all heavy atoms, including side chains. For all of these measures of accuracy, the predicted structure is first aligned to the reference crystal structure by a least-squares superposition of all C $\alpha$  atoms, excluding the loop being simulated. Thus, structural differences outside of the loop region being predicted can affect the reported measures of accuracy.

For the complex of phosphorylated CDK2 with cyclin A, we are able to reproduce the conformation of the activation loop with less than 2 Å RMSD overall. The phosphate group is placed with particularly high accuracy, less than 1 Å error as measured by the position of the phosphorous atom (Figure 2.2 and Figure 2.3, and Table 2.1). We also wish to highlight the differences between Figure 2.2B and Figure 2.2C, which plot the energy of different local minima identified during the four-stage prediction algorithm versus the backbone RMSD (Figure 2.2B) and the error in the phosphate position (Figure 2.2C). In the energy versus backbone RMSD plot, an energy “funnel” is clearly evident.

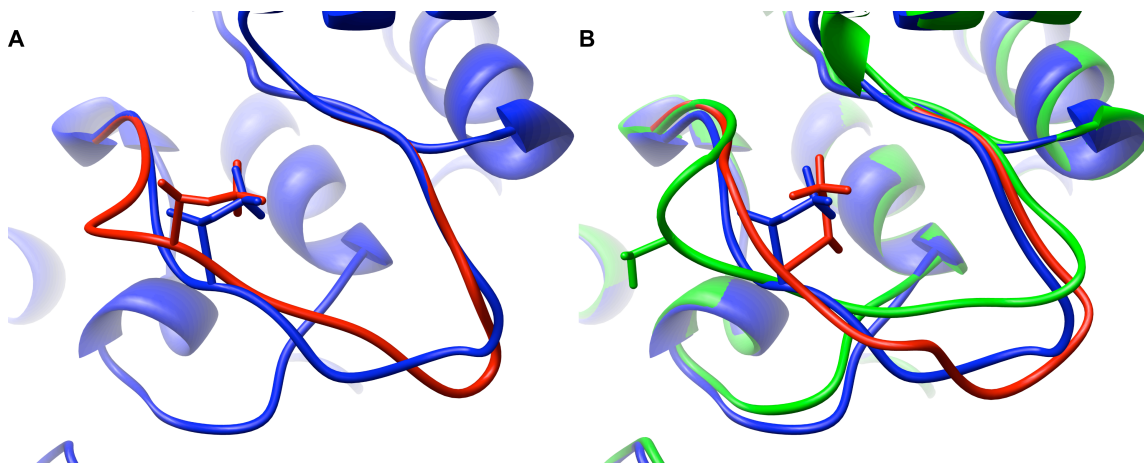


That is, the lowest energy identified decreases as the loop, as a whole, approaches native-like conformations. On the other hand, the phosphate group itself appears to respond to a narrow, deep energy well. That is, the phosphate locates the Arg cluster very early in the simulation (during the first of the four stages of the loop prediction algorithm), and remains there while the rest of the loop adjusts to the new position of the phosphate group.



**Figure 2.2 Example of the Hierarchical Loop Prediction, Applied to Reconstructing the Phosphorylated Activation Loop of CDK2/Cyclin A**

In this example, only the loop is predicted, and the remainder of the protein is held rigid in its crystallographic conformation. (A) Backbone traces for the ten lowest-energy loops sampled in the initial build-up stage (left) and the third and final refinement stage (right) of the hierarchical loop prediction protocol (figures prepared with Chimera [81]). (B) Energies as a function of backbone RMSD values of the 20 lowest-energy loops sampled in the four stages of the loop prediction protocol. (C) Energies as a function of the error in the phosphorus atom position, relative to the crystal structure, for the 20 lowest-energy loops sampled in the four stages of the loop prediction protocol.



**Figure 2.3 Loop Reconstruction and Prediction in CDK2/Cyclin A**

(A) Reconstruction: Crystal structure of phosphorylated CDK2/cyclin A (blue) and the predicted loop structure (red). The starting structure for the prediction was the phosphorylated structure, with the only difference being the loop region. (B) Prediction: Crystal structure of phosphorylated CDK2/cyclin A (blue), the crystal structure of the unphosphorylated CDK2/cyclin A (green), and the predicted loop structure upon in silico phosphorylation of the unphosphorylated CDK2/cyclin A structure (red). Figures prepared with Chimera [81].

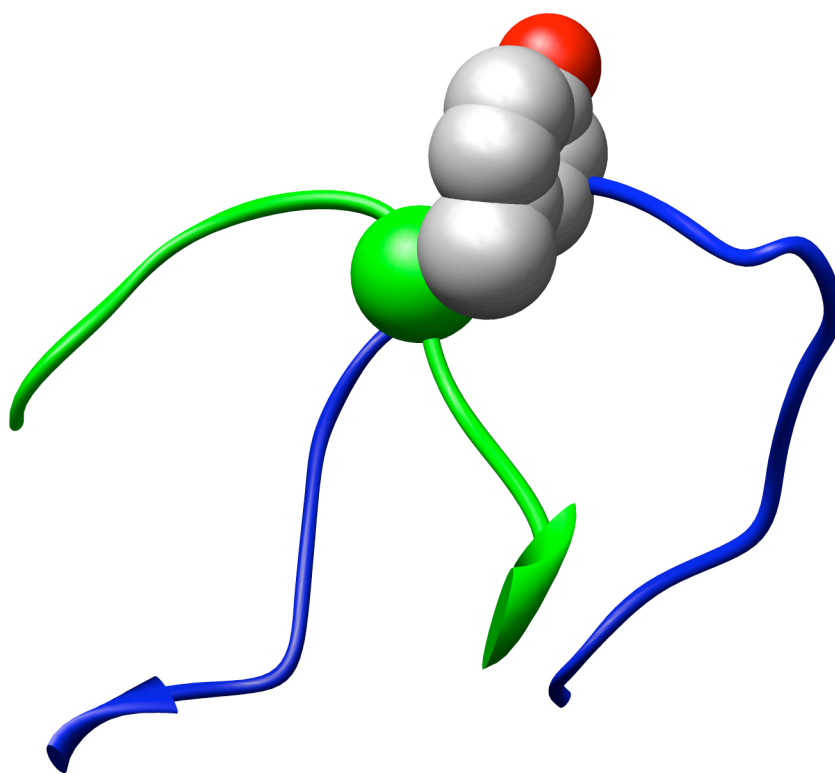
Case	Deviations from Phosphorylated Crystal Structure (Å)			
	Loop Only		Loop + Surroundings	
	P	BB	P	BB
Reconstruction	0.6	1.7	1.7	1.8
Prediction	8.8	6.7	0.8	2.9
CDK2/KAP	1.1	2.3	0.5	1.3

**Table 2.1 Results from the CDK2 Case Study**

The case study is summarized using two measures: the deviation of the phosphate group from its position in the crystal structure of phosphorylated CDK2 (“P”), and the overall backbone RMSD of the predicted loop compared to the same structure (“BB”). For the CDK2/KAP case, the reference structure is that of phosphorylated CDK2 in complex with KAP (1fq1), whereas the other cases are compared to the crystal structure of phosphorylated CDK2 bound to cyclin A (1JST). “Loop Only” refers to predicting only the loop in question; these results are provided only for comparison. “Loop + Surroundings” refers to the protocol in which we predict the loop residues as well as side chains of residues within 4.5 Å of any atom in the loop; these are the results that should be used to evaluate the success of the method. The different prediction cases are as follows. “Reconstruction” refers to predicting residues 152–163 in the crystal structure of phosphorylated CDK2 in complex with cyclin A. “Prediction” refers to the prediction of residues 152–163 after in silico phosphorylation of unphosphorylated CDK2/cyclin A (1fin). Finally, “CDK2/KAP” refers to predicting residues 155–165 in the structure of the phosphorylated activation loop of CDK2 when in complex with its phosphatase, KAP. It is important to note that the prediction methodology being used has no knowledge of the starting structure of the loop.

Next, we examine our ability to predict the phosphorylated loop conformation starting from the unphosphorylated structure. Naively attempting to predict the conformation of the phosphorylated activation loop from the unphosphorylated structure, holding the nearby side chains fixed, results in a poor prediction, with an error in the

position of the phosphorus atom of approximately 9 Å and a backbone RMSD of approximately 7 Å (Table 2.1). The reason for this failure is clear. A comparison of the phosphorylated and unphosphorylated structures reveals that the backbone of the phosphorylated conformation of the activation loop passes directly through the side chain of Tyr179 in the unphosphorylated form (Figure 2.4). This single misplaced side chain can prevent the loop from assuming the correct activated conformation. Thus, in this case, introduction of the phosphate group not only changes the activation loop conformation, but also rearranges the side chain hydrogen bonding network well beyond the loop itself.



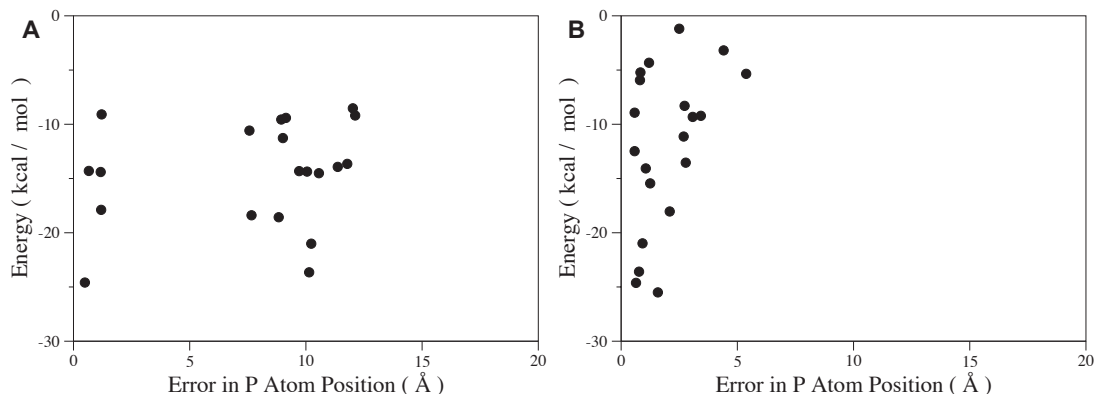
**Figure 2.4 Detailed View of One Portion of the Structural Superposition between the Phosphorylated and Unphosphorylated Crystal Structures of CDK2/Cyclin A**

The phosphorylated activation loop (blue) passes through the middle of Tyr179 (CPK) when inserted into the non-phosphorylated structure (green) of CDK2. Figure prepared with Chimera [81].

Thus, sampling of loop conformations must be coupled with, at least, the sampling of side chain conformations in the vicinity of the loop to permit accurate prediction. As described in Methods, we have adopted a strategy in which side chains near the loop are removed during the loop build-up, and side chains both on the loop and its vicinity are optimized and energy minimized simultaneously for the representative loop from each cluster. Applying this method to the loop reconstruction test increases the error by a small amount. Overall, the error in the phosphorus position increased by about 1 Å, and the backbone RMSD for the activation loop increased an insignificant amount (Figure 2.5 and Table 2.1). We note that Figure 2.5A clearly shows two major “basins” of attraction for the phosphorylated amino acid. The one with the lower energy places the phosphate in contact with the Arg cluster; the higher energy state places the phosphate in solution. These two conformations, which we will call “active” and “inactive,” respectively, will be examined further below.

When predicting the structural change upon phosphorylation of CDK2, this strategy permits Tyr179, as well as the three arginine residues that form salt bridges with pThr160, to re-optimize around the loop. The prediction with the simultaneous optimization of the loop and surrounding side chains results in a phosphorous atom error of 0.8 Å and a backbone RMSD of 2.9 Å (Figure 2.3 and Figure 2.5B, Table 2.1). The phosphorus atom position is quite accurate, whereas the remainder of the loop may require the optimization of other structural elements to assume the active conformation. The prediction clearly captures the qualitative change in conformation upon phosphorylation, however. Additionally, we performed the above calculations again, except this time substituting the three Arg residues of the arginine cluster with Lys

residues (unpublished results). In this case, the phosphate does not localize to the new “Lys cluster,” and we predict that the kinase would not be active. The preference of phosphorylated side chains for interacting with Arg, relative to Lys, has been noted previously, and we will explore this issue further in a separate publication.



**Figure 2.5 CDK2 Case Study**

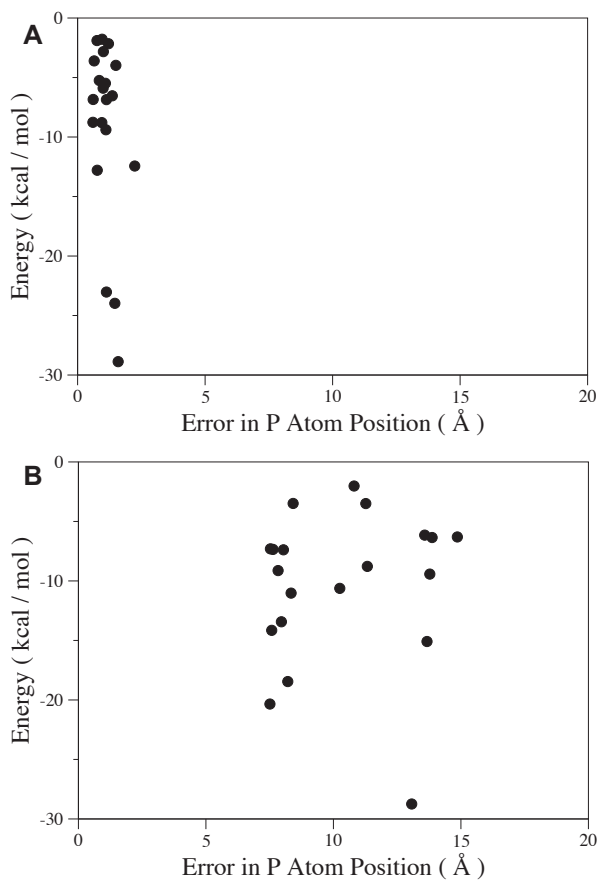
The 20 lowest-energy loops predicted for (A) reconstruction of residues 152–163 in phosphorylated, cyclin-bound CDK2; and (B) prediction of the structure of residues 152–163 upon in silico phosphorylation of Thr160 in the unphosphorylated, cyclin-bound CDK2. In each case, the loop and surrounding side chains are optimized simultaneously as described in Materials and Methods.

### 2.2.2. *CDK2 Phosphorylation in the Absence of Cyclin Binding*

Cyclin A binding is required for the full activation of CDK2 because phosphorylated CDK2 without cyclin A exhibits only 0.3% of the activity of the fully active, phosphorylated, cyclin A-bound structure [9]. The crystal structure of phosphorylated CDK2 in the absence of cyclin A shows that the activation loop is disordered [77]. The loop thus likely adopts several dynamically interconverting conformations with the pThr fully solvated, instead of bound by the arginine cluster. The activation loop in unphosphorylated CDK2 in the absence of cyclin A is ordered, albeit with somewhat elevated B-factors (roughly 60).

The results of predicting the loop in the phosphorylated, cyclin A-unbound structure [77], showed that the phosphate did not localize to the Arg cluster, which is qualitatively consistent with phosphorylation not contributing to activation (results not shown). However, the low-energy predicted loop conformations showed the phosphate localized at a different position, and the results did not provide any evidence that the loop should in fact be disordered. This is due to a deliberate bias in the loop prediction algorithm, which was originally designed to predict well-structured loops, not ensembles of structures characterizing a disordered loop. A full solution will require an algorithm that obeys detailed balance and thus correctly treats the loop conformational entropy; we have recently developed a Monte Carlo version of the loop sampling algorithm that accomplishes this goal, and we will report results in due course. As a preliminary step, we have performed the calculations with one key modification to the existing loop prediction algorithm. Specifically, the algorithm constrains the loop conformations to be relatively close to the body of the protein using a simple distance cutoff; more than 99% of loops in the Protein Data Bank (PDB) satisfy this criterion [72]. However, this restriction prevents disordered loops from exploring conformations in which the loop is well solvated and makes few contacts to the remainder of the protein. By simply turning off this screening, the loop prediction for phosphorylated CDK2 in the absence of cyclin A does show evidence of diverse low-energy conformations (Figure 2.6B), in the sense that the 20 lowest-energy conformations are highly diverse; the phosphorylated Thr160 is well solvated in most of these. Turning off the screening does not significantly affect the results for the in silico phosphorylation of CDK2 with cyclin A bound: pThr160 still localizes strongly to the Arg cluster (Figure 2.6A) in all 20 of the lowest-energy

conformations. Thus, in this case, the computer simulations can provide at least a qualitative prediction that phosphorylation of CDK2 in the absence of cyclin binding leads to disorder. More work is clearly required to test this capability in other cases of phosphorylation-induced disorder; the remainder of our results focuses exclusively on cases in which the phosphorylated loops are well ordered.



**Figure 2.6 Differences in Conformational Predictions for Phosphorylated CDK2 with and without Cyclin A Bound**

The 20 lowest-energy loops are considered for predicting the conformation of residues 152–163 upon *in silico* phosphorylation of Thr160 (A) in cyclin-bound CDK2 and (B) CDK2 in the absence of cyclin A. In (B), the x-axis represents the deviation of the phosphate from its position in the fully activated CDK2/cyclin A complex; the activation loop and pThr160 are disordered in the crystal structure of phosphorylated CDK2 in the absence of cyclin A. These calculations were performed with a consistent set of parameters that do not bias the results toward well-ordered loop structures. In the absence of cyclin A, the phosphate does not localize to the Arg cluster as in the cyclin bound case, and the 20 lowest-energy structures show considerable diversity in conformation.



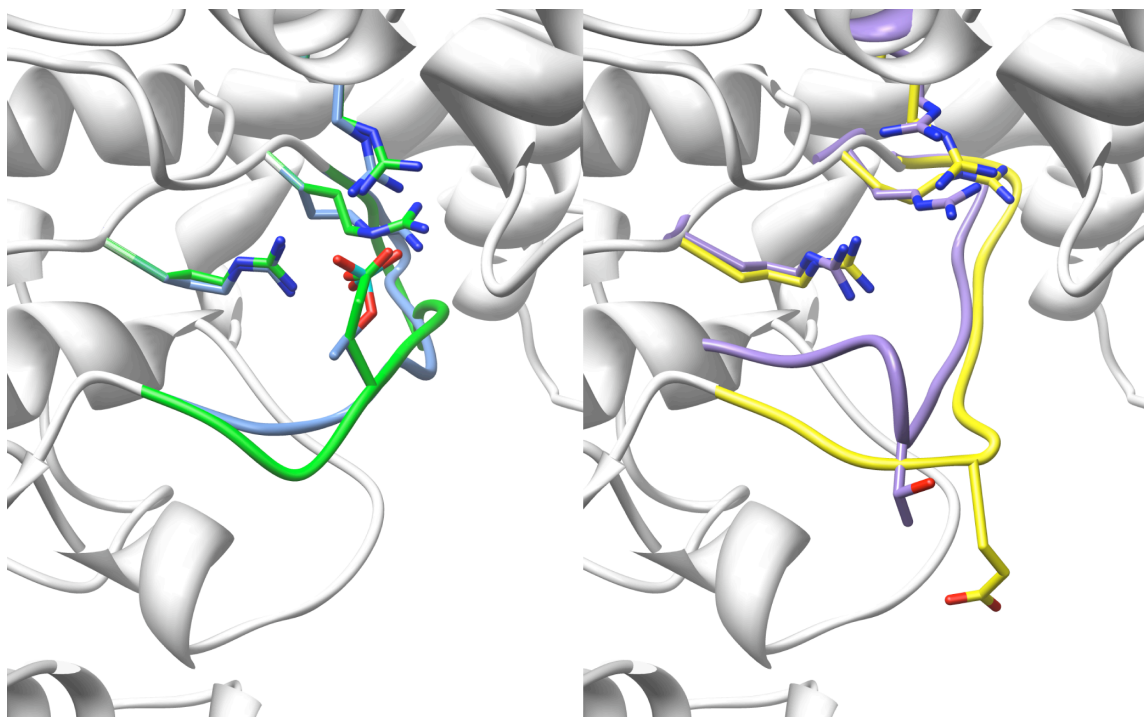
### 2.2.3. *Substitution of Glu for the Phosphorylated Thr in the Activation Loop of CDK2*

Among proteins that require phosphorylation for activation, there are many examples of a constitutively active mutant in which the phosphorylatable residue is substituted with either aspartate or glutamate [82,83,84,85,86]. Connell-Crowley et al., however, show that in the CDK2/cyclin A complex, a T160E mutation confers no additional activity to the complex [79]. To examine why, we use the active form of CDK2/cyclin A to create a CDK2/cyclin A T160E mutant *in silico*, and then use our prediction protocol to predict the conformation of the activation loop, residues 152–163 in the mutant protein. Using the protocol in which side chains surrounding the loop are optimized concurrently with the loop itself, the lowest-energy structure places the C $\beta$  atom of Glu160 9 Å away from its position in the wild-type active loop (Figure 2.7). The side chain of Glu160 is fully solvated, and the overall loop conformation is reminiscent of the “inactive” conformations identified in the loop prediction results with pThr160. On the other hand, the second lowest-energy conformation places the side chain of Glu160 in a position that is highly analogous to the position of pThr160 in the CDK2/cyclin A complex (i.e., contacting the Arg cluster; see Figure 2.7). We refer to this as the “active” conformation.

As shown in Table 2.2, the energy difference between these two states is small for the T160E mutant, with the inactive conformation slightly more stable than the active conformation. For the pThr160 structure, as discussed above, the active structure is correctly predicted to be lower in energy, and the energy gap is somewhat larger (18 kcal/mol). These relatively small differences in overall energies of the active and

inactive conformations mask very large differences in the individual energy components. Specifically, the active conformation is strongly favored by the Coulomb electrostatics term in the molecular mechanics energy (172 kcal/mol for the Glu160 case and 298 kcal/mol for the pThr160 case), primarily due to the interaction of the negatively charged side chain of Glu160 or pThr160 with the Arg cluster. On the other hand, the solvation free energy computed from the Generalized Born solvent model strongly favors the inactive conformations, primarily because both Glu160/pThr160 and the Arg cluster are much more solvent exposed in this conformation. The covalent term favors the inactive form; in other words, the active form involves significant internal strain that is relieved in the more open inactive state. In the Glu160 case, the energy differences in the Coulomb and solvation terms largely cancel, and the inactive state winds up slightly lower in energy. In the pThr160 case, the difference in the Coulomb term is significantly larger than the solvation term, causing the active state to be favored by a significant margin.

The differences between Glu160 and pThr160 of course derive primarily from the difference in the total charge on the side chains. The  $-2$  charge on pThr nearly doubles the favorable Coulombic electrostatic attraction to the Arg cluster relative to the singly charged carboxylate group on Glu. The more highly charged pThr side chain also has a more favorable solvation free energy in the inactive form, but the overall difference in solvation free energy does not change by as great a factor, in part because greater solvation of the Arg cluster in the inactive state represents a significant contribution to the difference in solvation free energy, and this component is essentially independent of whether Glu or pThr is present at residue 160.



**Figure 2.7 Active and Inactive Conformations of the CDK2 Activation Loop**

Left: Blue represents the crystal structures of the phospho-CDK2/cyclin A complex, and green represents the T160E-predicted active-like con-formation. The Arg cluster is shown in stick representation. The carboxylate group of Glu160 and the phosphate group of pThr160 are almost exactly superimposed. Right: Purple represents the crystal structure of unphosphorylated CDK2/cyclin A, and yellow represents the predicted inactive conformation of T160E. These two structures are qualitatively similar in that Thr160 and Glu160 both point out into solvent and the Arg cluster is better solvated. Figures prepared with Chimera [81].

Energy Component	$\Delta E = E_{\text{active}} - E_{\text{inactive}}$		
	pThr160	T160E	CDK2/KAP
Covalent	15.3	16.4	10.8
Coulomb	-298.2	-172.3	-110.9
Lennard-Jones	23.8	-2.4	30.1
Solvation	240.6	161.2	94.6
Total	-18.5	2.9	24.6

**Table 2.2 Energy Differences (kcal/mol) between “Active” and “Inactive” Conformations of the Activation Loop**

The energy differences are broken down by different components of the molecular mechanics energy: covalent (bonds, bond-angles, and torsions), the Coulombic electrostatic term, Lennard-Jones, and the solvation free energy. The active form is defined by pThr160 being localized near the Arg cluster. In the pThr160 case, this is the lowest-energy-predicted structure, i.e., as represented in Table 2.1. In the Thr160Glu (T160E) substitution, the lowest-energy structure has the Glu side chain pointed out into solution; we refer to this as the “inactive” conformation. The second highest-energy structure has the Glu side chain in a position analogous to the pThr side chain, and we refer to this as the “active” structure for T160E. For CDK2/KAP, active/inactive take on different meanings. The active conformation is analogous to the other active structures, i.e., the phosphate group is localized at the Arg cluster. The inactive conformation corresponds to lowest-energy conformation, in which the phosphate is correctly localized near the N-terminus of a helix in KAP.

#### 2.2.4. *Phosphorylated CDK2 in Complex with Its Phosphatase*

Finally, we examine conformational changes in the activation loop necessary for dephosphorylation to occur. KAP is responsible for the dephosphorylation, and subsequent inactivation, of CDK2 [9,78]. A crystal structure of the CDK2/KAP complex shows that pThr160 localizes to the N-terminus of a helix in the KAP protein, and interacts in that position with a single arginine side chain. The phosphate group must be stable enough in this position to allow the phosphatase to “pull” pThr160 from the Arg cluster on CDK2. Given the highly favorable electrostatic interactions between pThr160 and the three Arg side chains in the cluster, this did not seem intuitively obvious.

Nonetheless, the loop predictions in the CDK2/KAP complex did in fact place the phosphorous atom within 0.5 Å of its position in the crystal structure (Table 2.1); the lowest energy structure with pThr160 placed in contact with the Arg cluster is approximately 25 kcal/mol higher in energy. Table 2.2 helps to explain why. The phosphate placed close to the Arg cluster does in fact have much more favorable electrostatic interactions, as measured by the Coulombic term in the molecular mechanics energy function (111 kcal/mol difference). However, the solvation term strongly favors the pThr160 in its correct position in contact with KAP, due largely to the improved solvent accessibility of the Arg cluster after the pThr160 is removed from contact. In addition, the pThr160 is not as buried in the KAP complex. Finally, the covalent terms again indicate significantly higher internal strain in the activation loop when the pThr160 is in contact with the Arg cluster, and much of this strain is relieved when the loop adopts the phosphatase-bound conformation (this might be referred to as a spring-loaded mechanism).

### 2.2.5. Predicting Conformational Changes of Loops in Response to Phosphorylation

Encouraged by our results on CDK2, we have performed similar calculations on a larger, more diverse test set (Table 2.3). Our results show that in all cases of loop reconstruction, we reproduce the structure of the phosphorylated loop with high accuracy (Table 2.4). We reconstruct seven of the loops to within 1 Å backbone RMSD with respect to the original crystal structure and predict the phosphorus atom to within 0.5 Å of its crystallographic position. Although our predictions are not error free, the relatively small magnitude of the error gives us some confidence that we can predict the correct conformation of a phosphorylated loop using our hierarchical sampling methodology, all-atom force field, and Generalized Born solvation model.

Protein	Phosphorylated Structures (PDB ID)	Phosphorylated Residue(s)	Residues Predicted	Unphosphorylated Structures (PDB ID)
Hpr	1FU0 [87]	Ser46	42–48	1PTF [88]
SpoIIAA	1H4X [89]	Ser57	49–58	1H4Z [89]
Psp	1J97 [90]	Asp11	10–18	1L7O [91]
CDK2	1JST [76]	Thr160	152–163	1FIN [80]
ERK2	2ERK [92]	Thr183, Tyr185	172–186	1ERK [93]
FixJ	1D5W [94]	Asp54	79–99	1DCK [95]
Spo0A	1QMP [96]	Asp55	54–63	
GLM	1MKI	Ser74	64–75	
PMM	1K35 [97]	Ser108	105–114	
Pim1	2BIK	Ser261	251–262	
P38 $\gamma$	1CM8 [98]	Thr183,	177–189	
$\beta$ PGM	1LVH [99]	Asp8	7–16	
LCK	3LCK [100]	Tyr394	389–402	

**Table 2.3 Proteins Used in Our Test Set**

Unphosphorylated structures are listed only for the cases in which we attempt to predict the phosphorylated structure starting from the unphosphorylated structure. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the phosphorylated amino acid is not located within this region. We include this case to highlight the ability to extend our methods to regions larger than loops and their immediate surroundings.  $\beta$  PGM,  $\beta$ -phosphoglucomutase; GLM, probable glutaminase from *Bacillus subtilis*; LCK, human lymphocyte kinase; Pim1, proto-oncogene

serine/threonine protein kinase pim1; PMM, phosphomannomutase; Psp, phosphoserine phosphatase; Spo0A, stage 0 sporulation factor A.

In cases in which suitable unphosphorylated and phosphorylated structures of the same protein exist, we sought to extend this methodology to the more realistic situation in which we wish to predict the phosphorylated structure from the unphosphorylated structure. In these cases, we start from the crystal structure of the unphosphorylated protein, phosphorylate the residue of interest *in silico*, and predict the structure of the phosphorylated loop and its surroundings. We refer to this test as “loop prediction.” In this test, sampling only the conformation of the loop itself (Table 2.5, “loop only”) is successful in some cases, such as histidine-containing phosphocarrier protein (HPr), in which the conformational changes upon phosphorylation are small. In other cases, this strategy performs poorly because the phosphorylation induces substantial conformational changes in the surroundings of the loop, as discussed for CDK2 above. In order to capture these essential structural rearrangements in the region surrounding the loop, we also optimize the conformations of all side chains that have at least one atom within 4.5 Å of the loop, as described in Methods. This method permits all cases, excepting extracellular signal-regulated kinase 2 (ERK2), to be predicted with near-atomic accuracy, especially the phosphate group itself. Encouragingly, our molecular mechanics methods appear to be capable of not only predicting when significant conformational changes occur, as in CDK2, but also when phosphorylation induces little conformational change, as in HPr.

Protein	Loop Length	Loop Only				Loop + Surroundings			
		P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range
HPr	7	2.9	0.4	1.7	0.2–3.3	2.8	0.3	0.8	0.2–3.5
SpoIIAA	10	0.7(2.1)	0.3(0.4)	1.4	0.2–3.8	0.6(2.2)	0.3(0.8)	1.2	0.2–6.6
Psp	9	0.2	0.2	0.5	0.2–3.2	0.3	0.2	1.0	0.2–4.0
CDK2	12	0.6	1.7	3.5	1.2–8.4	1.7	1.8	1.7	1.0–12.5
ERK2	15	0.3/0.3	3.8	5.2	1.8–10.8	2.3/1.0	4.7	3.2	2.6–14.2
FixJ	8/10/5	0.1	0.5	1.2	0.4–4.7	0.3	0.8	1.0	0.4–4.7
Spo0A	10	0.3	0.2	0.8	0.2–3.1	0.8	0.4	0.9	0.2–4.1
GLM	12	0.1	0.2	1.0	0.2–6.2	0.2	2.1	1.6	0.4–5.8
PMM	10	0.3	0.4	1.0	0.3–5.4	0.3	1.0	1.1	0.4–5.9
Pim1	12	1.2	1.2	2.9	1.2–14.9	1.3	1.4	2.2	0.4–14.8
P38 $\gamma$	12	0.4/0.5	2.3	2.8	1.1–5.7	0.7/5.6	1.4	1.6	1.2–8.2
$\beta$ -PGM	10	0.9	0.6	1.3	0.2–7.1	0.6	0.2	0.6	0.2–8.8
LCK	14	1.2	3.0	3.3	1.6–10.1	2.5	2.7	2.4	1.8–11.9

**Table 2.4 Reconstruction of Phosphorylated Loop Conformations**

“Reconstruction” refers to the ab initio prediction of loop residues with the rest of the protein in the phosphorylated conformation. “Loop Only” refers to the prediction of only the loop in question, whereas “Loop + Surroundings” refers to the prediction of the loop in question in addition to surrounding side chains with at least one atom within 4.5 Å of any atom in the loop. The accuracy of the prediction is assessed by three measures: the error in the predicted phosphorus atom position (“P Error”), a backbone RMSD measure (“BB RMSD”), and the RMSD for all heavy atoms included in the prediction (“Heavy RMSD”). See the text for details. The “Sampled BB RMSD Range” column indicates the range of backbone RMSDs of sampled loops, emphasizing that our ab initio build-up procedure samples many different conformations of the loop to be predicted, and the final prediction is the lowest-energy conformation sampled. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the backbone and heavy atom RMSDs are calculated over the entire 23-residue region. For SpoIIAA, the predictions using a phosphate group with a -2 charge (as in the other test cases) gave poor results (in parentheses), whereas using a protonated phosphate group gave much better results; see text for details. Abbreviations used for the protein names are defined in Table 2.3.

ERK2 is unsuccessful because our existing sampling methods do not permit us to sample global structural changes like the changes in domain orientation that occur upon dual phosphorylation of ERK2 [9]. Given that we can accurately reconstruct the active conformation of the ERK2 activation loop, this inability to account for the domain reorientation is likely the cause of the poor prediction of the activation loop structure. This test case highlights a fundamental limitation of our current prediction strategy; the only solution is to expand the sampling to explicitly sample domain orientations, either in

a manner analogous to our helix prediction methods or by using normal modes or other methods capable of sampling low-frequency, global motions of the protein.

Protein	Loop Length	Loop Only				Loop + Surroundings			
		P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range
HPr	7	1.0	0.4	1.3	0.4–6.1	1.0	0.5	1.1	0.4–7.0
SpoIIAA	10	0.3(6.1)	0.7(5.7)	1.7	0.6–7.1	0.9(6.5)	0.8(3.9)	1.5	0.5–6.7
Psp	9	2.1	0.5	0.8	0.4–5.9	0.4	0.4	1.0	0.4–5.9
CDK2	12	8.8	6.7	8.1	3.2–9.9	0.8	2.9	3.6	2.0–11.0
ERK2	15	15.4/14.8	8.8	10.7	3.6–10.9	15.4/3.0	6.7	5.0	3.6–13.1
FixJ	8/10/5	1.3	1.1	1.5	0.4–4.8	1.0	1.6	1.7	0.4–4.8

**Table 2.5 Prediction of Phosphorylated Loop Conformations**

“Prediction” refers to the ab initio prediction of loop residues with the rest of the protein held in the unphosphorylated conformation. “Loop Only” refers to the prediction of only the loop in question; these results are provided only for comparison. “Loop + Surroundings” refers to the prediction of the loop in question in addition to surrounding side chains with at least one atom within 4.5 Å of any atom in the loop; these are the results that should be used in evaluating the success of our method. Definitions of the columns are provided in Table 4, and abbreviations used for the protein names are defined in Table 2.3. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the backbone and heavy atom RMSDs are calculated over the entire 23-residue region. For SpoIIAA, the predictions using a phosphate group with a -2 charge (as in the other test cases) gave poor results (in parentheses), whereas using a protonated phosphate group gave much better results; see text for details. ERK2, as described in the text, undergoes a domain reorientation as well as a loop movement, which is likely the cause of the poor predictions. The improvement of the CDK2 prediction when optimization of surrounding side chains is performed in concert with the loop prediction highlights the need to incorporate side chain flexibility in the calculation to successfully predict the phosphorylated loop conformation from the unphosphorylated structure.

The SpoIIAA case also requires explanation. The initial loop prediction and reconstruction tests on this case, using an unprotonated (-2) phosphate group as in the other cases, gave very poor results, listed in parentheses in Table 2.4 and Table 2.5. Examining the structure of the phosphorylated protein suggested a possible explanation. The pSer58 side chain lies within approximately 5 Å of the side chain of Asp56. The pKa of the phosphorylated amino acids is approximately 6, which implies that the predominant charge state is generally -2 under physiological conditions, and the good results we obtain for the other test cases using a -2 phosphate seem to confirm this assumption. However, it is well known that the pKa’s of titratable groups can be



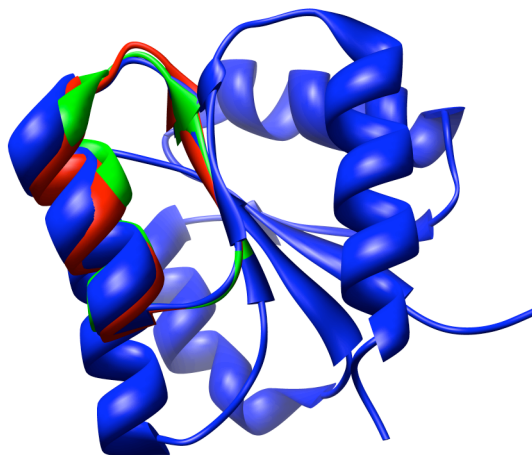
significantly shifted in macromolecules due to desolvation and/or the electrostatic field from the rest of the protein. In this case, the close proximity of Asp56 is likely to shift the pKa lower and favor the  $-1$ , protonated state of the phosphate group on pSer58 (the pH used in the crystallization conditions is 6.5). In fact, performing the loop predictions with a protonated phosphate group leads to excellent results. In current work, we are implementing a new version of our molecular mechanics algorithms that allows automatic identification of the optimal protonation state for phosphorylated amino acids and other titratable groups (especially His), using a thermodynamic cycle analogous to that employed by the numerous Poisson-Boltzmann based pKa prediction algorithms [101].

#### ***2.2.6. Predicting Conformational Changes of Helices in Response to Phosphorylation***

To test our ability to model conformational responses that extend beyond a single loop and its surroundings, we consider the repacking of helices that is known to occur upon phosphorylation of response regulator proteins of prokaryotes [9]. The response regulator proteins are the second part of the two-component signaling system that bacteria use to sense and respond to their environment. In most two-component systems, a sensor histidine kinase autophosphorylates a His upon sensing signal and transfers this phosphate to an Asp on a response regulator. Upon phosphorylation, certain structural changes occur, most notably a helix shift and loop rearrangement, that ultimately allow the response regulator to activate transcription of a set of genes.

We used our existing helix prediction algorithm to model the structural changes of the response regulator protein FixJ (from *Sinorhizobium meliloti* [94]), specifically, the

loop-helix-loop region that undergoes a helix shift and loop rearrangements. Another response regulator protein, Spo0A, has crystal structures in both the phosphorylated and unphosphorylated forms, but the crystal structure for the unphosphorylated form of Spo0A is a domain-swapped dimer, solved at a pH of 4.5. As discussed in Methods, the helix prediction algorithm employs rigid body sampling for the helix (residues 87–94) combined with the hierarchical loop sampling algorithm for predicting the connecting loops [74] (residues 79–86 and 95–99). The overall backbone RMSD for reconstructing this region in the phosphorylated crystal structure is 0.5 Å, with most of the error arising from the first flanking loop, which is longer, closest to the phosphate, and more exposed to solvent (Table 2.4 and Figure 2.8). The results from *in silico* phosphorylation of the unphosphorylated structure are also quite good, with an overall RMSD of 1.6 Å.



**Figure 2.8 Helix Reconstruction and Prediction of FixJ**

The loop-helix-loop region of FixJ was predicted starting from either the phosphorylated structure (blue) or the unphosphorylated structure (unpublished results). The reconstruction (starting from the phosphorylated structure) is in red, and the prediction (starting from the unphosphorylated structure) is in green. Figure prepared with Chimera [81].

## 2.3. Conclusion

We have described our initial efforts to predict and understand how the structure of a protein is modulated by post-translational phosphorylation. We believe that this work has practical significance in that we demonstrate that it is possible to make testable predictions concerning the structure of phosphorylated proteins, given the structure of the unphosphorylated protein and a known site of phosphorylation. In this work, we have restricted our efforts to predicting relatively modest, localized conformational changes, and we have assumed knowledge of which portions of the protein undergo significant conformational change (those portions closest to the phosphorylated amino acid). Despite these limitations, this modeling technology can be used to create hypotheses about mechanisms of regulation by phosphorylation that can then be tested experimentally, e.g., by site-directed mutagenesis. Applications of this type are underway. In addition, our in-depth case study of CDK2 illustrates how the computational methods can be used to obtain new insights into the energetic underpinnings of phosphorylation-induced conformational change even in a well-studied system.

## 2.4. Methods

### 2.4.1. *Dataset selection*

We searched the PDB [102] for phosphorylated protein structures determined by X-ray crystallography (with better than 2.5 Å resolution) that are phosphorylated on well-ordered loop structures that were less than 15 residues in length. We exclude structures in which phosphorylation causes a large, global rearrangement of the protein structure,

such as a hinge-bending movement or domain rearrangement, as is the case with glycogen phosphorylase [103] and insulin receptor tyrosine kinase [104,105]. In order to test the limitations of our method, we include one test case, ERK2, in which phosphorylation causes a domain rearrangement [9]. We also include a prokaryotic response regulator, FixJ, in which phosphorylation of an Asp induces a significant conformational change in the orientation of a helix [94,95]; this case is successfully treated with an extension of our methods described below. The test set is listed in Table 2.3.

We determined the loop length that we would predict for each phosphorylated structure using visual inspection. In the cases in which crystallographic structures are available for both the unphosphorylated and phosphorylated protein, these structures were superimposed and the residues to be predicted were defined as the portion of the loop that deviated in the superposition. For the reconstruction of phosphorylated structures without knowledge of the unphosphorylated form, the loop residues to be optimized were determined using a combination of visual inspection of secondary structure and crystallographic B-factors.

#### **2.4.2. *Molecular mechanics energy function***

All energy calculations use the OPLS-AA force field [69,106,107] and the Surface Generalized Born (SGB) model of solvation [14,15]. The molecular mechanics energy function represents electrostatics by a relatively simple model of fixed atomic partial charges interacting through the Coulomb approximation. The solvent model captures key effects of desolvation with relatively modest computational expense. Despite the simplicity of the energy function (i.e., it neglects polarizability contributions

to electrostatics, and implicit solvent models have well-known limitations), it performs well in predicting conformations of phosphorylated loops (see Results).

The force field parameters for the phosphorylated amino acids were generated by an automated atom-typing algorithm provided in the Impact software package. The atomic partial charges for the phosphorylated amino acid side chains were adjusted slightly from the default values by performing quantum chemistry calculations. The partial charges for phosphoserine (pSer) and phosphothreonine (pThr) were taken from previous work by Wong et al. [18], whereas charges for phosphotyrosine (pTyr) and phosphoaspartate (pAsp) were determined in their  $-2$  and  $-1$  charge states by performing quantum mechanical calculations with the software program Jaguar [108]. Methylbenzyl-phosphate was used to represent the pTyr side chain, and acetyl phosphate was used for pAsp. Geometry optimization of the phosphate ion was carried out at the HF/6-31G\*\* level, incorporating a condensed-phase environment via a self-consistent reaction field (SCRF) algorithm [109,110]. Single point calculations were performed at the LMP2/cc-pvtz(-f) level, also with SCRF treatment of solvation. Electrostatic potential fitting was used to determine the partial charges. The atomic partial charges for all four phosphorylated amino acids are provided in Tables S1–S4.

### **2.4.3. *Loop prediction methodology***

This study uses the method of Jacobson et al. [72] for predicting loop conformations. In brief, the loop prediction methodology uses an ab initio dihedral sampling scheme to enumerate conformations of the loop backbone that are free from steric clashes. Other methods have employed similar dihedral angle sampling schemes, including ICM [111], CONGEN [112], and the work of DePristo, et al [113]. Unlike

Monte Carlo and Molecular Dynamics sampling schemes, the algorithm itself has no knowledge of the starting conformation of the loop, and therefore, does not start predictions based on a starting structure of the loop. These closed backbone conformations are clustered, and a single member of each cluster is then selected for side chain addition and optimization, followed by complete energy minimization. The lowest energy structure is selected as the output of the loop prediction algorithm. The algorithm also permits explicit treatment of crystal packing. We have used this capability in this work (in the “loop reconstruction” cases, as described below), but did not identify any clear crystal-packing artifacts relevant to the conformations of the phosphorylated loops.

The Jacobson et al. paper [72] describes a hierarchical refinement procedure in which multiple iterations of the loop prediction algorithm are used to reduce errors caused by insufficient sampling. The parameters in this scheme have been slightly modified in this study due to the inclusion of some rather long loops (up to 15 residues) in our test set. The first stage allows for unrestrained sampling. After this stage is complete, the top ten lowest energy structures are passed to a first refinement stage in which more extensive sampling is performed around these low-energy basins. Specifically, loop conformations in this stage are only retained if all of the  $C\alpha$  atoms in the loop are within 10 Å of the starting loop structure (which is one of the ten lowest-energy structures from the initial stage). The five resulting lowest-energy structures from this stage are subjected to a second round of refinement, in which the maximum  $C\alpha$  deviation is restricted to 5 Å. Finally, the five lowest-energy structures from this stage are subjected to a third and final round of refinement, in which the maximum  $C\alpha$  deviation is restricted to 2.5 Å. In all, this procedure provides a rank-ordered list

consisting of about 250 loops and their associated energies in the context of the full protein. The final prediction is the loop conformation with the lowest energy.

Most published tests of loop prediction methods, including the Jacobson et al. paper [72], evaluate their success by the ability to reconstruct loops in a native protein structure. In these tests, all portions of the protein other than the loop in question remain in the native conformation during the simulation. Success of a prediction methodology in such a test is an important prerequisite for more realistic applications. We perform loop reconstruction tests in this work to assess the ability of the molecular mechanics energy function to identify correct conformations of phosphorylated loops, and to ensure that the sampling methods are sufficient to generate near-native conformations. However, predicting conformational changes induced by phosphorylation, i.e., by phosphorylating a protein *in silico*, is qualitatively more challenging. In the cases we consider, the sites of phosphorylation are located on loops, and most of the conformational change is localized to that loop. However, there is always some degree of conformational rearrangement in the vicinity of the loop, especially in the conformations of side chains contacting it. Similarly, predicting the conformation of a loop in a homology model is more challenging than reconstructing a loop in a native protein structure because the environment surrounding any given loop in a homology model contains errors that can affect the loop prediction accuracy. We address this issue by performing rotamer optimization and minimization of side chains in the immediate vicinity of the loop concurrently with the optimization of the side chains on the loop itself. In the test set presented here, this strategy performs well in predicting local structural changes, despite the fact that there are also some changes in backbone conformation in the surroundings.

We speculate that small changes in the conformations of the surrounding side chains can compensate for not explicitly allowing backbone relaxation. We also use this strategy in the control studies of reconstructing phosphorylated loops in which, interestingly, it sometimes improves the accuracy.

#### **2.4.4. *Helix prediction methodology***

The helix prediction algorithm used in this paper is based on the work of Li et al [74]. Briefly, the helix backbone is treated as a rigid body and sampled in six degrees of freedom (three translations and three rotations), and the two flanking loops are sampled using the loop prediction algorithm described above. Again, the method broadly samples the possible configurations of the helix and surrounding loops, independent of any starting configuration, and then hierarchically samples more finely around low-energy basins. As with the loop prediction algorithm, side chains on the loop-helix-loop region, and the surroundings if desired, are sampled using a rotamer-based optimization algorithm.

#### **2.4.5. *Development of “rotamer” libraries for phosphorylated residues***

The loop and helix prediction algorithms require the use of a rotamer library for sampling side chain conformations. However, the number of phosphorylated residues in the PDB is insufficient to construct these libraries by statistical analysis of observed side chain conformations. Instead, we obtained a rotamer library by exhaustively exploring the energy landscape of the side chains of phosphorylated amino acid dipeptides, and retaining all conformations below an energy threshold that ensures inclusion of all experimentally observed rotamers. We use the same energy function as in the rest of the



work, i.e., OPLS-AA/SGB. The pSer and pThr side chains are sampled at 10° resolution, whereas pTyr and pAsp are sampled at 30° resolution, due to the larger number of rotatable bonds. The total numbers of rotamers are 802 for pSer, 501 for pThr, 858 for pTyr, and 1,008 for pAsp. These values do not include the rotation of the phosphate group about the phosphoester bond, which is sampled uniformly at the same resolution as the other bonds.

# **Chapter 3 Energy-based analysis and prediction of the orientation between antibody light and heavy variable domains**

As discussed in Chapter 2, our predictions of changes in loop conformation upon phosphorylation were largely successful, with one primary exception - the case of predicting the conformational change upon phosphorylation in the mitogen-activated protein kinase ERK2 starting from the unphosphorylated structure. The error in this prediction was attributed to the fact that the unphosphorylated and phosphorylated structures differed not only in their loop conformation, but also in the relative orientation of the N- and C-terminal lobes of the kinase [92]. The change in lobe orientation between the inactive (often unphosphorylated) and active (often phosphorylated) states of protein kinases is not unique to ERK2, and is seen in crystallographic studies of other protein kinases as well (see [8,9,114]). Consequently, we sought to develop a method to predict the orientation of the protein kinase lobes. To address this problem, we turned to antibodies as a model system. The remainder of this chapter presents published work representing the beginnings of a method for the prediction of the relative orientation of

the heavy and light chain variable domains in antibodies, the N- and C-terminal lobes of protein kinases, and protein domains in general. This work was published with Ben Sellers, who 1) developed many of the coding structures upon which I built the methods used below, 2) created the main dataset, 3) built the multiple sequence alignments, and 4) engaged in extremely helpful discussions regarding antibody structure and aided in manuscript preparation. In addition, Matt Jacobson supervised and directed this research and co-wrote the manuscript.

### 3.1. Introduction

The sequence diversity in antibodies generated by the immune system is crucial for recognizing the almost infinite set of potential antigens. The overall fold of each antibody variable domain is well conserved, and the differences in sequence and structure within each variable domain are largely localized to the complementarity-determining regions (CDRs), six hypervariable loops presented on this structurally conserved framework that are critical for antigen binding. However, structural diversity can also be generated via differences in association between the variable domains of the heavy ( $V_H$ ) and light ( $V_L$ ) chains [115,116,117]. Since the paratope used for antigen binding is formed at the interface of the light and heavy chain variable domains, the relative orientations of these domains can significantly impact the antigen-binding properties of antibodies.

Several lines of evidence support the idea that the  $V_L:V_H$  orientation can affect antigen-binding properties, following the early observation by Stanfield, et al. of large antigen-induced changes in the relative orientation of variable domains [116]. One observation comes from antibody humanization efforts in which CDR loops from a high-

affinity murine antibody are grafted onto a human framework. Banfield et al. reported a humanized antibody with a 2-fold change in affinity in which there were minimal conformational changes in the CDR loops [118]. Instead, a significant difference in the  $V_L:V_H$  orientation resulted in large changes in the relative positions of the CDR loops, which was suggested to modulate the affinity for antigen. Nakanishi et al. also reported increases in binding affinity in a humanized antibody using domain interface mutations [119]. Further evidence is seen in biochemical experiments that found that alanine mutations of specific interface residues affected both affinity and specificity [120,121]. Since mutations at the  $V_L:V_H$  interface could have significant effects on the  $V_L:V_H$  orientation, these studies suggest that changes in domain orientation may alter antigen binding.

Variation in the relative orientation between domains has also been hypothesized to affect functional properties in other systems. Related members of the enolase superfamily have been observed to have different orientations between the C-terminal beta-barrel and N-terminal domains, and this has been hypothesized to play a role in their differing specificities [122]. The orientations between the C-terminal and N-terminal lobes of the catalytic domain of protein kinases can be altered upon phosphorylation [92,123], which may modulate binding of ATP as well as kinase inhibitors [124,125]. Antibodies represent a model system for the prediction of relative orientations between protein domains because of a) the well conserved framework of the individual domains, b) the ability to use knowledge-based rules to accurately model 5 of the 6 loops [126,127,128], c) the lack of a linker between the domains, and d) the availability of many crystallographic structures.

Antibodies have emerged in recent years as an important class of therapeutics, particularly for treating a variety of cancers [129,130,131]. Antibody humanization and phage-display technologies have played key roles in engineering therapeutic antibodies, but consideration of the three-dimensional structure of an antibody can also aid antibody engineering efforts [129,132,133,134,135]. Obtaining crystal structures can be expensive and time consuming. Homology modeling methods can be used to generate 3-D models of antibodies, and have also been used in structure-based rational design efforts [136,137,138], but the accuracy of the comparative models will directly affect the quality of hypotheses generated from them. Most of the work in comparative modeling of antibodies has focused on modeling the CDR loops [139,140,141,142]. However, large differences in the relative orientation of the variable domains between different antibodies would require that the preferred  $V_L:V_H$  orientations of the target antibody sequence be explicitly considered, as has recently been done [143].

Because the  $V_L:V_H$  orientation appears to be a functionally important manifestation of antibody sequence diversity, one would expect to see significant variation in the orientation among existing antibody crystal structures. While there has been some analysis of the interactions present at the interface of the light and heavy chain variable domains [117,144], there have been few studies analyzing variability in  $V_L:V_H$  orientation in different antibodies. Knappik, et al. note variability in orientation, particularly in antibodies containing  $V\lambda$  light chains [115]. Recently, Gray and colleagues clustered antibody structures and saw distinct clustering of antibodies containing  $V\lambda$  and  $V\kappa$  light chains [143].

Here, we first reexamine the diversity of  $V_L:V_H$  orientations in existing antibody crystal structures using 153 non-redundant sequences and find that the orientation can vary widely, confirming that the relative orientation of the heavy and light chain variable domains are a source of structural diversity in the set of possible antibodies. Through further structural analysis and molecular mechanics calculations, we show that this variation in the  $V_L:V_H$  orientation is sequence-dependent, and is not purely an artifact of crystal packing or antigen binding. This suggests that the orientation is encoded in sequence and can be predicted. The molecular mechanics calculations confirm that the observed orientations are close to the minimum energy, and define the features of the energy landscape governing the orientation of the heavy and light chain variable domains. We also extend this method to predict  $V_L:V_H$  orientation using comparative models of the individual domains with the aim of aiding the development of more accurate structural models of antibodies.

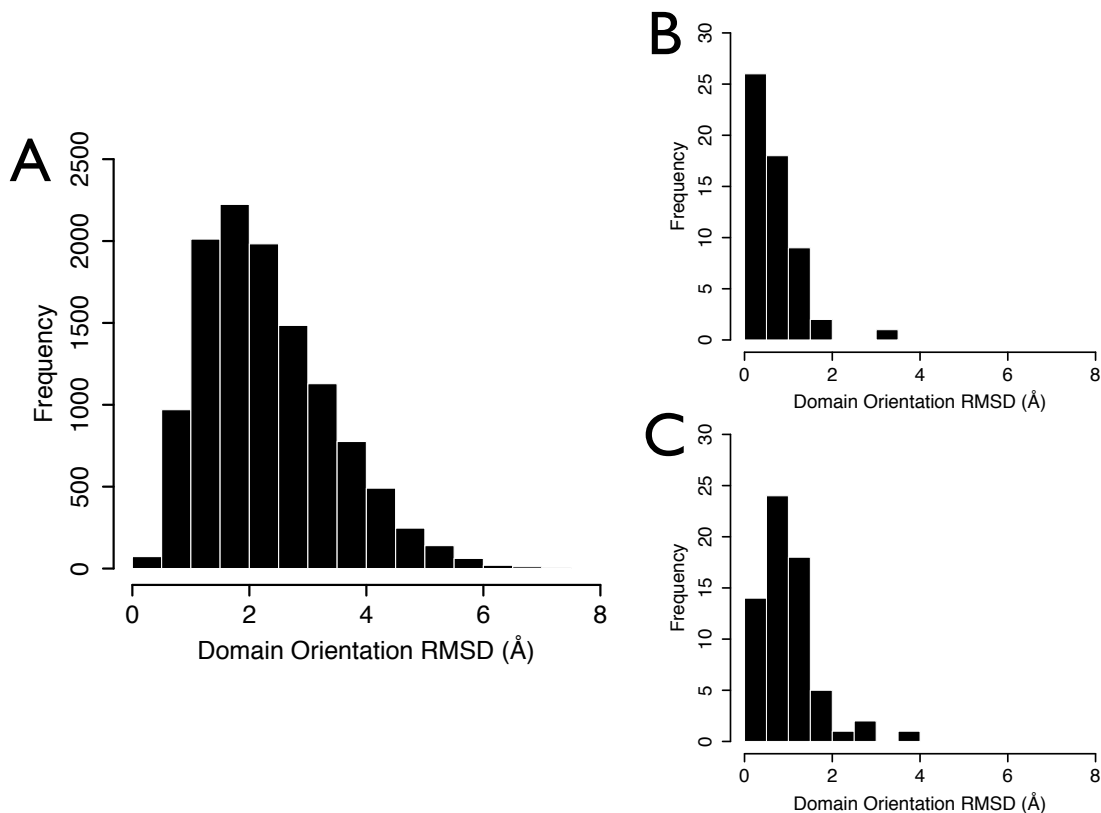
## 3.2. Results

### ***3.2.1. Extent of Variation of $V_L:V_H$ Orientation in Experimentally-Observed Antibody Structures***

To examine the extent of variability in the  $V_L:V_H$  orientation present in existing crystal structures of antibodies, we created a dataset of 300 antibody structures (Supplementary Table B1) of which a subset of 153 structures (Supplementary Table B2) was selected as a non-redundant set (filtered at 90% pairwise sequence identity) to be used for structural analysis (see Methods). The difference in domain orientation for each pair of structures was calculated by first aligning the heavy chains followed by optimal

superposition of the light chains. The difference in orientation was defined as the non-CDR alpha-carbon RMSD calculated between the light chain of one of the structures in its original and superimposed orientations. This approach ensures that conformational differences within the domains do not contribute to the RMSD.

We found the differences in domain orientation to be moderate overall, with a mean of 2.3 Å RMSD, consistent with a recent study [143]. However, as evidenced by a standard deviation of 1.1 Å, the differences in orientation between antibodies can often be quite large (Figure 3.1A). Because the differences in domain orientation lead to larger atomic deviations than are caused by differences in structure within the individual domains (~1 Å RMSD for non-CDR regions), the  $V_L:V_H$  orientation is an important source of structural diversity in antibodies. The number of pairs (out of 11628) exhibiting differences in domain orientation of 3.0, 4.0, and 5.0 Å RMSD or greater were 2,907 (25%), 985 (8%), and 238 (2%), respectively, with a maximum difference of 7.4 Å. While these pairs represent a minority of the data points, they reinforce the idea that the heavy and light chain variable domains can take on significantly different orientations between different antibodies.



**Figure 3.1 Distribution of differences in domain orientation**

Distribution of domain orientation differences calculated for all pairs of 153 antibodies with  $\leq 90\%$  identity over the Fv sequence (A), for 56 pairs of structures of unbound antibodies with identical sequence (B), and for 65 pairs of structures of antibodies with identical sequence, with each pair containing one unbound structure and one antigen-bound structure (C).

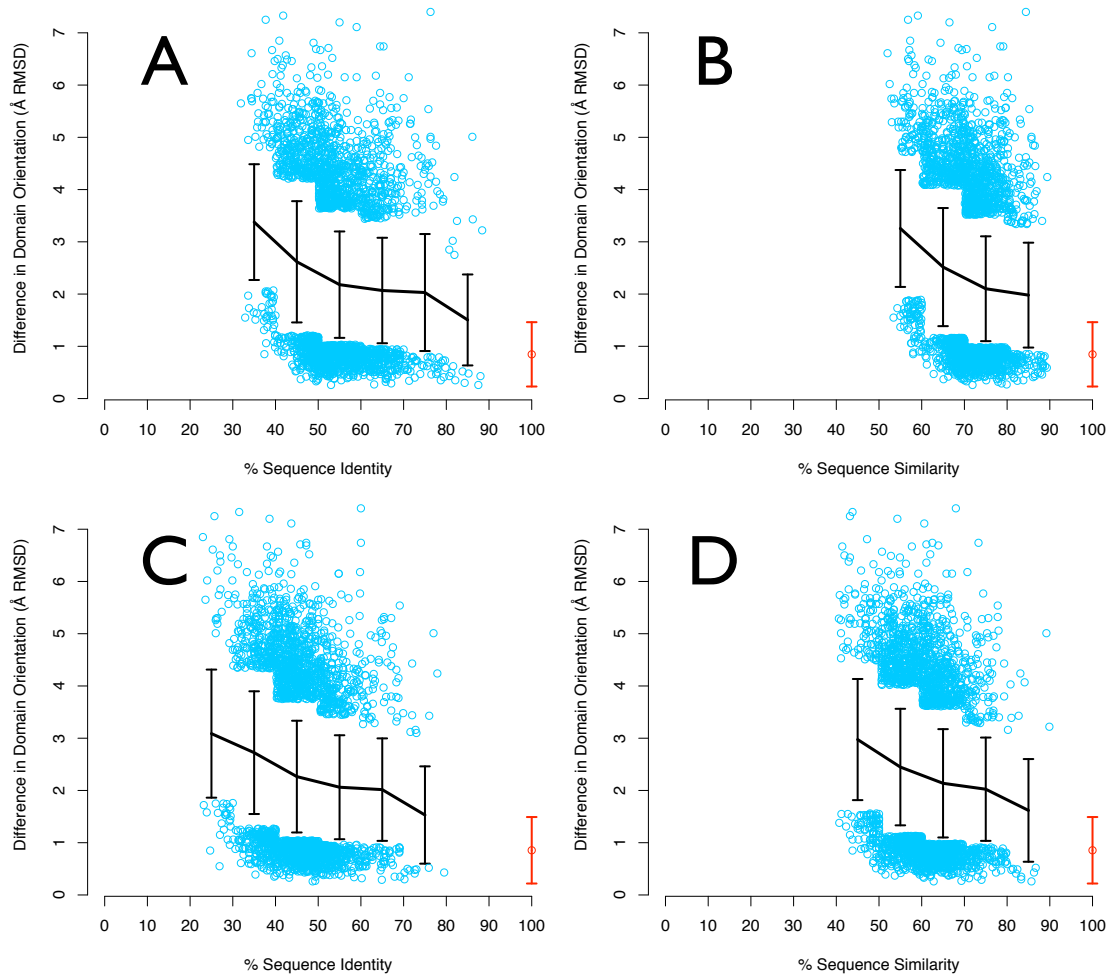
Crystal packing and conformational changes arising from antigen binding may give rise to some of the observed variation in orientation. We first examined the differences in domain orientation of 56 antibody pairs of identical sequence crystallized in the absence of ligand (Figure 3.1B). We found that the mean orientation difference was  $0.7 \text{ \AA}$  with a standard deviation of  $0.5 \text{ \AA}$ . We also compared the bound and unbound structures of 65 antibody pairs of identical sequence, as ligand binding has been shown to induce substantial rearrangements of the relative domain orientation [116]. The mean orientation difference including conformational changes due to antigen binding is  $1.0 \text{ \AA}$ , with a standard deviation of  $0.6 \text{ \AA}$  (Figure 3.1C). These smaller orientation differences in



complexes with identical interface sequences indicate that the larger differences observed in the non-redundant dataset are mostly due to sequence differences and only a relatively small amount of the variation is likely attributable to effects of crystal packing or antigen binding, though antigen binding does induce large changes in orientation in several cases (Supplementary Table B3-B6).

### ***3.2.2. Correlation of Differences in $V_L:V_H$ Orientation with Differences in Sequence***

We next investigated whether the variation between antibodies in our dataset was correlated with simple measures of the relatedness of their sequences. First, we calculated the sequence identity and similarity over the full length of both the light and heavy chain variable regions for all pairs of antibodies. The difference in domain orientation for each pair was plotted against the sequence identity or similarity (Figure 3.2). The mean difference in orientation is qualitatively correlated with both sequence identity and similarity. However, the range of observed orientations seen at a given sequence identity (Table 3.1) or similarity is large; even for pairs with 70-80% identity, 17% have 3.0Å RMSD or greater, and between 40 and 50% identity that number grows to 35%.



**Figure 3.2 Mean difference in domain orientation is correlated with simple measures of difference in sequence, but there is large variability at a given identity or similarity.**

The difference in domain orientation (as measured by non-CDR C-alpha RMSD, see Methods) for a non-redundant set of 153 antibodies is plotted against the combined sequence identity (a) or similarity (b) over both the heavy and light chain sequence; or the sequence identity (c) or similarity (d) over all residues that make inter-chain contacts in any of the structures in the dataset. Each data point was binned in 10% increments of sequence identity or similarity, and the mean was calculated. A line was plotted for the means of all bins containing greater than 50 data points (black line). Data points that were in the lowest or highest 10% of domain orientation differences are plotted as points (blue). The mean and standard deviations of comparisons involving identical antibodies (over the whole Fv region (A,B) or the interface residues (C,D) from different crystal structures (both unbound-unbound pairs and unbound-bound pairs) is shown in red at 100% identity or similarity.

Distribution of Orientation Differences By Sequence Identity of Fv Residues						
Range(%)	% of Pairs	Median	% < 1Å	% < 2Å	% ≥ 3Å	% ≥ 4Å
90-100	0.0	N/A	N/A	N/A	N/A	N/A
80-90	0.8	1.3	31.8	78.4	6.8	2.3
70-80	3.4	1.8	17.0	57.1	17.2	6.5
60-70	19.1	1.9	11.8	54.0	17.3	5.1
50-60	45.0	2.0	9.2	49.6	20.0	6.0
40-50	28.2	2.5	5.1	33.8	35.4	12.7
30-40	3.4	3.3	0.5	8.8	60.0	27.8
20-30	0.0	N/A	N/A	N/A	N/A	N/A
10-20	0.0	N/A	N/A	N/A	N/A	N/A
0-10	0.0	N/A	N/A	N/A	N/A	N/A
Distribution of Orientation Differences By Sequence Identity of Fv Interface Residues						
Range(%)	% of Pairs	Median	% < 1Å	% < 2Å	% ≥ 3Å	% ≥ 4Å
90-100	0.0	N/A	N/A	N/A	N/A	N/A
80-90	0.1	1.0	54.5	90.9	9.1	0.0
70-80	0.7	1.3	28.6	79.8	9.5	3.6
60-70	4.9	1.9	13.1	55.8	15.0	5.1
50-60	20.8	1.9	11.5	54.5	17.2	4.8
40-50	51.5	2.1	8.6	46.7	23.4	7.5
30-40	19.9	2.6	4.6	30.1	38.5	15.0
20-30	2.1	2.9	1.2	15.2	43.2	16.9
10-20	0.0	2.8	0.0	0.0	50.0	0.0
0-10	0.0	N/A	N/A	N/A	N/A	N/A

**Table 3.1 The differences in orientation between antibodies are more pronounced at lower sequence identities.**

The percentage of antibody pairs exhibiting differences in orientation less than 1 Å, less than 2 Å, more than 3 Å, and more than 4 Å RMSD are shown for ranges of sequence identity calculated over all Fv residues (top) or all Fv interface residues (bottom), along with the median and the percentage of total antibody pairs.

Since the residues at the interface of the two domains are expected to play a larger role than non-interface residues in determining the  $V_L:V_H$  orientation, we examined whether differences in domain orientation were correlated with sequence identity and similarity of the interface residues. Once again, the mean values of differences in orientation are correlated with the sequence identity or similarity of the pairs, and there is large variation in orientation differences at a given identity or similarity (Figure 3.2C and

D). Among pairs of antibodies with 60-70% identity in the interface, 23% have differences in orientation  $>3.0\text{\AA}$ ; for pairs with 20-30% identity, 46% have differences in orientation  $>3.0\text{\AA}$  (Table 3.1).

These data support the hypothesis that the domain orientation of an antibody is largely encoded by its sequence. However, there is a large amount of variation in the orientation at any given sequence identity, suggesting that even at relatively high sequence identity, there are not tight bounds on the difference in orientation of a pair of antibodies.

### **3.2.3. *Energy-Based Prediction of $V_L:V_H$ Orientation***

The variation in orientation observed in the pairwise comparison of different antibody structures led to the hypotheses that the  $V_L:V_H$  orientation is determined by contacts made by interface residues, and that the orientation of an antibody represents the global minimum of the energy landscape. In order to test these hypotheses, we developed a tool for predicting the relative orientations of the heavy and light chain variable domains based on molecular mechanics-based energy calculations. Due to the fact that— at any given sequence identity or similarity—the variation in domain orientation is large, knowledge-based methods for predicting the orientation that are based on simple sequence-based metrics may have limited utility. Our approach is to sample  $V_L:V_H$  orientations found in existing crystal structures and evaluate the energy of these orientations after optimizing side-chain rotamers for residues at the interface of the variable domains. The energy of each orientation is evaluated with a molecular mechanics force field (OPLS [69,106,107]) in conjunction with an implicit solvent model (Generalized Born [14,15]). Although this energy function has some limitations, it has

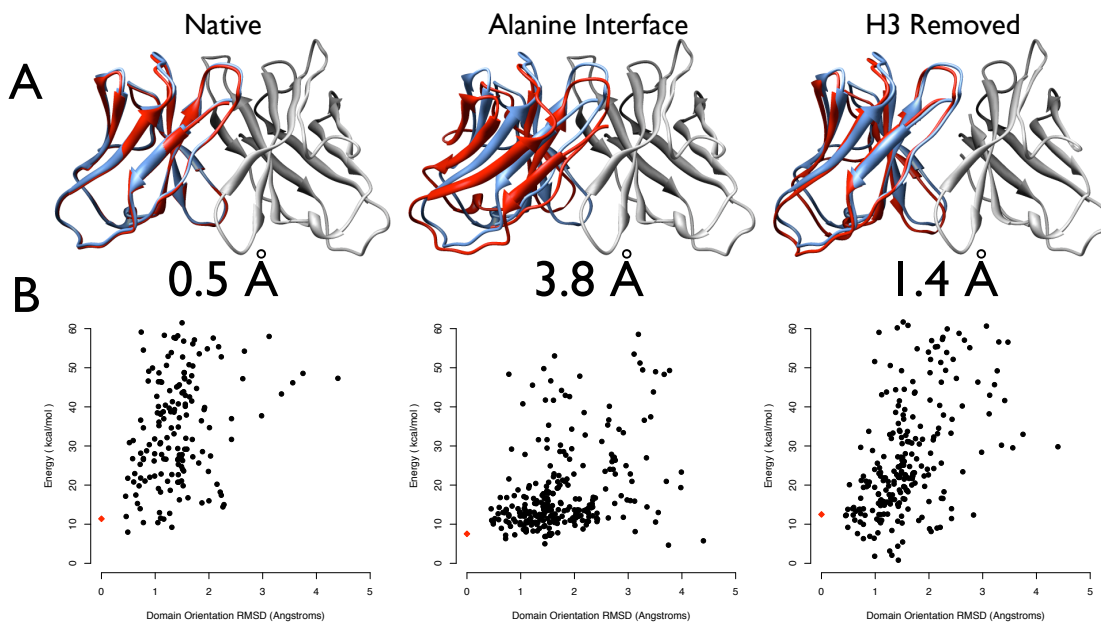
previously been shown to perform well in predicting protein side chain and loop conformations [69,70,72,145], as well as the positions and orientations of helices in proteins [74].

We illustrate the results using a case study of a humanized variant of anti-p185<sup>HER2</sup> antibody 4D5 [146] (PDB [102] ID: 1fvc) and provide summary statistics on a larger dataset consisting of twenty antibodies. While this is clearly a small test set relative to the number of existing antibody crystal structures, the mean, median, and standard deviations of 2.6, 2.4, and 1.2 Å RMSD difference in orientation between structures in this test set suggest that this set represents a reasonably diverse set of orientations.

#### *3.2.3.1. Native-like $V_L:V_H$ orientations correspond to the lowest energy among decoys*

In the first set of calculations, we placed the heavy and light chain variable domains of each antibody in 300 different orientations, obtained from 300 existing crystal structures. The conformations of all side chains in the interface (see Methods) were subjected to rotamer optimization and energy minimization, prior to calculating the energy of each orientation. The results show that this method selects either the native or a near-native conformation as the lowest-energy. For the 1fvc case study, the lowest energy orientation differs from the native by 0.5 Å RMSD (Figure 3.3A), and the energy landscape is funnel-like (Figure 3.3B), i.e., there is a strong correlation between energy and RMSD from native. Table 3.2 shows that this result is general across the test set of twenty antibodies using the same protocol. A native or near-native (<1 Å RMSD) orientation is selected in all but 4 cases, with an average deviation from the native orientation of 0.6 Å. When excluding the crystallographic conformation, the average

RMSD increases to 0.9 Å. This provides further evidence that the orientations observed in antibody crystal structures are encoded by sequence, and represent the global minimum in the energy landscape. It also demonstrates that it is possible to predict antibody  $V_L:V_H$  orientations using sampling and energy calculations with good accuracy. A more realistic and challenging test, using homology models of the domains, is discussed below.



**Figure 3.3 Near-native orientations of the light and heavy chain variable domains represent global energy minima using a molecular mechanics scoring function.**

Orientations and energy landscapes predicted using unmodified crystal structures (Native) or with all interface residues mutated to alanine (Alanine Interface) or with the H3 loop removed (H3 removed) prior to prediction. (a) The predicted orientation of the light chain (red) vs. the native orientation (blue) of the 4D5 antibody (PDB ID 1fvc) and the corresponding domain orientation RMSDs. The crystallographic heavy chain is shown in gray in all cases. (b) Predicted energy vs. domain orientation RMSD plots of the decoy orientations sampled. The native orientation is shown as a red square on the y-axis.

### *3.2.3.2. Interface Side-Chain Contacts Define the Energy Landscape Governing $V_L:V_H$ Orientation*

As a control, when the same protocol was applied after all of the interface residues were mutated to alanine, the lowest-energy orientation for 1fvc was far from the crystallographic orientation, with a difference in orientation between the predicted and crystallographic orientations of 3.8 Å RMSD (Figure 3.3A). Furthermore, while the energy landscape of the native sequence shows a distinct funnel-like shape, the energy landscape of the sequence with interface residues mutated to alanine results in a much flatter energy landscape with little preference for orientation (Figure 3.3B). In the other cases of our test set, we also found that mutating the interface residues to alanine generally results in prediction of a non-native orientation (Table 3.2) and/or a flattened energy landscape (Supplementary Figure B1) with respect to the relative orientation of the heavy and light chain variable domains.

### *3.2.3.3. The role of contacts involving the H3 loop in determining the $V_L:V_H$ orientation*

The H3 loop is the most variable CDR loop in terms of length, sequence, and structure, and is located at the interface of the heavy and light chain variable domains where it is involved in a large number of inter-chain contacts and is hypothesized to play a key role in the association of the variable domains [115,117]. As discussed above, crystal structures of identical antibodies with and without antigen can show some differences in domain orientations and, although no causal effect has been established, the length of the H3 loop has been linked to these domain orientation changes [116]. We tested the hypothesis that the H3 loop may be essential to determining the  $V_L:V_H$  orientation.

PDB	Range	Mean (Std. Dev.)	Including crystal conformation	Excluding crystal conformation	Alanine Interface	H3 Removed
1a6u	0.0-5.5	2.2(0.8)	0.8	0.8	1.5	1.6
1a7n	0.0-5.6	2.3(1.2)	0.2	0.2	0.8	0.2
1bvl	0.0-5.2	2.4(1.0)	1.0	1.0	2.1	1.2
1dlf	0.0-7.4	4.4(1.2)	0.6	0.6	5.7	0.6
1dql	0.0-4.3	1.9(0.8)	0.8	0.8	4.2	0.9
1dsf	0.0-5.2	2.6(0.8)	0.0	1.0	0.0 (1.6)	2.5
1fvc	0.0-4.5	1.8(0.8)	0.5	0.5	3.8	1.4
1igm	0.0-4.6	2.0(0.9)	0.6	0.6	3.6	1.6
43c9	0.0-5.0	2.0(0.8)	0.0	0.3	1.3	0.7
1mam	0.0-4.6	2.3(0.8)	1.7	1.7	3.4	1.7
1dfb	0.0-7.1	3.4(1.3)	0.0	1.9	3.0	2.3
1m7d	0.0-5.6	2.9(1.0)	0.6	0.6	2.1	2.7
1mnu	0.0-4.6	2.3(1.0)	1.3	1.3	1.8	0.5
1nca	0.0-4.9	2.6(1.1)	0.7	0.7	0.7	0.8
1a14	0.0-5.8	3.0(1.2)	0.0	1.5	1.6	1.3
1dn0	0.0-5.0	2.2(1.1)	0.0	0.4	0.8	1.2
1i9j	0.0-4.8	1.9(0.9)	0.4	0.4	1.3	0.6
1bz7	0.0-4.8	2.2(0.8)	1.0	1.0	2.5	2.9
1ai1	0.0-5.1	2.7(1.0)	1.5	1.5	1.0	1.9
1fgn	0.0-6.3	2.9(1.3)	1.8	1.8	3.1	0.0 (2.8)
Average		2.5(1.2)	0.7	0.9	2.2 (2.3)	1.3 (1.5)

**Table 3.2 The crystallographic orientation of light and heavy chain antibody variable domains is defined by the interface sequence and is at or near the global minimum of the energy landscape.**

For each of the twenty antibodies shown, we sampled orientations of the heavy and light chain variable domains from those observed in our full 300-antibody dataset. The difference in orientation was calculated between each of these orientations and the crystallographic orientation. The range, the mean, and the standard deviation of differences between the native orientation and the sampled orientations are shown (the crystallographic orientation is included, so the minimum RMSD is 0.0). Three calculations were performed, one where we tried to reproduce the crystallographic orientation without modifying the individual domains (Native), one where we mutated all of the interface residues in both domains to alanine (Alanine Interface), and one where we removed the H3 loop prior to sampling and scoring (H3 Removed). The differences in orientation between the predicted (lowest-energy) and crystallographic orientations are shown. For the calculations using the native variable domains, the crystallographically observed  $V_L:V_H$  conformation was selected by energy in many cases (i.e., 0.0 RMSD); results excluding the crystallographic conformation are also shown (“excluding crystal conformation”). When the interface was mutated to alanine or the H3 loop was deleted, the crystallographic conformation was chosen only in 1 case, and the lowest energy conformation excluding the native is shown in parenthesis.

We removed the H3 loop from the 1fvc heavy chain and again sampled domain orientations. The difference in orientation between the predicted and crystallographic



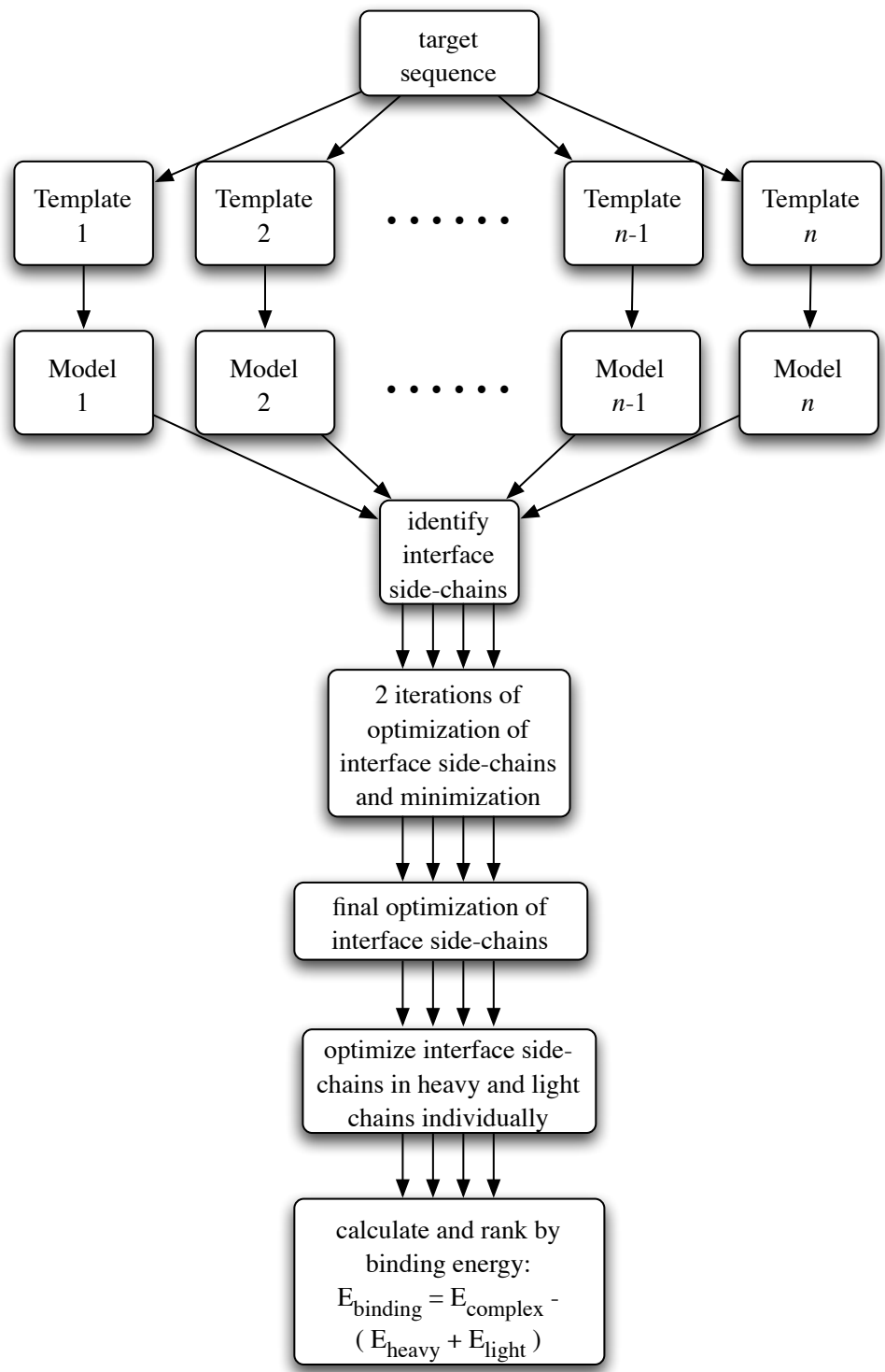
orientations increased from 0.5 Å in the calculations on the unmodified domains to 1.4 Å RMSD after the H3 loop is removed (Figure 3.3A and B). Similar results were obtained for most other antibodies in the test set, with a few exceptions (Table 3.2), some of which (e.g. PDB IDs 1dsf and 1m7d) may be the result of a flattening of the energy landscape upon removal of the H3 loop (Supplementary Figure B1). Overall, removing the H3 loop somewhat decreases the ability to identify near-native orientations, with an average deviation from the native orientation of 1.3 Å RMSD when including the crystallographically observed conformation, and 1.5 Å RMSD when excluding it. In most cases, the loss of contacts made by the H3 loop has a moderate effect on the relative orientation of the heavy and light chain variable domains. We emphasize, however, that this is a lower-bound on the effect of native H3 loop contacts on the  $V_L:V_H$  orientation because we have retained the native backbone conformation of the individual domains in these calculations.

#### *3.2.3.4. Predicting the $V_L:V_H$ orientation for comparative models of antibodies*

Our results support incorporating  $V_L:V_H$  orientation sampling as an important step in creating accurate models of antibody structure, as has been done in one recent study[143]. We created homology models of the Fv regions as described in Methods. The H3 loop was omitted because, as demonstrated above, the prediction accuracy when using native backbones was generally modestly affected by the absence of H3 loop contacts. Conversely, accurate prediction of the H3 loop depends critically on an accurate orientation of heavy and light chain variable domains (data not shown). In our modeling protocol (Figure 3.4), homology models of the light and heavy chains of antibody 1fvc were created using all structures in the full dataset that had 55-75%

sequence identity with 1fvc as template structures. This range of sequence identity was chosen to ensure that a large variety of orientations, including some that were near the native orientation, were available to sample. However, it also creates a challenging test due to the relatively low sequence identity. The range of orientations sampled from this set of templates for 1fvc ranged from 0.4 to 4.5 Å RMSD difference in domain orientation.

We estimated the “interface energy” between the two domains by subtracting the energies of the individual domains from the energy of the complex after optimization of the side-chain rotamers of the interface residues. For 1fvc, the predicted orientation (by lowest interface energy) was 2.8 Å from the native orientation, significantly worse than the results using the native backbone conformation. Overall, for the 20 protein test set, we predict the domain orientation with an average of 2.1 Å RMSD and a median of 1.6 Å, which is somewhat worse on average (by 0.8 Å) than the predicted orientations using the native backbone with the H3 loop removed (Table 3.3). In several cases, near-native domain orientations score well, just not quite as well as conformations that are farther from the native conformation. In some cases, this may reflect genuine flexibility of the interface that can contribute to “induced fit” conformational changes upon antigen binding [116], but perhaps more likely represents errors associated with the energy function and limitations of the sampling. Among all 20 antibodies, the average and median of the best model within the 5 best-ranking models by interface energy were 1.5 Å and 1.3 Å from the native orientation, respectively, and were 1.3 Å and 1.1 Å within the top 10 models ranked by interface energy.



**Figure 3.4 Method for predicting V<sub>L</sub>:V<sub>H</sub> orientation in antibody comparative modeling**

First, we create homology models using a non-redundant set of antibodies with crystal structures (in this case, all antibodies in our dataset with 55–75% sequence identity to our target sequence). Side-chains at the interface in any of the modeled structures are identified, followed by two iterations of rotamer optimization of these side-chains in the Fv structure and minimization of all atoms (including backbone) of the interface residues and the residues N- and C-terminal to all interface residues, followed by a final rotamer optimization. Each chain was also separately energy minimized after the second side-chain optimization of the variable region. The “interface energy” was estimated and used to rank the models. The orientation of the model with the lowest predicted interface energy is selected as the predicted orientation.

PDB	Range	Mean (SD)	Highest Sequence Identity	Predicted by Energy	Best of 5 Lowest Energy Predictions	Best of 10 Lowest Energy Predictions	Energy / Sequence Identity Hybrid
1a6u	1.0-4.5	2.3(0.7)	2.4	2.4	2.4	2.2	2.4
1a7n	0.4-4.0	1.8(0.8)	3.4	1.2	0.8	0.8	1.2
1bvl	0.6-4.8	2.2(0.9)	2.2	1.3	1.3	1.2	1.2
1dlf	1.4-7.2	4.3(1.2)	4.7	4.8	4.0	3.6	4.0
1dql	0.6-4.3	2.0(0.8)	1.3	2.0	1.8	1.5	1.5
1dsf	1.0-5.0	2.5(0.7)	2.9	1.6	1.6	1.6	2.0
1fvc	0.4-4.5	1.9(0.8)	3.7	2.8	0.8	0.8	0.8
1igm	0.4-4.6	2.1(0.9)	1.8	0.9	0.4	0.4	0.4
43c9	0.7-5.0	1.9(0.8)	0.7	1.7	1.4	0.9	1.6
1mam	0.7-4.6	2.3(0.9)	3.2	1.7	1.7	1.6	1.6
1dfb	0.7-7.1	3.0(1.3)	3.3	4.2	1.6	1.6	4.2
1m7d	0.8-5.3	2.7(0.9)	2.7	2.4	2.4	1.7	2.7
1mnu	0.6-5.0	2.0(0.9)	0.6	1.6	1.3	1.0	1.8
1nca	0.6-4.7	2.2(1.0)	1.7	1.3	1.3	1.3	1.4
1a14	0.6-5.3	2.8(0.9)	2.4	3.1	2.3	1.8	1.8
1dn0	0.4-4.2	1.6(0.7)	1.0	0.9	0.7	0.7	0.7
1i9j	0.4-4.1	1.8(0.8)	4.1	0.8	0.8	0.4	2.5
1bz7	0.4-4.4	2.0(0.9)	1.0	0.9	0.6	0.6	1.0
1ai1	1.0-3.8	2.2(0.6)	1.4	1.5	1.0	1.0	1.4
1fgn	0.6-6.2	3.1(1.3)	0.6	4.0	1.0	1.0	1.5
Average		2.3(1.1)	2.3	2.1	1.5	1.3	1.8
Median			2.3	1.6	1.3	1.1	1.6

**Table 3.3 Predicted Orientations After Building Comparative Models From Multiple Templates.**

For each of the twenty target antibodies listed, we built comparative models using each antibody structure with 55–75% sequence identity with the target sequence as a template. The difference in orientation between each of the resulting models and the crystallographic orientation was calculated (Range). For points of comparison, the mean and standard deviations of the domain orientation differences between the crystallographic orientation and the sampled orientations are shown, as well as the difference in orientation of the model chosen based on having the highest sequence identity to the target sequence. The accuracy of the model with the lowest predicted binding-energy between the domains and the lowest RMSD of the 5 and 10 predicted lowest-energy models for each target sequence are also shown. The final column shows the difference in orientation between the model chosen by highest template sequence identity from the top 10 models ranked by binding energy. The results in this final column show significant improvement over the mean difference in orientation (Wilcoxon Signed-Ranks one-tailed test,  $p=0.003$ ) and the accuracy of the orientation chosen based on sequence identity alone ( $p=0.05$ ).

As a control, we compared the accuracy of our predicted orientations to the accuracy obtained by choosing the orientation using a simple sequence-based metric, specifically choosing the highest sequence identity antibody in the 55-75% range for each of the 20 antibodies in our test set. This simple approach generates average and median domain orientation RMSDs of 2.3 Å (Table 3.3). In most cases, the energy-based prediction results in a comparable prediction accuracy. In four cases (PDBIDs 1a7n, 1dsf, 1mam, 1i9j), the energetic calculations provide large improvement, and in only one case (1fgn) does the calculation result in substantially worse predictions. Nonetheless, the energy-based prediction method alone does not result in significant improvement over simply choosing based on sequence identity.

It is clear that correctly predicting the domain orientation using homology models is a difficult problem, and further improvements in accuracy would be desirable. We believe that the relatively poor predictions obtained for a few of the cases are due to some combination of 1) omitting the H3 loop, and 2) subtle changes in backbone structures among antibodies, which indirectly affects side chain packing in the interfaces. More robust results will require combined or iterative sampling of domain orientations and backbone structures, as has been explored, for example, in protein-protein docking [147,148].

Given the current limitations of the energy-based prediction of the  $V_L:V_H$  orientation encountered here and the limited accuracy of choosing an orientation based on sequence identity alone, we examined the possibility of combining the interface energy and sequence identity metrics to achieve better accuracy. Re-ranking the top 10 orientations predicted using the interface energy by sequence identity showed weakly

significant improvement over using sequence identity alone ( $p = 0.05$ , Wilcoxon signed-ranks one-tailed test), resulting in a mean of 1.8 Å and a median of 1.5 Å RMSD difference in orientation (Table 3.3). For 1fvc, this improved the accuracy of the predicted orientation to 0.8 Å from the crystallographic orientation. These results suggest that the energy-based method used in combination with re-ranking based on sequence identity may be helpful in identifying a template defining the orientation of the two domains that can then be used to aid the construction of the model of the variable fragment.

### 3.3. Discussion

We first analyzed >10,000 pairs of antibody structures to quantify the significant variation in the  $V_L:V_H$  orientation (Figure 3.1, Table 3.1), which increases on average with decreasing sequence similarity for residues in the interface between the domains (Figure 3.2). We find that near-native orientations correspond to the global minimum in the energy landscape defined by the sequence (Table 3.2, Supplementary Figure B1), and only a relatively small amount (on average) of  $V_L:V_H$  orientation variability can be ascribed to crystal packing effects or conformational changes associated with antigen binding. These findings support the idea that diversification of the  $V_L:V_H$  orientation may be an important part of the ability of the immune system to recognize a tremendously diverse set of antigens. Indeed, the diversity in antibody orientation allows loops of the same structure and/or sequence to present a different combining site to the antigen, allowing for a greater diversity of recognition [115,116,117,120,121,144]. Coupled with the newly-emerging idea that the heavy chain constant region can also affect the affinity and specificity of antigen-binding [149,150,151,152,153], there is

increasing evidence that there are multiple ways in addition to CDR diversity which the immune system uses in order to convert the sequence diversity generated through somatic recombination and affinity maturation into functional diversity.

The large diversity in the  $V_L:V_H$  orientation suggests that this degree of freedom needs to be taken into account when building comparative models of antibodies. We have developed a method to predict the orientation of a given antibody using many different crystal structures of antibodies to define possible  $V_L:V_H$  orientations, side chain optimization for interface residues, and a molecular mechanics energy function to choose the optimal orientation. Using crystallographically determined variable domain backbone structures, we are able to predict the  $V_L:V_H$  orientation to better than 1 Å RMSD on average (excluding the crystallographic orientation from sampling). Predicting the  $V_L:V_H$  orientations using homology models (at 55-75% sequence identity) is more challenging, but in combination with re-ranking of the top orientations by sequence identity, we predict the  $V_L:V_H$  orientations with an average difference between the predicted and crystallographic orientations of 1.8 Å and a median of 1.5 Å RMSD. This suggests that this method may be useful in selecting a template to define the  $V_L:V_H$  orientation that can then be used in further construction of a model of the variable fragment. Gray and colleagues incorporate modeling of the  $V_L:V_H$  orientation in recent efforts to develop an accurate method to build antibody comparative models. They use sequence similarity over the Fv region of both chains to select an initial orientation and rigid-body minimization to refine the orientation following modeling of the H3 loop [143]. Our method of using multiple templates allowing for backbone-orientation coupling and energy-based calculations offers an alternative method to choose a template

for defining the  $V_L:V_H$  orientation and may be useful at lower sequence identity or similarity values.

While antibody structure modeling is important in its own right, we can also consider the modeling of the relative orientation of the light and heavy chains in antibodies as a subset of the problem of comparative modeling of inter-domain interactions in multi-domain proteins in general. Approaches similar to those explored here may be useful for predicting how relative domain orientations vary among homologous multi-domain proteins. However, the results also suggest that this will be challenging, largely because sequence can not only encode differences in domain orientation but also subtle (and occasionally not-so-subtle) changes in protein backbone conformations near the domain interface.

## 3.4. Methods

### 3.4.1. *Full antibody dataset*

We first obtained all experimentally-determined structures of antibodies by searching the Protein Databank [102] for entries with titles containing “antibody”, “Fv”, or “Fab”. Single-domain (ex. Camelid), single-chain (scFv) and homo-dimers were removed, along with structures determined at 3.0Å or greater, resulting in a list of 459 antibody structures. A few other structures were eliminated due to missing residues and other technical problems. This reduced the number of structures in our dataset to 300 (Supplementary Table B1). Each structure in this dataset was reduced to an Fv fragment by removing all residues more than 8 residues beyond the end of the third CDR on each chain.



### **3.4.2. *Dataset for energy-based analysis and comparative modeling***

In order to examine the energy landscape governing the  $V_L:V_H$  orientation and to test our ability to predict the orientation in comparative models, we constructed a test set of antibodies that were crystallized as Fv fragments in the absence of antigen, such that the presence of the constant regions or antigen could not affect the  $V_L:V_H$  orientation found in the crystal structures. In our full dataset, we found 12 structures that fit these criteria. We filtered this set for redundancy such that all pairs of structures had <90% sequence identity over the full Fv region and over the interface positions (below). This resulted in 9 antibodies. Eleven more randomly chosen antibodies were added from the non-redundant dataset (produced as described below) to bring the total number of antibodies comprising the dataset used for energy-based analysis and comparative modeling to twenty (Supplementary Table B2).

### **3.4.3. *Non-redundant dataset***

We also wanted to examine the variation in  $V_L:V_H$  orientation found in antibodies. This required a large set of structures of different antibodies. We used the 9 antibodies in our test set as a seed to generate a non-redundant set of 153 antibodies from our full dataset that had less than 90% sequence identity between any pair (Supplementary Table B2).

### **3.4.4. *Dataset to determine crystal packing and antigen effects***

For comparison to the difference in orientation found in antibodies of different sequences, we determined the amount of variation due to crystal packing and antigen binding effects (Supplementary Table B3-6). Of structures crystallized in the absence of

antigen, we identified 56 sets of structures with identical antibody sequences and 58 sets of structures with identical residues in positions at the interface of the heavy and light chain variable domains (see below). We also identified structural pairs of antibodies identical over the whole Fv region (65 pairs) or over all interface residues (66 pairs) in which one was crystallized in the absence of antigen and one was crystallized bound to antigen. Because we wanted to determine an upper bound for non-sequence-based effects on orientation, for each set of structures we chose the pair of structures with the largest difference in orientation for analysis.

#### **3.4.5. *Multiple sequence alignments for the heavy and light chains***

The heavy and light chain sequences were aligned as follows. Sequences were extracted from the SEQRES section of each PDB file in the full dataset and were cut to Fv length which we define as 8 residues past the C-terminal end of the H3 loop. In cases where multiple antibodies exist in a single unit cell, a single heavy/light-chain pair was chosen arbitrarily. Heavy chain and light chain sequences were then aligned separately using CLUSTAL [154], producing a heavy chain multiple sequence alignment (MSA) and a light chain MSA. The MSAs were manually modified to ensure alignment of the CDRs L1, L3, H1, and H2 according to structure-based alignments [126,127,128]. Gaps in the CDR regions for H3 were placed in the middle of the CDR region, and for L2, the alignment was preserved (only 7FAB has a deletion in L2, which extends into the following framework region).

#### **3.4.6. Calculation of difference in domain orientation**

To compare the domain orientation between two antibodies, **A** and **B**, the following protocol was applied. Structural alignments were performed using the non-CDR alpha carbons unless otherwise noted. First, the heavy chains of **A** and **B** were aligned using the SVDSuperimposer class in Biopython [155]. The light chain of **A** was then aligned to the light chain of **B** to obtain **A'**, the light chain of **A** in the configuration it would adopt if it had the same domain orientation as **B**. The root mean square deviation (RMSD) of the non-CDR alpha carbons was then calculated between the light chain of **A'** and **A** to give the difference in domain orientation. Using this method, RMSD calculations are unaffected by structural *intra-chain* differences between **A** and **B**, except to the extent that these differences affect the structure-based alignments. We chose to use an RMSD measure because it reflects both translations and rotations of the domains, because it is independent of the choice of axis system, and because RMSD measures are commonly used in comparative modeling studies.

To examine the variation in the  $V_L:V_H$  orientation of antibodies, we calculated the difference in orientation using this protocol between all pairs of our 153 antibody non-redundant dataset.

#### **3.4.7. Calculation of sequence identity and similarity**

Sequence identity and similarity of all pairs within our 153 antibody non-redundant dataset were calculated by combining the identity and similarity calculations for the heavy and light chains. Both the identity and similarity were calculated using the same multiple sequence alignments used to perform the structural alignments. For each position in the alignment, the number of positions that were identical or similar (as

defined by a positive score in the Blosum62 [156] matrix) was divided by the total number of positions compared. Any column in which either of the sequences being compared contained a gap was not used in the comparison.

To calculate the identity and similarity over the residues involved in the interface between the heavy and light chain variable domains, we identified all residues with any side-chain heavy atom  $< 4 \text{ \AA}$  from any heavy atom in the opposite chain.

### **3.4.8. *Energy-based scoring of domain orientations using crystal structures***

We aimed to develop a method for sampling and scoring different orientations of the heavy and light chain variable domains in order to test hypotheses about the determinants of the orientation. Due to the large number of antibody structures available, we decided to use the  $V_L:V_H$  orientations in existing antibody crystal structures as states to sample from. We refer to these orientations as “decoys”, as in the context of the current work, these are orientations that are designed to test the ability to predict the known native orientation for a given antibody. The decoy orientations are taken from all 300 antibody structures found in our full dataset. For each of the 20 non-redundant Fv structures in our test set (above), we separated the crystallographic heavy and light chains variable domains and aligned these individually to the corresponding domains in each structure of our full dataset to generate the decoy orientations.

Simply reorienting the domains can lead to severe clashes of the side-chains and, occasionally, the backbone. We made the approximation of a fixed-backbone for simplicity, and while there are possible problems with this [135,147,148,157,158,159], we believed that it would still lead to a useful model in order to probe the interactions that are important for determining the orientation. However, the side-chains at the

interface were sampled in order to find their optimal orientations in the context of the given orientation of the domains.

To define the interfacial side-chains, we determined the residues that had any side-chain heavy atom within 4 Å of any heavy atom of the opposing domain in *any* of the native or decoy structures sampled. This set of side-chains, comprising the union of the side-chains at the interface in any of the native or decoy structures, was optimized in all of the native or decoy structures, thus providing a fair comparison of each orientation. The side-chain optimization was performed using the Protein Local Optimization Program (PLOP) [69,70] with the OPLS all-atom energy function [69,106,107] and the Surface Generalized Born solvent model with a surface-area correction [14,15].

#### **3.4.9. *Energy-based prediction of domain orientations in comparative models***

We developed a method for the prediction of the  $V_L:V_H$  orientation in comparative models of antibodies (Figure 3.4). Taking into consideration the idea that the backbone conformations of the individual domains and the relative orientation between them may be linked to some degree, we decided to sample orientations and backbone conformations concomitantly by constructing multiple comparative models of the target antibody sequence. To construct each full comparative model, we built structural models of the heavy and light chains individually using one template for the framework region and one template for each of the CDR loops, except for CDR H3, which was not included in the models. Framework templates were chosen as those members of the full 300-member dataset with 55–75% sequence identity with the target sequence. This range of sequence identity was chosen to ensure that a large variety of orientations, including some that were near the native orientation, were available to

sample. The loop templates were selected from a database of non-redundant, high-resolution antibody structures using canonical loop rules defined by Martin et al [128].

The model for each domain was then constructed based on the loop and framework templates using the multiple-template homology modeling feature in PLOP. Briefly, PLOP parses a multiple sequence alignment containing the target sequence, framework template sequence, and each loop template sequence along with information on which template to use at each target residue position. The backbone and side chain positions of residues that are identical in the target and template sequences are copied from the appropriate template. Insertions, gaps, and transitions between templates are then minimally refined using the loop prediction feature in PLOP [72] and all non-conserved residue side chains are optimized [69] followed by all-atom energy minimization [160]. The models of the individual domains were then oriented as in the template used for the framework regions, and steric clashes were relieved through minimization, producing the initial comparative model for the given framework template.

We first refined these models by optimizing rotamers of the interface side-chains followed by minimization of all atoms (including backbone) of the interface residues as well as residues immediately adjacent in sequence. A second iteration of rotamer optimization of the interface side-chains was performed, followed by another minimization. From the resulting models, we then performed a third rotamer optimization (to produce the final models) as well as in the individual heavy and light chain variable domains. The relative interface energy was then calculated as the energy of the  $V_L:V_H$  complex minus the unbound, individual chain energies. The predicted model was chosen as that with the most favorable interface energy.

# Chapter 4 Phosphorylation-induced conformational changes of the Arp2/3 complex

This chapter presents currently unpublished results using molecular dynamics simulations to study conformational changes of the Arp2/3 complex upon phosphorylation of the Arp2 subunit. This work will eventually include biochemical data from Lawrence LeClaire III, Diane Barber, who received assistance from Matt Welch's group at UC-Berkeley, though the results presented here are from the strictly computational work that I have been directly involved with.

## 4.1. Introduction

Spatial and temporal control of the assembly and disassembly of actin filaments is crucial for a number of distinct cell processes, including endocytosis and cell migration [161]. The spontaneous assembly of actin filaments from a pool of actin monomers requires the formation of an unstable actin trimer nucleus, from which further polymerization is thermodynamically favorable [162]. To achieve fast filament assembly, nature has evolved classes of proteins that act as actin nucleators, constituting one level of spatiotemporal regulation. While the spire proteins and the formins nucleate

linear (unbranched) filaments [163,164], the Arp2/3 complex nucleates a new actin filament (daughter filament) from the side of an existing filament (the mother filament) [165,166]. The resulting branched filament networks are required to build many cellular structures including lamellipodia at the leading edge of motile cells and phagocytic cups, and to move endosomes and intracellular pathogens (reviewed in [161,167,168]). As such, aberrant Arp2/3 complex function has been implicated in a number of disease conditions, most notably cancer metastasis [167,169].

Arp2/3 complex function is also highly regulated, leading to a second level of regulation of actin filament assembly. The Arp2/3 complex is composed of seven subunits – the actin-related proteins Arp2 and Arp3, and ARPC1-5. In addition to binding of the mother filament, full activity of the Arp2/3 complex requires ATP binding to Arp2 and Arp3 [170,171], and binding of a nucleation promoting factor (NPF) (see [172]). Examples of NPFs include WASP [173,174], N-WASP [175], and SCAR/WAVE [176], which couple Arp2/3 complex activity to that of Rho-family GTPases [177,178], and the pathogenic proteins ActA from *Listeria monocytogenes* [179] and RickA from *Rickettsia* [180,181]. An additional, more recently identified requirement for activating nucleation by the Arp2/3 complex is threonine or tyrosine phosphorylation of the Arp2 subunit [182]. Mass spectrometry of purified Arp2/3 complex revealed phosphorylation of Arp2 Thr237 and Thr238; though no phosphorylated tyrosine was identified, mutagenesis studies pointed to Arp2 Tyr202 as the likely phosphorylation site. The structures of the apo and nucleotide-bound crystal structures were revealed by X-ray crystallography [114,183], and the structure of the active Arp2/3 complex at the junction of the mother filament and the newly nucleated daughter filament (the branch junction)



was recently revealed in cryo-EM reconstructions [184,185]. Docking of the inactive Arp2/3 complex crystal structure into the branch junction density revealed substantial rearrangements of subunits, particularly of the Arp2 and ARPC3 subunits [185]. Arp3 and Arp2 must undergo a large change in their relative orientation from their arrangement in the inactive crystal structure to their conformation in the branch junction density, in which they appear to mimic the short-pitch dimer assembly of actin subunits and incorporate into the daughter filament. Mapping of the Arp2 phosphorylation sites onto the crystal structure reveals that these residues are near the interface of Arp2, Arp3, and ARPC4, and consequently, phosphorylation of these residues could play a role in the large conformational changes expected upon activation [182]. Biochemical assays suggest that Arp2 phosphorylation causes conformational changes in the complex that are similar to those necessary for activation [182].

Computational methods, molecular dynamics methods in particular, have been used previously to study conformational changes upon phosphorylation (reviewed in [3]). Examples include the study of structural changes in the activation and glycine-rich loops of protein kinases [1,20,28,38], changes in peptide conformations [25,26], and in membrane proteins such as phospholamban [54]. Here, we use molecular dynamics simulations of the unphosphorylated and Arp2 threonine phosphorylated Arp2/3 complex to characterize their structural differences and gain insight into the mechanism by which phosphorylation induces these conformational changes. We also perform simulations of mutants to test aspects of our model of phosphoregulation. We find large conformational changes upon phosphorylation, involving the reorientation of Arp2 relative to Arp3 towards the short-pitch dimer orientation. Our simulations suggest a mechanism by

which phosphorylation alters a network of salt-bridge interactions at the Arp2/Arp3/ARPC4 interface that weakens the auto-inhibitory interactions holding Arp2 and Arp3 in an inactive orientation.

## 4.2. Results

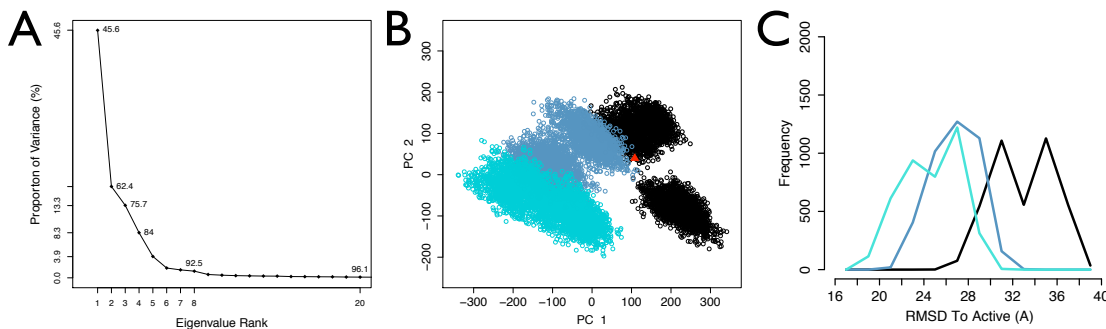
### ***4.2.1. Overall structural changes upon phosphorylation of wild-type Arp2/3 complex***

We ran duplicate 30 ns molecular dynamics simulations of the unphosphorylated Arp2/3 complex as well as complexes phosphorylated on Arp2 Thr237 or Thr238 to study conformational changes induced by phosphorylation. Observing large conformational changes using molecular dynamics simulations is difficult because of the gap between experimental and computationally feasible timescales, and we do not expect to see the full range of structural change in these simulations. Nonetheless, we believe that these simulations could reveal hints about the conformational changes due to phosphorylation and draw useful hypotheses about the interactions mediating phosphoregulation of the Arp2/3 complex.

To analyze the overall differences in structure observed between the unphosphorylated and phosphorylated simulations, we found the atomic displacement vectors best describing the observed variation by performing Principal Component Analysis [186,187] on the C $\alpha$  coordinates in each snapshot (every 10 ps) of each simulation over the final 20 ns of simulation time after aligning subdomains 1 and 2 of Arp3 (residues 1-39, 51-151, 376-410). The first and second principal components (PCs) account for 62.4% of the variation, and the first 4 principal components account for

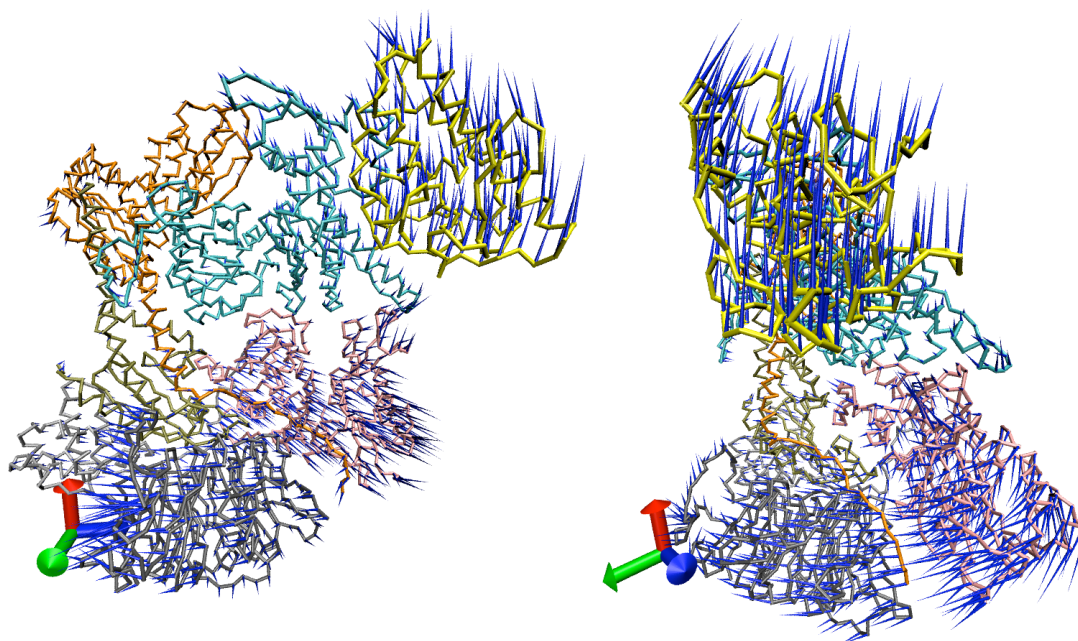
84.0% (Figure 4.1A). The large-scale differences between the unphosphorylated and phosphorylated simulations are largely localized to the first PC, with the second PC showing variation between the duplicate simulations of the complex in the same phosphorylation state (Figure 4.1B). In particular, the phosphorylated simulations sampled conformations that had more negative values of PC1 relative to the unphosphorylated state.

The first principal component primarily involves changes in the orientation of Arp2, ARPC1, and ARPC3 relative to the rest of the complex (Figure 4.2). In particular, the more negative PC1 values associated with the conformations sampled in the phosphorylated states relative to the unphosphorylated state coincide with movement of the Arp2 subunit relative to the Arp3 subunit toward its active position as a mimic of an actin short-pitch dimer [185]. This is quantified in histograms of the C $\alpha$  root mean square deviation (RMSD) of the Arp2 subunit between individual snapshots over the last 20 ns and Arp2 in its active orientation with respect to Arp3 following alignment of experimentally resolved C $\alpha$  atoms of Arp3 subdomains 1 and 2 (Figure 4.1C). This finding is consistent with the idea that phosphorylation induces conformational changes that mimic a nucleation-competent form [182].



**Figure 4.1 Principal component 1 differentiates between unphosphorylated and phosphorylated simulations**

A) Scree plot showing the proportion of variance in atomic displacement accounted for by each principal component (PC), sorted from highest to lowest eigenvalue. The total proportion of variance accounted for by all PCs with equal or greater eigenvalue than a given PC are indicated next to points on the plot. B) Snapshots from the last 20 ns of the duplicate unphosphorylated simulations (black), Arp2 T237 phosphorylated simulations (slate blue), and Arp2 T238 phosphorylated simulations (turquoise) were projected onto the first and second principal components describing the variation in atomic displacements. Discrimination of unphosphorylated and phosphorylated states is achieved along the first principal component, while the second principal component exhibits variation between the two independent simulations in each phosphorylation state. The projection of the starting crystal structure (PDB 1K8K [183]) on this set of principal components is shown as a red triangle. C) Histograms of the RMSD of the  $C\alpha$  coordinates of Arp2 in unphosphorylated (black), Arp2 Thr237 phosphorylated (slate blue), and Arp2 Thr238 phosphorylated (turquoise) simulation snapshots to their positions after placing Arp2 relative to Arp3 to mimic an actin short-pitch dimer. These RMSDs were calculated after the alignment of the experimentally-resolved  $C\alpha$  atoms of subdomains 1 and 2 of Arp3.

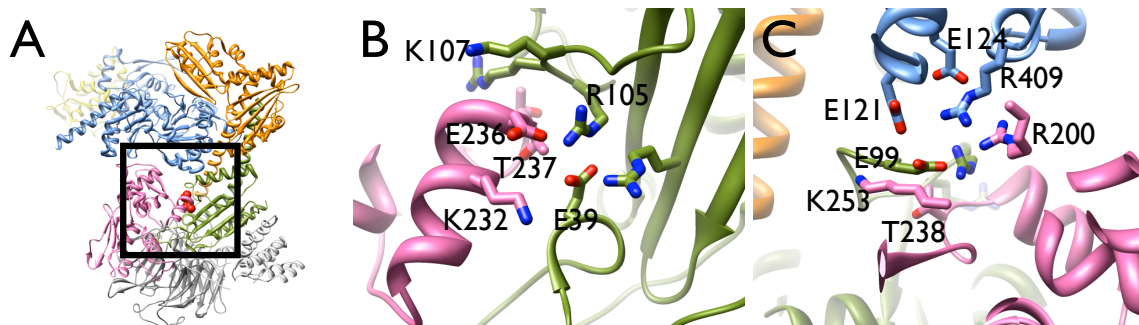


**Figure 4.2 First principal component characterized by large subunit rearrangements of Arp2, ARPC1, and ARPC3.**

Porcupine plot of the first principal component showing large subunit rearrangements of Arp2, ARPC1, and ARPC3 relative to the rest of the complex. The  $C\alpha$  coordinates displaced by one standard deviation of the conformer distribution from the average structure (of the unphosphorylated, Arp2 pThr237, and Arp2 pThr238 simulations) along the positive direction of PC1 are shown as a chain trace, and cones are drawn to the  $C\alpha$  coordinates displaced by one standard deviation of the conformer distribution in the negative direction of PC1. Since the states sampled by the phosphorylated simulations sample more negative values of PC1, the cones reflect the direction and relative size of atomic displacements needed to progress from the unphosphorylated structural states to the phosphorylated structural states. This figure was produced using Dynamite [188] and VMD [189].

#### ***4.2.2. Salt-bridge and hydrogen-bonding interactions in unphosphorylated and phosphorylated wild-type complexes***

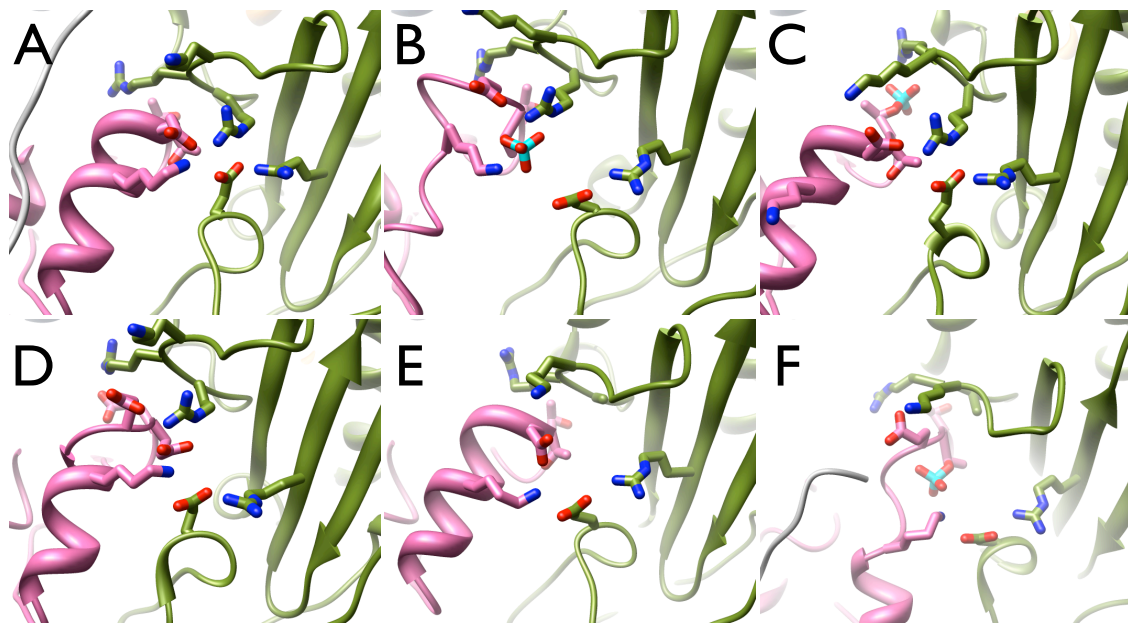
Acknowledging the difficulty in extrapolating the large but limited structural changes observed over this 30 ns timescale to functionally relevant observations on timescales of seconds and greater [190], our simulations suggest that phosphorylation on Arp2 Thr237 or Thr238 induces structural changes, in particular, reorienting Arp2 relative to Arp3 toward the nucleation-competent state. We examined the vicinity of the phosphorylation sites in detail to identify the interactions that mediated these structural changes and to draw hypotheses about the basis of phosphoregulation. As mentioned above, the crystal structure of the unphosphorylated Arp2/3 complex shows that the phosphorylation sites of Arp2 Thr237 and Thr238 are centrally located, near residues of the Arp2, Arp3, and ARPC4 subunits (Figure 4.3, [182,183]). This region contains a number of charged and polar groups that contribute to a strong electrostatic character to the interactions between Arp3, ARPC4, and the threonine phosphorylation sites of Arp2 (Supplementary Figure C1). Arg105 on ARPC4 is located near Arp2 Thr237 and interacts with Arp2 Glu236 and ARPC4 Glu39, which forms a salt bridge with ARPC4 Arg71. Arp2 Lys232 and ARPC4 Lys107 are also in this vicinity. Arg106 on ARPC4 is near Arp2 Thr238, and hydrogen bonds to ARPC4 Glu99, which is in close proximity to Arp3 Arg409 and Glu121 and Arp2 residues Arg200 and Lys253. LeClaire, et al. hypothesized that ARPC4 Arg105 and Arg106 mediate the effects of phosphorylation at Arp2 Thr237 and Thr238, respectively [182].



**Figure 4.3 Arp2 phosphorylation sites Thr237 and Thr238**

A) Overall structure showing the location of the Arp2 phosphorylation sites Thr237 and Thr238 at their central location in the starting Arp2/3 complex crystal structure, at the interface of Arp2, Arp3, ARPC4, and, to a lesser extent, ARPC2. Also shown are close-ups of the crystallographic electrostatic interactions near Arp2 Thr237 (B) and Thr238 (C). Arp2 - magenta, Arp3 - cyan, ARPC4 - olive green. This and all other structural figures were prepared with Chimera [81] except where indicated.

We hypothesized that the structural changes induced by phosphorylation may be initiated through a perturbation of the electrostatic and hydrogen-bonding network at this interface. As a control, we examined the differences in interaction between our unphosphorylated simulations and the crystal structure. We observe a large degree of conservation of the crystallographic interactions in the unphosphorylated simulations in the vicinity of the Arp2 Thr237 site (Figure 4.4A). The unphosphorylated simulations show direct interactions between Arp2 Glu236 and Arp2 Lys232 and ARPC4 Lys107 not found in the crystal structure, though these residues are close in space. The two unphosphorylated simulations differ in other aspects of the interaction network, particularly in the vicinity of Arp2 Thr238. One simulation exhibits ARPC4 Glu99 interacting with ARPC4 Arg106 as in the crystal structure, as well as Arp2 Arg200. The second simulation, on the other hand, shows breakage of these interactions. Overall, these simulations show a fairly high degree of conservation of the crystallographic interactions. Therefore, we believe the simulations provide a reliable frame of reference from which we can interpret the simulations of phosphorylated and mutant complexes.



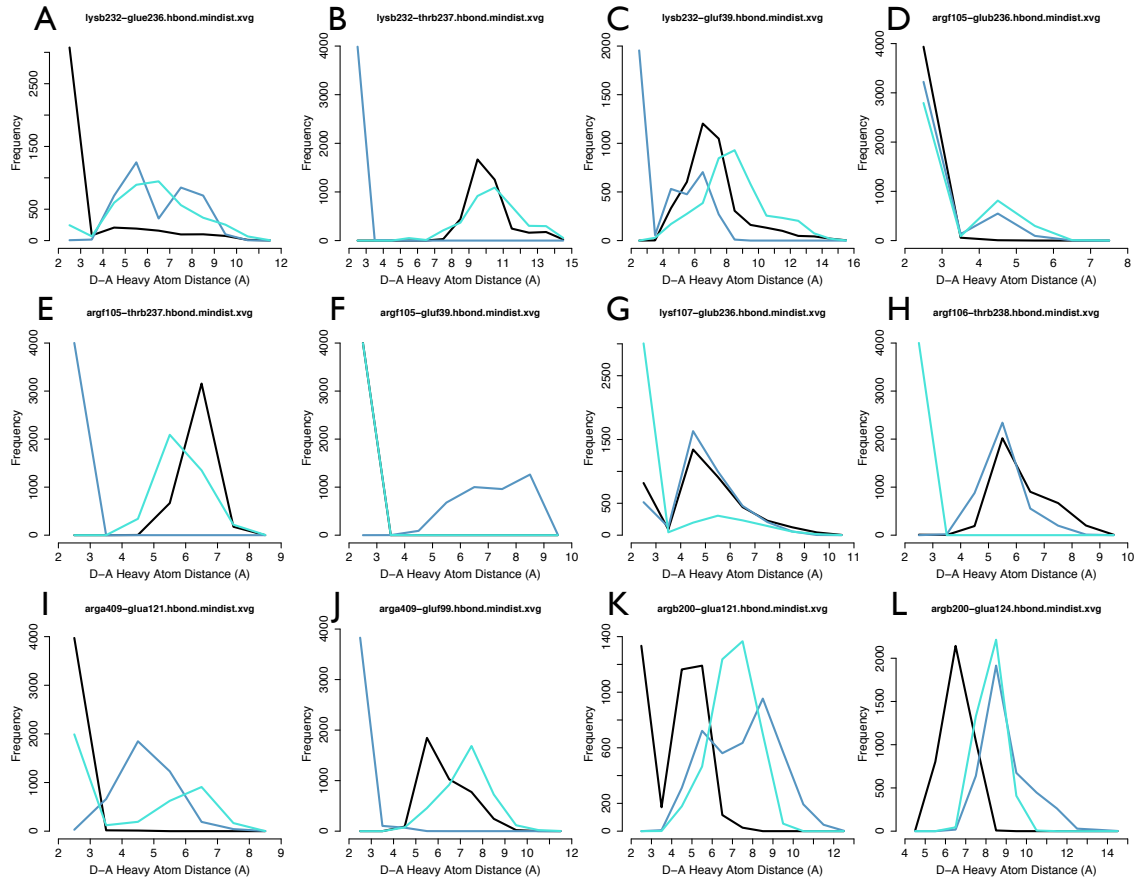
**Figure 4.4 Interactions in the vicinity of Arp2 Thr237 in simulations**

Snapshots from the last nanosecond of simulations of the unphosphorylated (A), Arp2 pThr237 (B) and Arp2 pThr238 (C) wild-type complexes, as well as the Arp2 Thr237Glu/Thr238Glu double mutant (D) and the unphosphorylated (E) and Arp2 Thr237 phosphorylated (F) ARPC4 Arg105Ala mutant complexes, with a focus on the interactions near Arp2 Thr237. Arp2 residues Lys232, Glu236, Thr/pThr237, and Thr238 are shown as well as ARPC4 residues Glu39, Arg71, Arg/Ala105, and Arg106 are shown in stick representation. Arp2 – magenta, ARPC4 – olive green.

We examined the effects of Arp2 phosphorylation on the electrostatic network in the vicinity of Arp2 Thr237 (Figure 4.5A-G). In the unphosphorylated simulations, this network includes interactions between ARPC4 Arg105 and ARPC4 Glu39 and Arp2 Glu236, and Arp2 Glu236 interacting with Arp2 Lys232. In both simulations of the Arp2/3 complex phosphorylated on Arp2 Thr237, ARPC4 Arg105 is recruited to interact with the phosphothreonine. In the process, its bidentate interaction with ARPC4 Glu39 is broken. The interaction between Arp2 Glu236 and ARPC4 Arg105 shows slight weakening, but is largely intact, while Arp2 Lys232 bridges between Arp2 pThr237 and ARPC4 Glu39. In both simulations of the Arp2/3 complex phosphorylated on Arp2 Thr238, the network found in the unphosphorylated simulations is largely intact, with the interaction between Arp2 Lys232 and Arp2 Glu236 being broken, as in the simulations of



the Arp2 pThr237 complex, and the formation of a stronger interaction between Arp2 Glu236 and ARPC4 Lys107.



**Figure 4.5 Changes in electrostatic interactions upon phosphorylation of Arp2 Thr237 or Thr238**

Histograms of the D(onor)-A(ceptor) side-chain heavy atom distances between select residue pairs in unphosphorylated (black), Arp2 pThr237 (slate blue), and pThr238 (turquoise) simulations. Distances are calculated for the following residue pairs: A) Arp2 Lys232 – Arp2 Glu236; B) Arp2 Lys232 – Arp2 Thr/pThr237; C) Arp2 Lys232 – ARPC4 Glu39; D) ARPC4 Arg105 – Arp2 Glu236; E) ARPC4 Arg105 – Arp2 Thr/pThr237; F) ARPC4 Arg105 – ARPC4 Glu39; G) ARPC4 Lys107 – Arp2 Glu236; H) ARPC4 Arg105 – Arp2 Thr/pThr 238; I) Arp3 Arg409 – Arp3 Glu121; J) Arp3 Arg409 – ARPC4 Glu99; K) Arp2 Arg200 – Arp3 Glu121; L) Arp2 Arg200 – Arp3 Glu124.

In the vicinity of Arp2 Thr238, the unphosphorylated simulations and crystal structure show a network of interactions involving Arp3 residues Glu124, Arg409, and Glu121, Arp2 residues Arg200 and Lys253, and ARPC4 residues Glu99 and Arg106. Phosphorylation of either Arp2 Thr237 or Arp2 Thr238 alters the interactions of these residues (Figure 4.5H-L). Phosphorylated Arp2 Thr238 forms a strong interaction with

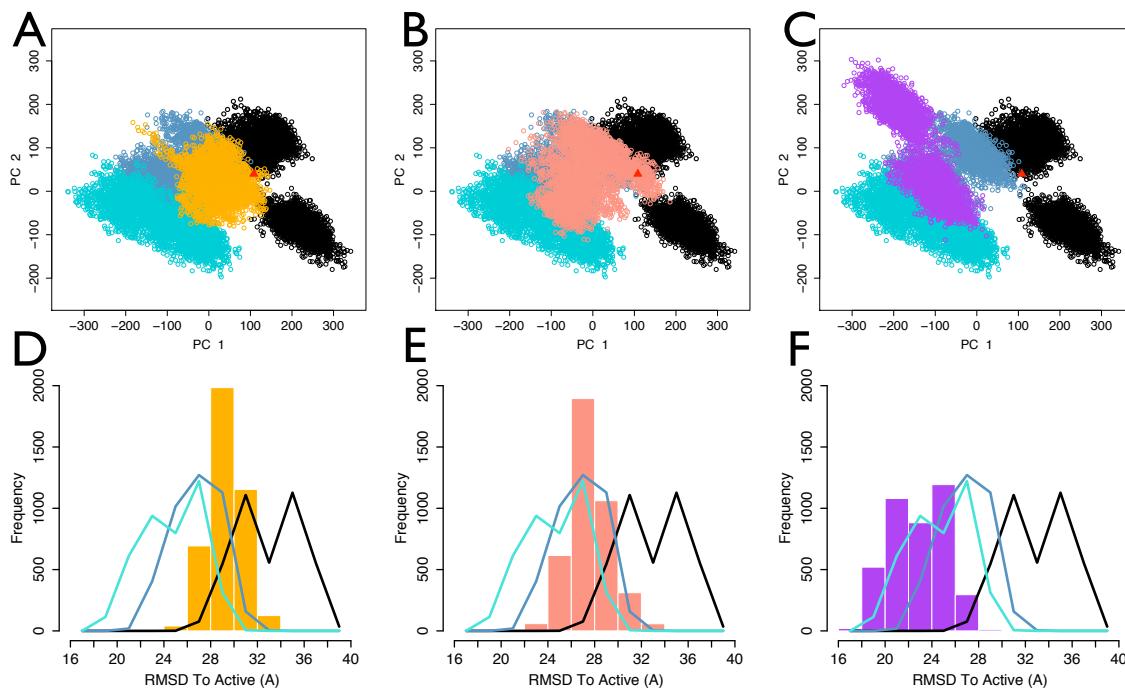
ARPC4 Arg106 initially breaking its interaction with ARPC4 Glu99, but this interaction is recovered later in the simulation. Additionally, the interactions of Arp3 Glu121 with Arp3 Arg409 and Arp2 Arg200 are weakened relative to the unphosphorylated simulations. Phosphorylation of Arp2 Thr237 appears to induce direct hydrogen bonding interactions between Arp3 Arg409 and ARPC4 Glu99, as well as weakening interactions between Arp3 Arg409 and Glu121, as in the Arp2 pThr238 simulations. As in the unphosphorylated simulations, ARPC4 Arg106 hydrogen bonds to ARPC4 Glu99 in one of the simulations, but not in the other.

#### **4.2.3. *Structural effects of mutation of the Arp2/3 complex***

While there are some differences in the particular interactions that are altered by phosphorylation of Arp2 Thr237 or Thr238 over this timescale, our simulations suggest a general model in which phosphorylation at either site causes a perturbation that alters the network of interactions near the phosphorylation site, and consequently leads to large-scale activating structural rearrangements of the complex. To examine aspects of this model, we created mutants of the Arp2/3 complex *in silico* and ran two independent 30 ns simulations for each. We examined these mutants from the perspective of large-scale structure by projecting the snapshots of the last 20 ns of each trajectory onto the principal components calculated from the unphosphorylated, Arp2 pThr237, and pThr238 simulations, as well as focusing on the atomistic interactions in the vicinity of the wild-type phosphorylation sites.

Glutamate can often, but not always, serve as a mimic of phosphothreonine, and we examined the effects of substituting glutamate residues for Arp2 Thr237 and Thr238. The structures sampled in these simulations were at somewhat more negative values of

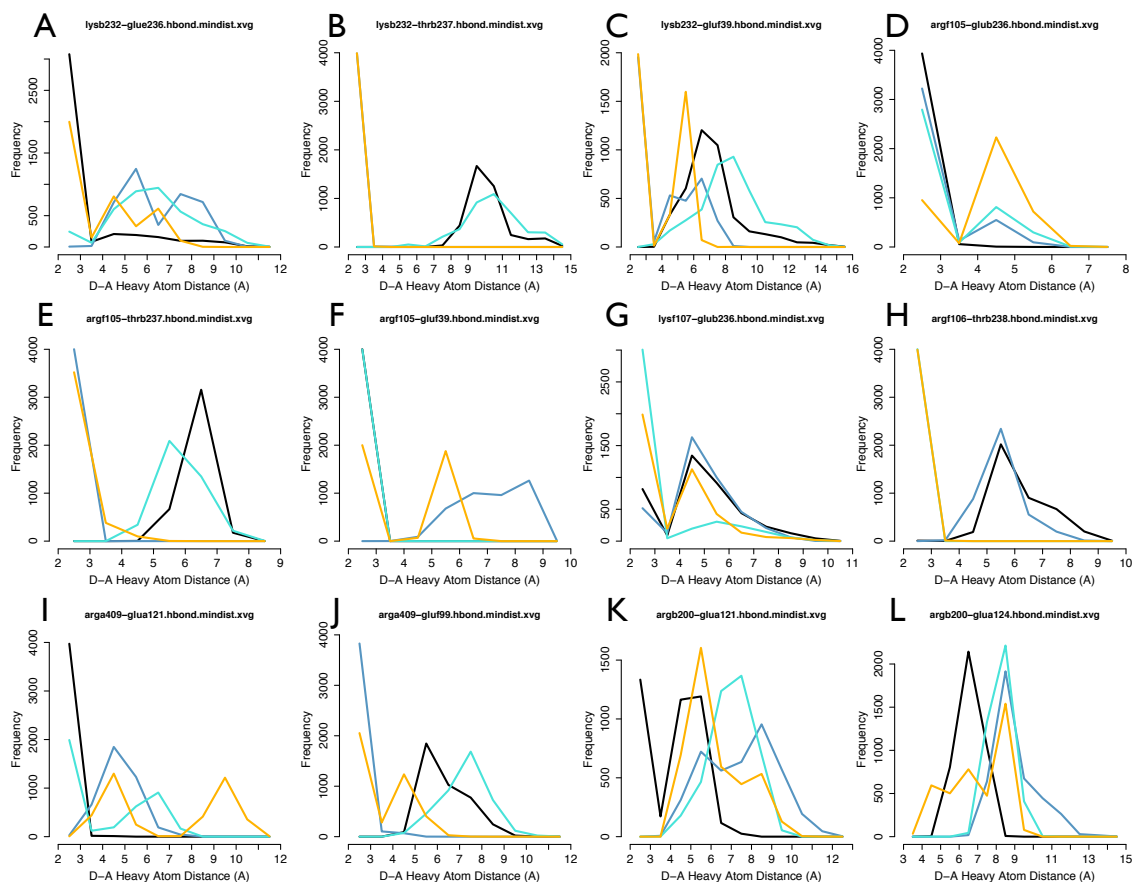
PC1 than the unphosphorylated structures, but more positive than the majority of the phosphorylated structures sampled (Figure 4.6A). Similarly, the movement of Arp2 relative to Arp3 toward the short-pitch dimer orientation was also of an intermediate nature (Figure 4.6D). An examination of the interactions surrounding the Arp2 Thr237Glu mutation reveals a network configuration that resembles that of the Arp2 pThr237 simulations more than the unphosphorylated simulations (Figure 4.7), with Arp2 Thr237Glu interacting with ARPC4 Arg105, breaking its interaction with ARPC4 Glu39. Similarly, interactions near Arp2 Thr238Glu show some characteristics of both the Arp2 pThr237 and the pThr238 simulations. Arp2 Thr238Glu interacts with ARPC4 Arg106 analogously to the interaction made by Arp2 pThr238, as well as the weakening of interactions involving Arp3 Glu121.



**Figure 4.6 Overall structural changes of mutant complexes**

Projections of simulation snapshots onto the first two principal components calculated from the unphosphorylated, Arp2 pThr237, and pThr238 simulations and the Arp2 C $\alpha$  RMSD from its position in the proposed active, short-pitch dimer orientation were calculated for the Arp2 Thr237Glu/Thr238Glu double mutant (A and C, orange) and the unphosphorylated (B and D, salmon) and Arp2 Thr237 phosphorylated (C and E, purple) ARPC4 Arg105Ala mutant complexes. Unphosphorylated – black; Arp2

pThr237 – slate blue; Arp2 pThr238 – turquoise; projection of starting crystal structure onto principal components – red triangle.

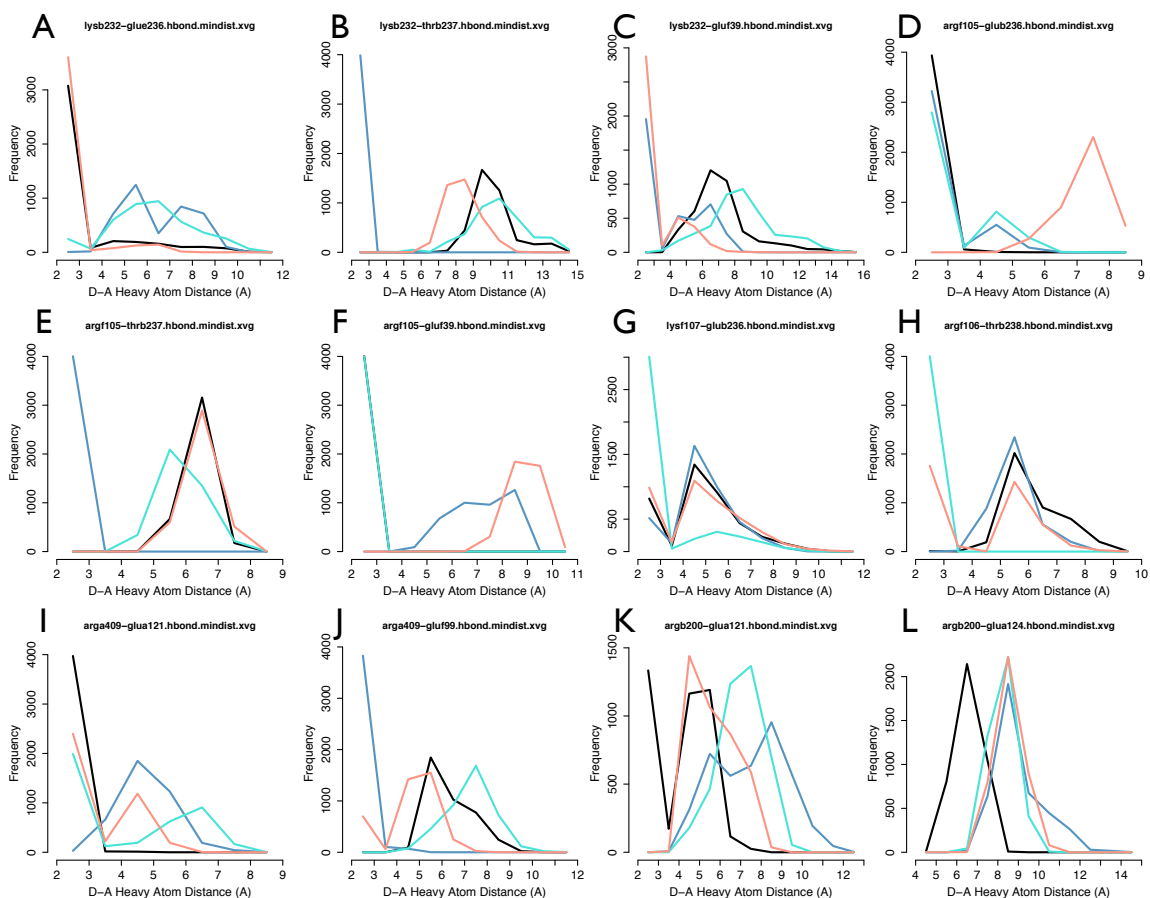


**Figure 4.7 Changes in electrostatic interactions in the Arp2 Thr237Glu/Thr238Glu double mutant simulations**

Histograms of the D(onor)-A(acceptor) side-chain heavy atom distances between select residue pairs in the Arp2 Thr237Glu/Thr238Glu double mutant simulations (orange) compared with unphosphorylated (black), Arp2 pThr237 (slate blue) and pThr238 (turquoise) simulations. Distances are calculated for the following residue pairs: A) Arp2 Lys232 – Arp2 Glu236; B) Arp2 Lys232 – Arp2 Thr/pThr237; C) Arp2 Lys232 – ARPC4 Glu39; D) ARPC4 Arg105 – Arp2 Glu236; E) ARPC4 Arg105 – Arp2 Thr/pThr237; F) ARPC4 Arg105 – ARPC4 Glu39; G) ARPC4 Lys107 – Arp2 Glu236; H) ARPC4 Arg105 – Arp2 Thr/pThr 238; I) Arp3 Arg409 – Arp3 Glu121; J) Arp3 Arg409 – ARPC4 Glu99; K) Arp2 Arg200 – Arp3 Glu121; L) Arp2 Arg200 – Arp3 Glu124.

In the Arp2 pThr237 simulations, we observed the breakage of a bidentate salt-bridge interaction between ARPC4 residues Arg105 and Glu39 as Arg105 is recruited to form strong interactions with Arp2 pThr237. We examined the effects of mutating ARPC4 Arg105 to alanine for its effects on mediating conformational changes due to phosphorylation of Arp2 at Thr237. We performed simulations on the ARPC4

Arg105Ala mutant complex in the unphosphorylated and Arp2 pThr237 states. Like the Arp2 Thr237Glu/Thr238Glu double mutant, the ARPC4 Arg105Ala mutant produced a structural change intermediate to that of phosphorylation at Arp2 Thr237 or Thr238, suggesting the possibility of partial activity, along with some movement of Arp2 towards the short-pitch dimer orientation (Figure 4.6B and E). Phosphorylation of Arp2 Thr237 in the context of the ARPC4 Arg105Ala mutant produced even larger structural changes than phosphorylation of Arp2 Thr237 (or Thr238) alone (Figure 4.6C and F). Besides the obvious breakage of salt-bridge and hydrogen-bonding interactions involving ARPC4 Arg105 by mutation to alanine, other interactions are only modestly perturbed relative to the unphosphorylated wild-type complex in the unphosphorylated ARPC4 Arg105Ala mutant simulations (Figure 4.8). The changes in the interaction network are more pronounced in the ARPC4 Arg105Ala / Arp2 pThr237 simulations (Figure 4.9). In both the unphosphorylated and phosphorylated mutant simulations, formation of the Arp2 Lys232 – ARPC4 Glu39 interaction is favored even more than in the Arp2 pThr237 simulations. An interesting observation is that while the unphosphorylated ARPC4 Arg105Ala mutant shows only modest change in the Arp2 Arg200 – Arp3 Glu121 interaction, as does the Arp2 Thr237Glu/Thr238Glu double mutant above, the phosphorylated wild-type and mutant simulations show a weakened interaction. These simulations suggest that, perhaps contrary to expectation, mutation of ARPC4 Arg105, which coordinates the phosphate of Arp2 pThr237 in the wild-type simulations, to alanine may in fact act as an activating or partially activating mutation.

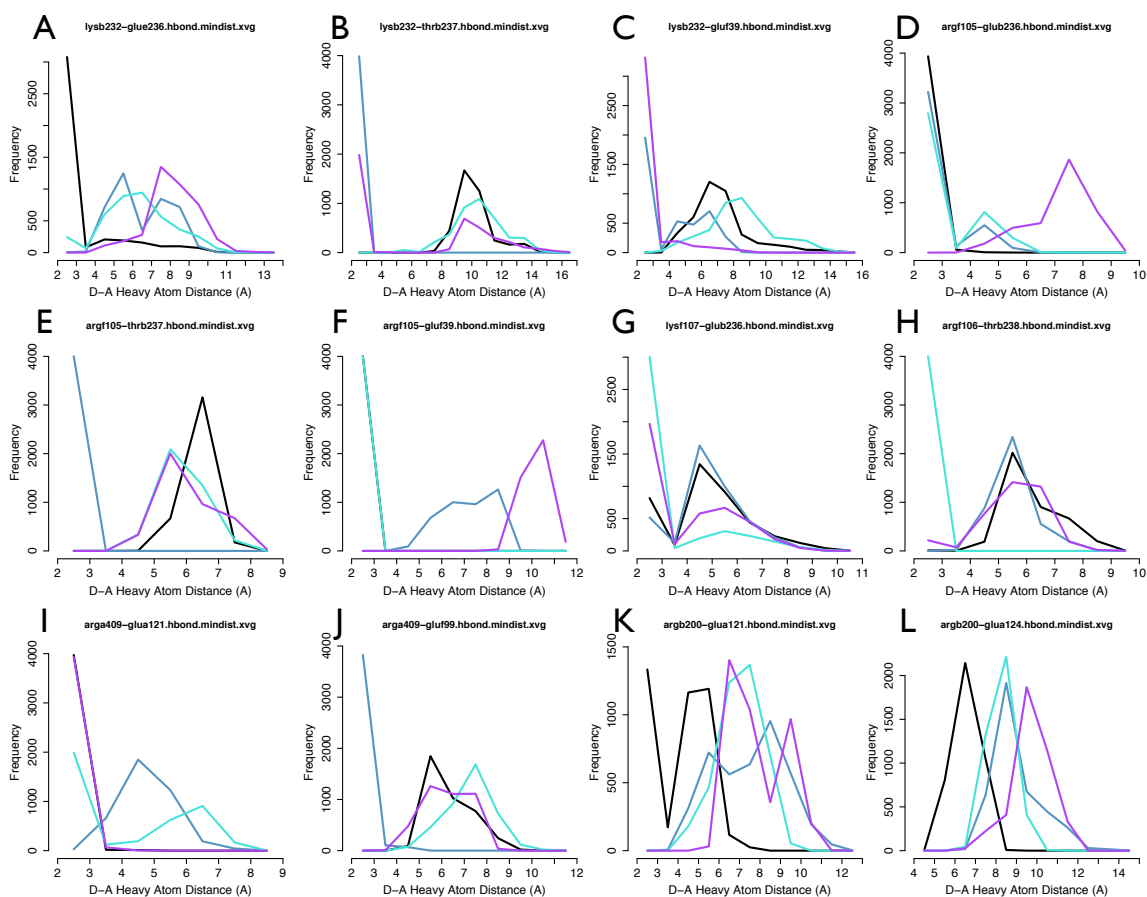


**Figure 4.8 Changes in electrostatic interactions in the unphosphorylated ARPC4 Arg105Ala mutant simulations**

Histograms of the D(onor)-A(ceptor) side-chain heavy atom distances between select residue pairs in the unphosphorylated ARPC4 Arg105Ala mutant simulations (salmon) compared with unphosphorylated (black), Arp2 pThr237 (slate blue) and pThr238 (turquoise) simulations. Distances are calculated for the following residue pairs: A) Arp2 Lys232 – Arp2 Glu236; B) Arp2 Lys232 – Arp2 Thr/pThr237; C) Arp2 Lys232 – ARPC4 Glu39; D) ARPC4 Arg105 – Arp2 Glu236; E) ARPC4 Arg105 – Arp2 Thr/pThr237; F) ARPC4 Arg105 – ARPC4 Glu39; G) ARPC4 Lys107 – Arp2 Glu236; H) ARPC4 Arg105 – Arp2 Thr/pThr238; I) Arp3 Arg409 – Arp3 Glu121; J) Arp3 Arg409 – ARPC4 Glu99; K) Arp2 Arg200 – Arp3 Glu121; L) Arp2 Arg200 – Arp3 Glu124.

We examined whether the alterations of the electrostatic and hydrogen bonding network at the interface of Arp2, Arp3, and ARPC4 caused by phosphorylation destabilizes the auto-inhibited Arp2-Arp3 orientation. To examine this possibility, we calculated the number of Arp2-Arp3 contacts (see Methods) for snapshots every 100 ps over the last 20 ns of each simulation. Consistent with the model of weakening Arp2-Arp3 interactions by phosphorylation, the number of atomic contacts between Arp2 and

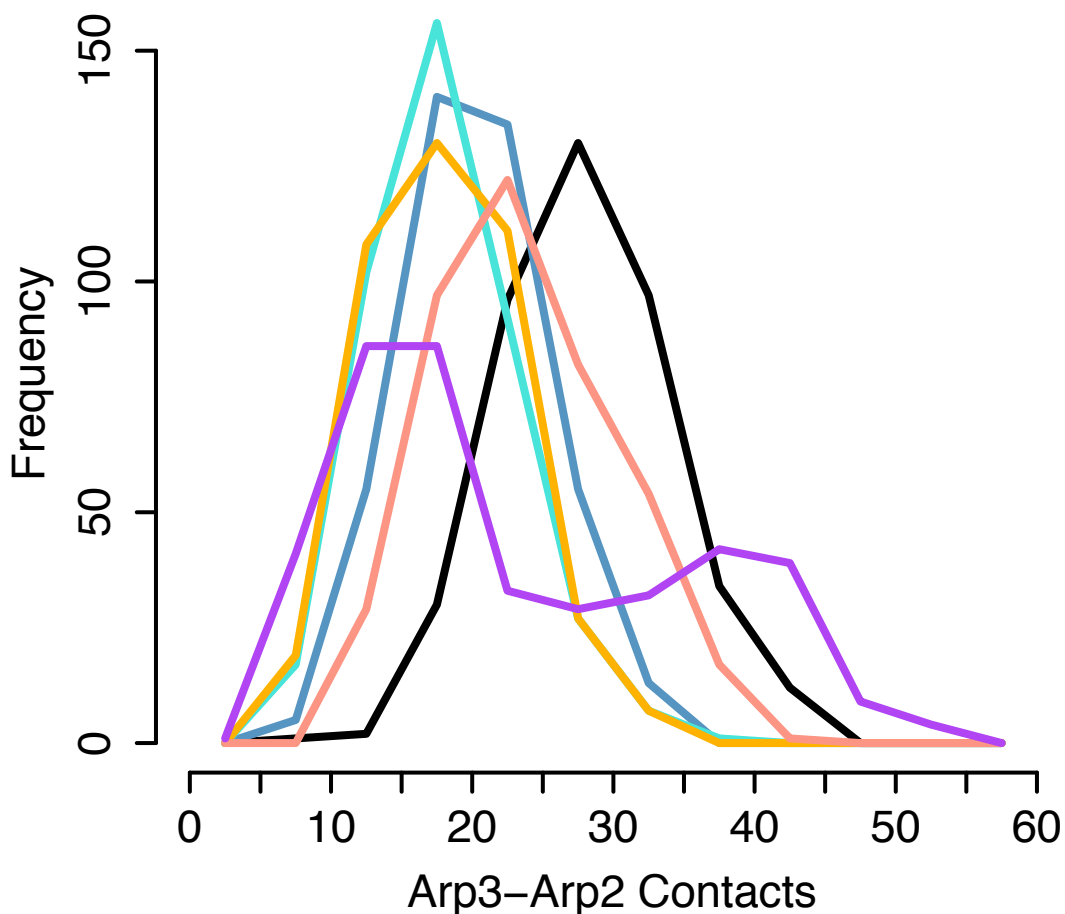
Arp3 decreases from an average of 28.4 in the unphosphorylated simulations to averages of 20.7 and 18.4 in the Arp2 pThr237 and Arp2 pThr238 simulations (Figure 4.10). Both the Arp2 Thr237Glu/Thr238Glu double mutant and the ARPC4 Arg105Ala mutant complexes show a decrease in the number of Arp2-Arp3 contacts as well (18.4 and 24.1 contacts on average, respectively). The simulations of the Arp2 pThr237 / ARPC4 Arg105Ala mutant complex showed a bimodal distribution with one peak with fewer Arp2-Arp3 contacts and one with a larger number of contacts than is found in the unphosphorylated simulations.



**Figure 4.9 Changes in electrostatic interactions in the Arp2 Thr237 phosphorylated ARPC4 Arg105Ala mutant simulations**

Histograms of the D(onor)-A(acceptor) side-chain heavy atom distances between select residue pairs in the Arp2 Thr237 phosphorylated ARPC4 Arg105Ala mutant simulations (purple) compared with unphosphorylated (black), Arp2 pThr237 (slate blue) and pThr238 (turquoise) simulations. Distances are calculated for the following residue pairs: A) Arp2 Lys232 – Arp2 Glu236; B) Arp2 Lys232 – Arp2

Thr/pThr237; C) Arp2 Lys232 – ARPC4 Glu39; D) ARPC4 Arg105 – Arp2 Glu236; E) ARPC4 Arg105 – Arp2 Thr/pThr237; F) ARPC4 Arg105 – ARPC4 Glu39; G) ARPC4 Lys107 – Arp2 Glu236; H) ARPC4 Arg105 – Arp2 Thr/pThr 238; I) Arp3 Arg409 – Arp3 Glu121; J) Arp3 Arg409 – ARPC4 Glu99; K) Arp2 Arg200 – Arp3 Glu121; L) Arp2 Arg200 – Arp3 Glu124.



**Figure 4.10 Loss of Arp2-Arp3 contacts upon phosphorylation and mutation**

Histograms of the number of heavy atom contacts between Arp2 and Arp3. Unphosphorylated – black; Arp2 Thr237 phosphorylated – slate blue; Arp2 Thr238 phosphorylated – turquoise; Arp2 Thr237Glu/Thr238Glu double mutant – orange; unphosphorylated ARPC4 Arg105Ala – salmon; Arp2 Thr237 phosphorylated ARPC4 Arg105Ala mutant – purple.



### 4.3. Discussion

We performed molecular dynamics simulations to examine structural changes in the Arp2/3 complex upon phosphorylation of Arp2 Thr237 or Thr238. Our results show that phosphorylation induces the reorientation of Arp2 relative to Arp3, pushing it towards the short-pitch dimer-like orientation found in cryo-EM reconstructions of the branch junction (Figure 4.1B and C) [185,191]. This suggests that phosphorylation leads to activating conformational changes that allow Arp2 and Arp3 to assume an orientation that more closely resembles the nucleation-competent state, consistent with biochemical data as well as the slight increase in actin polymerization observed upon addition of untreated, but not phosphatase-treated, Arp2/3 complex [182].

Analysis of the electrostatic potential of the Arp2/3 complex apo crystal structure (PDB 1K8K [183]) revealed the highly charged nature of the Arp2-Arp3-ARPC4 interface (Supplementary Figure C1). Further analysis of the interactions in the vicinity of the phosphorylation site revealed the presence of a number of salt-bridge and hydrogen bond interactions (Figure 4.3B and C). Phosphorylation of both Arp2 Thr237 and Thr238 resulted in perturbations of the network of interactions (Figure 4.4-4.6, 4.10), and we believe these play a crucial role in the large-scale structural changes observed. On the basis of these observations, we posit a model in which the network of crystallographically-observed electrostatic interactions, which are also conserved to a high degree in the simulations of the unphosphorylated complex, help to stabilize the auto-inhibited state of the complex, in particular, holding Arp2 in an inactive orientation relative to Arp3. Phosphorylation helps to alter this network of interactions, destabilizing

the auto-inhibited Arp2-Arp3 orientation. This hypothesis is consistent with our simulations of the Arp2 Thr237Glu/Thr238Glu double mutant (Figure 4.7) and the unphosphorylated ARPC4 Arg105Ala mutant complexes (Figure 4.8), which show some degree of activating conformational change (Figure 4.6 and Figure 4.10), though not as much as the phosphorylated wild-type complexes, and the phosphorylated ARPC4 Arg105Ala mutant (Figure 4.6, Figure 4.9, Figure 4.10), which shows a large activating conformational change. Experimental validation of these mutants is being pursued currently, and mutation of other residues in this electrostatic network may also provide useful tests to this model of activation by relief of auto-inhibition.

The model above may only constitute a part of the overall mechanism of phosphoregulation of the Arp2/3 complex. In addition to promoting activating conformational changes via disruption of auto-inhibitory interactions, phosphorylation of Arp2 Thr237 or Thr238 may also form interactions that lend stability to a more active-like conformation. In the Arp2 pThr238 simulations, while ARPC4 Arg106 is coordinating the phosphate, Arp3 Arg409 and Arp2 Arg200 are both, perhaps, in close enough proximity to form an arginine cluster with which Arp2 pThr237 may be able to interact. The arginine cluster is often found forming an enthalpically-favorable pocket for a phosphorylated moiety [1], such as the basic RD pocket found in many protein kinases [10]. Our simulations may be too short to observe the formation of these interactions and their effects, but it is conceivable that these may lead to further activating conformational changes. Studies of the effects of mutation at these residues may lead to answers as to whether direct interactions of the phosphothreonines with these residues play a role in the conformational change induced by phosphorylation.

Additionally, Arp2 Tyr202, which evidence suggests is a phosphorylation site [182], is in this region and may interact with this arginine cluster as well.

The negative PC1 values associated with structures sampled in the Arp2 pThr237 and pThr238 simulations also involve an increase in the distance between the C-termini of ARPC1 and ARPC3. This is in contrast to FRET studies, in which CFP and YFP were fused to the C-termini of ARPC1 and ARPC3, respectively, that showed a decrease in the inter-chromophore distance upon binding of the activating factors ATP and NPFs [170]. Our results are difficult to compare to macromolecular FRET data due to the fact that our simulations are missing CFP and YFP, as well as being on a much shorter timescale. However, the FRET study above did not examine phosphatase-treated and -untreated Arp2/3 complex conformational changes. Our simulations suggest that such a study may show an increased distance between the chromophores in the phosphorylated state compared with the unphosphorylated state achieved after phosphatase treatment. This conformational change may be required in order to allow the short-pitch dimer Arp2-Arp3 orientation to form.

Due to the difficulty in producing and studying mutants of the Arp2/3 complex, computational studies have the potential to play a valuable role in understanding aspects of Arp2/3 complex function and regulation. While the impact of computation has been limited thus far due to the large system size, molecular dynamics simulations have been used to examine dynamics of the ATP binding cleft in Arp2 and Arp3 [192,193], and homology modeling of the structures of the Arp2/3 complex from different species has generated hypotheses about functionally important surfaces [194]. We hope the model

derived from these simulations can serve as a framework for guiding experimental studies to further examine the role of phosphoregulation of the Arp2/3 complex.

## 4.4. Methods

### 4.4.1. *Molecular dynamics simulations*

Systems were prepared for molecular dynamics simulations starting from the crystal structure of the apo bovine Arp2/3 complex (PDB 1K8K [183]). A model of the unphosphorylated, wild-type bovine Arp2/3 complex was generated using the Protein Local Optimization Program (PLOP) [69,70,72] by building in all atoms missing in the electron density (except Arp2 subdomains 1 and 2). Subdomains 1 and 2 of Arp2 were modeled in based on homology to the actin monomer structure (PDB 1ATN [195]). The 15-residue unstructured extension at the end of ARPC2 was minimized, as were residues 39-51 of Arp3, residues 288-297 and 309-319 of ARPC1, and 41-43 and 65-67 of the Arp2 model. All phosphorylated and mutant models were generated from the unphosphorylated model by removing all side chain atoms from the unmodified residue and optimizing the positions of the side-chain atoms of the modified residue. These models were then solvated in TIP3P water [13] and monovalent counterions were added to neutralize the system using Maestro (Schrodinger LLC).

The full system was then minimized using DESMOND [196] in five stages with the following atoms restrained to their positions in the starting model: 1) all heavy atoms; 2) all backbone (N-C $\alpha$ -C-O) heavy atoms and experimentally determined side-chain heavy atoms; 3) all experimentally determined heavy atoms; 4) all experimentally determined backbone atoms; 5) no restraints. Minimizations were performed with at

least 100 steps of Steepest Descent minimization followed by L-BFGS optimization after reaching a gradient of  $10.0 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$  up to a total of 10,000 steps or a gradient of  $0.1 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$ . After full minimization of the system, an equilibration was performed. First, the systems were annealed to a temperature of 300K using Langevin dynamics at constant temperature and volume over 50 ps with all heavy atoms restrained. Subsequently, Langevin dynamics at constant temperature and pressure with a target temperature and pressure of 300 K and 1 atm were performed in stages: 1) 50 ps with all heavy atoms restrained with  $50 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$  force constants; 2) 50 ps with all backbone heavy atoms and experimentally determined side-chain atoms restrained with  $50 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$  force constants; 3) 150 ps with all experimentally-determined heavy atoms restrained with force constants reduced over the course of the simulation from 25 to  $5 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$ ; 4) 100 ps of simulation restraining only the experimentally determined backbone heavy atoms, over which the force constants of the restraints were brought to 0 from  $5.0 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$ ; 5) 100 ps of the unrestrained system. All Langevin dynamics simulations were performed with a  $100 \text{ ps}^{-1}$  damping constant.

Each system was then simulated for 30 ns using the Martyna-Tobias-Klein integrator [197] with a reference temperature of 300 K and a reference pressure of 1 atm. The barostat mass was set with a time constant of 2 ps and an equilibrium temperature of 300 K. The masses of all chain variables were set using a time constant of 1.0 ps. Both the Langevin dynamics and standard molecular dynamics simulations were performed with all bonds involving hydrogens constrained, a 2 fs time step for the bonded and short-range nonbonded interactions and updating of long-range nonbonded interactions every 4 fs using the RESPA multiple time step approach. Nonbonded interactions were tapered

using force-switching starting at a distance of 9.0 Å to an interaction cutoff of 9.5Å. Pairlists were constructed using a distance of 10.5 Å and a migration interval of 12 ps. These parameters were tested in short simulations in the NVE ensemble to ensure good energy conservation.

#### **4.4.2. *Principal component analysis***

Coordinates of the C $\alpha$  atoms from the last 20 ns of each unphosphorylated, Arp2 pThr237, and Arp2 pThr238 simulation were collected into a single trajectory on which Principal Component Analysis was performed using the Bio3D package [198] for the R statistical software package [199]. The C $\alpha$  atoms for the duplicate simulations of the Arp2 Thr237Glu/Thr238Glu and the unphosphorylated and Arp2 Thr237 phosphorylated ARPC4 Arg105Ala mutants were projected onto the full set of principal components. All C $\alpha$  coordinates were used after the superposition of the C $\alpha$  atoms of Arp3 subdomains 1 and 2 resolved in the starting crystal structure (residues 3-39, 51-151, 376-410) of each frame in each trajectory.

#### **4.4.3. *Calculation of number of Arp2-Arp3 contacts***

A contact between the Arp2 and Arp3 subunits was defined as the number of heavy atoms in Arp3 that were within 3.5 Å of any heavy atom in Arp2. The number of contacts between the Arp2 and Arp3 subunits was calculated for every 10<sup>th</sup> frame of each simulation. The results for the duplicate simulations of each wild-type or mutant complex were then pooled and compared.

# Chapter 5 Future Directions

## 5.1. Scoring and sampling

The two fundamental challenges in dealing with any system computationally are in sampling and scoring relevant conformations, and work is ongoing to improve both aspects. Some issues regarding the energetics of phosphorylated amino acids have been explored in model systems [17], but it is clear that treating the -2 charge of the phosphate presents a major challenge for molecular mechanics energy functions. Electronic polarizability, which is not treated by commonly used fixed charge force fields, is potentially a significant factor for portions of a macromolecule interacting with a phosphate group carrying a -2 charge. Polarizable force fields can, in principle, capture such effects. Alternatively, QM/MM methods may be a promising (although computationally costly) approach, particularly for comparison to purely MM methods.

Adequate sampling can also limit the ability to fully elucidate conformational and dynamical responses to phosphorylation. The timescale required to observe conformational changes induced upon phosphorylation are unknown and need not be fast relative to the length of most molecular dynamics simulations. Particularly when some assumptions can be made about the extent or nature of conformational change, Monte Carlo-type methods may be used to sample more efficiently and consequently reduce the

computational cost. Numerous other sampling methods, such as those reviewed by Scheraga and coworkers [200], may also be useful for studying the mechanisms of electrostatic switches.

## 5.2. Phosphoregulation of ion channels

While phosphorylation-induced conformational changes have been better studied in cytosolic proteins than in membrane proteins both experimentally and computationally, phosphorylation is an important regulatory mechanism in membrane proteins. Ion channels are integral membrane proteins that can be regulated by phosphorylation through changes in receptor trafficking, leading to decreases in the number of channels in the membrane, or through alteration of the conducting properties of the channel itself. Notable examples of phosphorylation-mediated changes in ion channel trafficking include AMPA [201] and NMDA [202] receptors, which play key roles in synaptic transmission and long-term potentiation. (In the G-protein coupled receptor family of membrane proteins, phosphorylation affects the trafficking of mu-opioid receptors, the internalization of which is thought to be a cause for addiction to opiates such as morphine [203].) In contrast to altering the number of channels, phosphorylation can also induce changes in single-channel activity through conformational changes, as discussed below.

### 5.2.1. *Kir1.1*

Kir1.1 (also known as ROMK1), an inward-rectifying potassium channel Kir1.1 that functions in the kidney to maintain potassium homeostasis, has been shown to be phosphorylated on serine residues 44, 219, and 313 by PKA [204]. Phosphorylation on at



least two of these sites is required for channel activity, while maximal conduction seems to require phosphorylation of all 3 sites. Single-channel patch clamp recordings on the closely related and similarly regulated splice variant Kir1.2 (ROMK2) revealed that the equivalent of Ser44 in Kir1.1 altered the number of channels at the membrane, while dephosphorylation of the latter two sites led to a decrease in the open channel probability and conductance [205]. The effect of phosphorylation at Ser219 and Ser313 were shown to increase Kir1.1 binding to PIP2, the presence of which is required for active channels [206]. Because the -2 charged phosphates repel the -4 charged PIP2, the fact that phosphorylation increases the binding affinity for PIP2 is surprising, and suggests that phosphorylation induces a conformational change that promotes PIP2 binding. These conformational changes are apparently not reproduced upon mutation of Ser313 to Asp, and mutation of Ser219 to Asp actually renders the channel even less sensitive to PIP2, opposing the effects of phosphorylation [206].

Kir1.1 assembles as a homotetramer in the functional channel, and consequently is a large system to treat with computational approaches. Another difficulty arises in the fact that no full-length eukaryotic Kir channel crystal structure has been solved. However, the crystal structures of the bacterial homologue KirBac1.1 [207][KuoDoyle2003] and a chimeric channel containing the transmembrane regions of the bacterial KirBac1.3 fused to the cytoplasmic regions of Kir3.1 [208]. Nonetheless, several homology modeling and MD studies have been performed on these channels and have produced valuable insights and hypotheses regarding interactions with agonists as well as gating mechanisms. Through homology modeling of the ATP-sensitive Kir6.2 channel, the authors were able to make predictions about the location of the ATP binding

site, which were supported through mutagenesis studies [209]. Further work on the Kir6.2 channel, employing ligand docking and MD simulations provided a candidate binding site for the PIP2 agonist that is a common activator for all Kir channels [210]. Notably, both of these binding sites contacted residues from distinct subunits, and this binding of regulators at subunit interfaces was suggested as an efficient way to induce large structural changes, such as those involved in channel opening/closing. A comparative analysis of MD simulations of Kir1.1, Kir3.1, and Kir6.2 models was also performed [211]. One study has been performed directly examining the effect of phosphorylating Ser219, aimed at identifying the conformational changes that may give rise to the increased PIP2 affinity [212]. Comparison of structures at the beginning and end of their simulations, which were performed in the absence of a membrane, suggest that phosphorylation decreases the distance between pSer219 and Arg217 while increasing the distance to Lys218, moving it closer to the membrane where it can form a stronger interaction with PIP2. Longer simulations with an explicit lipid membrane may provide increased understanding of phosphorylation-induced structural changes and the interactions mediating those changes.

Kir1.1 is also inhibited by intracellular acidification, and much work has been focused on identifying the pH sensor. The pH sensor was thought to be Lys80 based on the fact that mutation of this residue to Met results in a lower pH being required for closure of the channel [213]. However, recent work has shown convincingly that Lys80 is part of the gating mechanism rather than the pH sensing mechanism, and mutation of Lys80 destabilizes the closed state leading to a lower pH being required to close the channel [214,215,216]. To my knowledge, the identity of the titratable pH sensor is

unknown. Constant-pH molecular dynamics simulations may provide insight into the relevant titratable residue and the subsequent conformational changes.

### **5.2.2. *Other ion channels regulated by phosphorylation***

A few additional examples of membrane proteins that are regulated by phosphorylation are presented here for consideration for future projects. I have not researched these as extensively as I have researched Kir1.1 phosphorylation, but hopefully these can provide starting points for further reading.

The cystic fibrosis transmembrane conductance regulator (CFTR) is a chloride channel that is crucial to transport of chloride ions across epithelial cells [217]. Phosphorylation of the regulatory R domain of CFTR on multiple sites by PKA is essential to channel conductance, though no individual site is absolutely necessary [217]. The R domain, which is unstructured when expressed isolated in solution, has been shown to associate with other domains of CFTR only in the presence of phosphorylation [218]. Recent work has used secondary structure predictions to identify marginally-stable regions of secondary structure, and several of these map to the vicinity of phosphorylation sites [219]. Therefore, the authors suggest that phosphorylation may induce ordering of local secondary structure. The same study uses Discrete Molecular Dynamics [220,221] to produce a low-energy ensemble of conformations of the R domain, and suggest that phosphorylation increases the radius of gyration [219]; naively, this seems contradictory to the hypothesis of phosphorylation-induced ordering.

Connexin 43 (Cx43) forms hexameric “hemichannels” at gap junctions, allowing cell-to-cell communication of small solutes. (Two hemichannels, one contributed by each cell, allow the formation of the fully functional channel.) Phosphorylation of

Protein Kinase C (PKC) inhibits the channel activity, blocking cell-to-cell solute passage. Limited proteolysis and tryptophan fluorescence revealed that phosphorylation of Ser368 induces a conformational change by which parts of the C-terminus are stabilized, though one or more tryptophan residues are more exposed in the unphosphorylated state [222].

### **5.2.3. *Channelopathies***

Single-nucleotide polymorphisms (SNPs) are markers of human genetic variation and can be associated with susceptibility to diseases. Ion-channel related diseases (termed channelopathies) could result from incorrect phosphoregulation, which could affect the trafficking or conductive properties of channels. Consequently, Armstrong and co-workers have identified 16 candidate SNPs in ion channels that lead to the introduction or destruction of a predicted phosphorylation site [56]. They have investigated two channels that they have found to exhibit altered function upon insertion of a phosphorylation site, the voltage-gated calcium channel CaV1.2 and the voltage-gated potassium channel Kv11.1 encoded by the hERG1 gene. CaV1.2 exhibits the Gly406Arg mutation that is associated with the neurological disorder Timothy Syndrome [223]. The authors studied the homologous mutation (Gly436Arg) in the rabbit CaV1.2 channel and found that this allowed phosphorylation of Ser439 (rabbit numbering) by CaMKII, and that this phosphorylation led to overly-activated channels [55]. The Kv11.1 channel is important for the rhythmic contractions of cardiac muscle. Armstrong and co-workers find that the SNP Lys897Thr causes a loss in the PKN phosphorylation of Thr895, while introducing an Akt (PKB) phosphorylation site at Thr897 [56]. This phosphorylation led to the inhibition of channel activity, as did the wild-type Thr895

phosphorylation site, though these led to differences in responses to upstream signaling events.

As with most phosphorylation events, the structural basis for the effects of these altered phosphorylation sites is unclear, and consequently, atomic-level computation may be able to provide insight into the relevant interactions that mediate these functional changes. In addition, the fact that these are phosphorylation sites introduced by mutation provides an interesting starting point from which to think about design principles of phosphoregulation.

### 5.3. Lysine methylation

Several other post-translational modifications also modulate protein charge in a site-specific manner, and hence may also function as “electrostatic switches”. These include sulfonation [224], lysine acetylation, and glutamate methylation, the latter two of which are important histone modifications [225,226,227]. Compared with phosphorylation, relatively little is known about these modifications, especially at a structural level. This implies that there is a potentially large role for computation to play, although the dearth of data also makes it challenging to evaluate the suitability of different computational methods.

Lysine methylation is an extremely interesting post-translational modification in that it preserves the positive charge on the amine, but loses hydrogen bonding capability with increased methylation. The extreme case is the tri-methylated lysine, which carries a positive charge but cannot make any hydrogen bonds. Lysine methylation and other histone modifications mediate the binding of interacting domains [228], and linking of these interacting domains to recognize multiple modifications allows the modification of

chromatin structure [229]. The presence and degree of lysine methylation are read out by domains that are members of the Royal-superfamily, including chromodomains and its variants, and several crystal structures now exist of these domains bound to peptides containing mono-, di-, and tri-methylated lysines [230,231,232,233,234,235,236,237,238]. Crystal structures show di- and tri-methylated lysines bind in an “aromatic cage” in a shallow surface groove, where the tertiary or quaternary amine is bound through cation- $\pi$  interactions with enclosing aromatics. In contrast, the domains reading out un- or mono-methylated states apparently use size exclusion and hydrogen bonding in a deep cavity for the side chain. It appears that the binding preferences for different methylation states, at least for the chromodomain HP1, are not strict – while tri- and mono-methylated lysines can differ in their dissociation constant by a factor of 10, this difference can be a factor of 2 between di- and tri-methylated lysines [234,235]. However, all of the modified lysines bind significantly better than the unmodified lysine.

While some of these preferences can be rationalized on the basis of the crystal structures, a more thorough thermodynamic analysis is warranted. There are a number of basic physicochemical comparisons to make regarding the different methylation states of lysine. In correctly representing binding thermodynamics, the correct treatment of the unbound state is important to obtaining accurate absolute and relative free energies of binding. Consequently, a comparison between the solvated states of lysine side chains in different methylation states could be useful to understand solvent structure and relative solvation energies. Additionally, interactions could be compared to different groups, including carboxylates, phosphates, and, due to their prominent interactions in crystal

structures and the important role of cation- $\pi$  interactions [239], aromatics to assess the magnitude of the effects of methylation on different types of interactions. Further development could involve the calculation of relative affinities of different methylation states with different model binding sites mimicking those found in crystal structures. Through an improved understanding of the atomic-level interactions in model systems, insight into the recognition properties of existing chromodomains as well as into the possibility of altering or designing specificity of chromodomains and other methyl-lysine binders might be gained.

## References

1. Groban ES, Narayanan A, Jacobson MP (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol* 2: e32.
2. Narayanan A, Sellers BD, Jacobson MP (2009) Energy-based analysis and prediction of the orientation between light- and heavy-chain antibody variable domains. *J Mol Biol* 388: 941-953.
3. Narayanan A, Jacobson MP (2009) Computational studies of protein regulation by post-translational phosphorylation. *Curr Opin Struct Biol* 19: 156-163.
4. Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res* 27: 237-239.
5. Craig AL, Burch L, Vojtesek B, Mikutowska J, Thompson A, et al. (1999) Novel phosphorylation sites of human tumour suppressor protein p53 at Ser20 and Thr18 that disrupt the binding of mdm2 (mouse double minute 2) protein are modified in human cancers. *Biochem J* 342 ( Pt 1): 133-141.
6. Chrivia JC, Kwok RP, Lamb N, Hagiwara M, Montminy MR, et al. (1993) Phosphorylated CREB binds specifically to the nuclear protein CBP. *Nature* 365: 855-859.
7. Parker D, Ferreri K, Nakajima T, LaMorte VJ, Evans R, et al. (1996) Phosphorylation of CREB at Ser-133 induces complex formation with CREB-binding protein via a direct mechanism. *Mol Cell Biol* 16: 694-703.
8. Huse M, Kuriyan J (2002) The conformational plasticity of protein kinases. *Cell* 109: 275-282.
9. Johnson LN, Lewis RJ (2001) Structural basis for control by phosphorylation. *Chem Rev* 101: 2209--2242.
10. Nolen B, Taylor S, Ghosh G (2004) Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* 15: 661--675.
11. Blume-Jensen P, Hunter T (2001) Oncogenic kinase signalling. *Nature* 411: 355-365.



12. Ostman A, Hellberg C, Bohmer FD (2006) Protein-tyrosine phosphatases and cancer. *Nat Rev Cancer* 6: 307-320.
13. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* 79: 926-935.
14. Gallicchio E, Zhang LY, Levy RM (2002) The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of Computational Chemistry* 23: 517-529.
15. Ghosh A, Rapp CS, Friesner RA (1998) Generalized born model based on a surface integral formulation. *Journal of Physical Chemistry B* 102: 10983-10990.
16. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, et al. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem* 23: 128-137.
17. Mandell DJ, Chorny I, Groban ES, Wong SE, Levine E, et al. (2007) Strengths of hydrogen bonds involving phosphorylated amino acid side chains. *J Am Chem Soc* 129: 820--827.
18. Wong SE, Bernacki K, Jacobson M (2005) Competition between intramolecular hydrogen bonds and solvation in phosphorylated peptides: simulations with explicit and implicit solvent. *J Phys Chem B* 109: 5249--5258.
19. Luo R, David L, Hung H, Devaney J, Gilson MK (1999) Strength of solvent-exposed salt-bridges. *Journal of Physical Chemistry B* 103: 727-736.
20. Banavali NK, Roux B (2007) Anatomy of a structural pathway for activation of the catalytic domain of Src kinase Hck. *Proteins* 67: 1096-1112.
21. Lei M, Velos J, Gardino A, Kivenson A, Karplus M, et al. (2009) Segmented transition pathway of the signaling protein nitrogen regulatory protein C. *J Mol Biol* 392: 823-836.
22. Ozkirimli E, Post CB (2006) Src kinase activation: A switched electrostatic network. *Protein Sci* 15: 1051--1062.

23. Ozkirimli E, Yadav SS, Miller WT, Post CB (2008) An electrostatic network and long-range regulation of Src kinases. *Protein Sci* 17: 1871--1880.
24. Zou J, Wang Y-D, Ma F-X, Xiang M-L, Shi B, et al. (2008) Detailed conformational dynamics of juxtamembrane region and activation loop in c-Kit kinase activation process. *Proteins* 72: 323--332.
25. Hamelberg D, Shen T, McCammon JA (2005) Phosphorylation effects on cis/trans isomerization and the backbone conformation of serine-proline motifs: accelerated molecular dynamics analysis. *J Am Chem Soc* 127: 1969--1974.
26. Hamelberg D, Shen T, McCammon JA (2007) A proposed signaling motif for nuclear import in mRNA processing via the formation of arginine claw. *Proc Natl Acad Sci U S A* 104: 14947--14951.
27. Cheng Y, Zhang Y, McCammon JA (2005) How does the cAMP-dependent protein kinase catalyze the phosphorylation reaction: an ab initio QM/MM study. *J Am Chem Soc* 127: 1553-1562.
28. Cheng Y, Zhang Y, McCammon JA (2006) How does activation loop phosphorylation modulate catalytic activity in the cAMP-dependent protein kinase: a theoretical study. *Protein Sci* 15: 672--683.
29. De Vivo M, Cavalli A, Bottegoni G, Carloni P, Recanatini M (2006) Role of phosphorylated Thr160 for the activation of the CDK2/Cyclin A complex. *Proteins* 62: 89--98.
30. Endicott JA, Noble ME, Tucker JA (1999) Cyclin-dependent kinases: inhibition and substrate recognition. *Curr Opin Struct Biol* 9: 738-744.
31. Morgan DO (1995) Principles of CDK regulation. *Nature* 374: 131-134.
32. Matsuura I, Wang JH (1996) Demonstration of cyclin-dependent kinase inhibitory serine/threonine kinase in bovine thymus. *J Biol Chem* 271: 5443-5450.
33. Zukerberg LR, Patrick GN, Nikolic M, Humbert S, Wu CL, et al. (2000) Cables links Cdk5 and c-Abl and facilitates Cdk5 tyrosine phosphorylation, kinase upregulation, and neurite outgrowth. *Neuron* 26: 633-646.
34. Bartova I, Koca J, Otyepka M (2008) Regulatory phosphorylation of cyclin-dependent kinase 2: insights from molecular dynamics simulations. *J Mol Model* 14: 761--768.

35. Bartova I, Otyepka M, Kriz Z, Koca J (2004) Activation and inhibition of cyclin-dependent kinase-2 by phosphorylation; a molecular dynamics study reveals the functional importance of the glycine-rich loop. *Protein Sci* 13: 1449-1457.
36. Bartova I, Otyepka M, Kriz Z, Koca J (2005) The mechanism of inhibition of the cyclin-dependent kinase-2 as revealed by the molecular dynamics study on the complex CDK2 with the peptide substrate HHASPRK. *Protein Sci* 14: 445--451.
37. Welburn JPI, Tucker JA, Johnson T, Lindert L, Morgan M, et al. (2007) How tyrosine 15 phosphorylation inhibits the activity of cyclin-dependent kinase 2-cyclin A. *J Biol Chem* 282: 3173--3181.
38. Zhang B, Tan VBC, Lim KM, Tay TE (2007) The activation and inhibition of cyclin-dependent kinase-5 by phosphorylation. *Biochemistry* 46: 10841--10851.
39. Young MA, Gonfloni S, Superti-Furga G, Roux B, Kuriyan J (2001) Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell* 105: 115-126.
40. Mendieta J, Gago F (2004) In silico activation of Src tyrosine kinase reveals the molecular basis for intramolecular autophosphorylation. *J Mol Graph Model* 23: 189-198.
41. Solt I, Magyar C, Simon I, Tompa P, Fuxreiter M (2006) Phosphorylation-induced transient intrinsic structure in the kinase-inducible domain of CREB facilitates its recognition by the KIX domain of CBP. *Proteins* 64: 749--757.
42. Zor T, Mayr BM, Dyson HJ, Montminy MR, Wright PE (2002) Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. *J Biol Chem* 277: 42241-42248.
43. Wright PE, Dyson HJ (2009) Linking folding and binding. *Curr Opin Struct Biol* 19: 31-38.
44. Espinoza-Fonseca LM, Kast D, Thomas DD (2007) Molecular dynamics simulations reveal a disorder-to-order transition on phosphorylation of smooth muscle myosin. *Biophys J* 93: 2083--2090.
45. Espinoza-Fonseca LM, Kast D, Thomas DD (2008) Thermodynamic and structural basis of phosphorylation-induced disorder-to-order transition in the regulatory

- light chain of smooth muscle myosin. *Journal of the American Chemical Society* 130: 12208-12209.
46. Smart JL, McCammon JA (1999) Phosphorylation stabilizes the N-termini of alpha-helices. *Biopolymers* 49: 225-233.
  47. Lu KP, Zhou XZ (2007) The prolyl isomerase PIN1: a pivotal new twist in phosphorylation signalling and disease. *Nat Rev Mol Cell Biol* 8: 904-916.
  48. Mazanetz MP, Fischer PM (2007) Untangling tau hyperphosphorylation in drug design for neurodegenerative diseases. *Nat Rev Drug Discov* 6: 464--479.
  49. Lee HJ, Srinivasan D, Coomber D, Lane DP, Verma CS (2007) Modulation of the p53-MDM2 interaction by phosphorylation of Thr18: a computational study. *Cell Cycle* 6: 2604--2611.
  50. Brown CJ, Srinivasan D, Jun LH, Coomber D, Verma CS, et al. (2008) The electrostatic surface of MDM2 modulates the specificity of its interaction with phosphorylated and unphosphorylated p53 peptides. *Cell Cycle* 7: 608--610.
  51. Lindahl E, Sansom MSP (2008) Membrane proteins: molecular dynamics simulations. *Current Opinion in Structural Biology* 18: 425-431.
  52. Metcalfe EE, Traaseth NJ, Veglia G (2005) Serine 16 phosphorylation induces an order-to-disorder transition in monomeric phospholamban. *Biochemistry* 44: 4386-4396.
  53. Li JH, Bigelow DJ, Squier TC (2003) Phosphorylation by cAMP-dependent protein kinase modulates the structural coupling between the transmembrane and cytosolic domains of phospholamban. *Biochemistry* 42: 10674-10682.
  54. Sugita Y, Miyashita N, Yoda T, Ikeguchi M, Toyoshima C (2006) Structural changes in the cytoplasmic domain of phospholamban by phosphorylation at Ser16: a molecular dynamics study. *Biochemistry* 45: 11752--11761.
  55. Erxleben C, Liao Y, Gentile S, Chin D, Gomez-Alegria C, et al. (2006) Cyclosporin and Timothy syndrome increase mode 2 gating of CaV1.2 calcium channels through aberrant phosphorylation of S6 helices. *Proc Natl Acad Sci U S A* 103: 3932-3937.

56. Gentile S, Martin N, Scappini E, Williams J, Erxleben C, et al. (2008) The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. *Proc Natl Acad Sci U S A* 105: 14704-14708.
57. Shi Y, Chen X, Wu Z, Shi W, Yang Y, et al. (2008) cAMP-dependent protein kinase phosphorylation produces interdomain movement in SUR2B leading to activation of the vascular KATP channel. *J Biol Chem* 283: 7523--7530.
58. Young MA, Gonfloni S, Superti-Furga G, Roux B, Kuriyan J (2001) Dynamic coupling between the SH2 and SH3 domains of c-Src and hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell* 105: 115-126.
59. Roche P, Mouawad L, Perahia D, Samama JP, Kahn D (2002) Molecular dynamics of the FixJ receiver domain: movement of the beta 4-alpha 4 loop correlates with the in and out flip of Phe101. *Protein Science* 11: 2622-2630.
60. Peters GH, Frimurer TM, Andersen JN, Olsen OH (2000) Molecular dynamics simulations of protein-tyrosine phosphatase 1B. II. Substrate-enzyme interactions and dynamics. *Biophysical Journal* 78: 2191-2200.
61. Stultz CM, Levin AD, Edelman ER (2002) Phosphorylation-induced conformational changes in a mitogen-activated protein kinase substrate - Implications for tyrosine hydroxylase activation. *Journal of Biological Chemistry* 277: 47653-47661.
62. Powell DW, Rane MJ, Joughin BA, Kalmukova R, Hong JH, et al. (2003) Proteomic identification of 14-3-3zeta as a mitogen-activated protein kinase-activated protein kinase 2 substrate: role in dimer formation and ligand binding. *Mol Cell Biol* 23: 5376-5387.
63. Phan-Chan-Du A, Hemmerlin C, Krikorian D, Sakarellos-Daitsiotis M, Tsikaris V, et al. (2003) Solution conformation of the antibody-bound tyrosine phosphorylation site of the nicotinic acetylcholine receptor beta-subunit in its phosphorylated and nonphosphorylated states. *Biochemistry* 42: 7371-7380.
64. Schneider ML, Post CB (1995) Solution structure of a band 3 peptide inhibitor bound to aldolase: A proposed mechanism for regulating binding by tyrosine phosphorylation. *Biochemistry* 34: 16574-16584.
65. Feng MH, Philippopoulos M, MacKerell AD, Lim C (1996) Structural characterization of the phosphotyrosine binding region of a high-affinity SH2

- domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. *Journal of the American Chemical Society* 118: 11265-11277.
66. Tomoo K, Shen X, Okabe K, Nozoe Y, Fukuhara S, et al. (2003) Structural features of human initiation factor 4E, studied by x-ray crystal analyses and molecular dynamics simulations. *Journal of Molecular Biology* 328: 365-383.
  67. Fu Z, Aronoff-Spencer E, Backer JM, Gerfen GJ (2003) The structure of the inter-SH2 domain of class IA phosphoinositide 3-kinase determined by site-directed spin labeling EPR and homology modeling. *Proceedings of the National Academy of Sciences of the United States of America* 100: 3275-3280.
  68. Wozniak-Celmer E, Oldziej S, Ciarkowski J (2001) Theoretical models of catalytic domains of protein phosphatases 1 and 2A with Zn<sup>2+</sup> and Mn<sup>2+</sup> metal dications and putative bioligands in their catalytic centers. *Acta Biochimica Polonica* 48: 35-52.
  69. Jacobson M, Kaminski G, Friesner R, Rapp C (2002) Force field validation using protein side chain prediction. *Journal of Physical Chemistry B* 106: 11673-11680.
  70. Jacobson MP, Friesner RA, Xiang Z, Honig B (2002) On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* 320: 597-608.
  71. Guallar V, Jacobson MP, McDermott A, Friesner RA (2004) Computational Modeling of the Catalytic Reaction in Triose Phosphate Isomerase. *Journal of Molecular Biology* 337: 227-239.
  72. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, et al. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* 55: 351-367.
  73. Coutsiias EA, Seok CL, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. *Journal of Computational Chemistry* 25: 510-528.
  74. Li X, Jacobson MP, Friesner RA (2004) High-resolution prediction of protein helix positions and orientations. *Proteins-Structure Function and Genetics* 55: 368-382.
  75. Harper JW, Adams PD (2001) Cyclin-dependent kinases. *Chemical Reviews* 101: 2511-2526.
  76. Russo AA, Jeffrey PD, Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nature Structural Biology* 3: 696-700.

77. Brown NR, Noble MEM, Lawrie AM, Morris MC, Tunnah P, et al. (1999) Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. *Journal of Biological Chemistry* 274: 8746-8756.
78. Song HW, Hanlon N, Brown NR, Noble MEM, Johnson LN, et al. (2001) Phosphoprotein-protein interactions revealed by the crystal structure of kinase-associated phosphatase in complex with PhosphoCDK2. *Molecular Cell* 7: 615-626.
79. Connell-Crowley L, Solomon MJ, Wei N, Harper JW (1993) Phosphorylation independent activation of human cyclin-dependent kinase 2 by cyclin A in vitro. *Mol Biol Cell* 4: 79-92.
80. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, et al. (1995) Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* 376: 313-320.
81. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605-1612.
82. Huang W, Erikson RL (1994) Constitutive activation of Mek1 by mutation of serine phosphorylation sites. *Proc Natl Acad Sci U S A* 91: 8960-8963.
83. McCabe TJ, Fulton D, Roman LJ, Sessa WC (2000) Enhanced electron flux and reduced calmodulin dissociation may explain "calcium-independent" eNOS activation by phosphorylation. *J Biol Chem* 275: 6123-6128.
84. Klose KE, Weiss DS, Kustu S (1993) Glutamate at the site of phosphorylation of nitrogen-regulatory protein NTRC mimics aspartyl-phosphate and activates the protein. *J Mol Biol* 232: 67-78.
85. Charbon G, Breunig KD, Wattiez R, Vandenhoute J, Noel-Georis I (2004) Key role of Ser562/661 in Snf1-dependent regulation of Cat8p in *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. *Mol Cell Biol* 24: 4083-4091.
86. Kassenbrock CK, Anderson SM (2004) Regulation of ubiquitin protein ligase activity in c-Cbl by phosphorylation-induced conformational change and constitutive activation by tyrosine to glutamate point mutations. *J Biol Chem* 279: 28017-28027.

87. Audette GF, Engelmann R, Hengstenberg W, Deutscher J, Hayakawa K, et al. (2000) The 1.9 Å resolution structure of phospho-serine 46 HPr from *Enterococcus faecalis*. *J Mol Biol* 303: 545-553.
88. Jia Z, Vandonselaar M, Hengstenberg W, Quail JW, Delbaere LT (1994) The 1.6 Å structure of histidine-containing phosphotransfer protein HPr from *Streptococcus faecalis*. *J Mol Biol* 236: 1341-1355.
89. Seavers PR, Lewis RJ, Brannigan JA, Verschueren KH, Murshudov GN, et al. (2001) Structure of the *Bacillus* cell fate determinant SpoIIAA in phosphorylated and unphosphorylated forms. *Structure (Camb)* 9: 605-614.
90. Cho H, Wang W, Kim R, Yokota H, Damo S, et al. (2001) BeF<sub>3</sub>(-) acts as a phosphate analog in proteins phosphorylated on aspartate: structure of a BeF<sub>3</sub>(-) complex with phosphoserine phosphatase. *Proc Natl Acad Sci U S A* 98: 8525-8530.
91. Wang W, Cho HS, Kim R, Jancarik J, Yokota H, et al. (2002) Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states. *J Mol Biol* 319: 421-431.
92. Canagarajah BJ, Khokhlatchev A, Cobb MH, Goldsmith EJ (1997) Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell* 90: 859-869.
93. Zhang F, Strand A, Robbins D, Cobb MH, Goldsmith EJ (1994) Atomic structure of the MAP kinase ERK2 at 2.3 Å resolution. *Nature* 367: 704-711.
94. Birck C, Mourey L, Gouet P, Fabry B, Schumacher J, et al. (1999) Conformational changes induced by phosphorylation of the FixJ receiver domain. *Structure Fold Des* 7: 1505-1515.
95. Gouet P, Fabry B, Guillet V, Birck C, Mourey L, et al. (1999) Structural transitions in the FixJ receiver domain. *Structure Fold Des* 7: 1517-1526.
96. Lewis RJ, Brannigan JA, Muchova K, Barak I, Wilkinson AJ (1999) Phosphorylated aspartate in the structure of a response regulator protein. *J Mol Biol* 294: 9-15.
97. Regni C, Tipton PA, Beamer LJ (2002) Crystal structure of PMM/PGM: an enzyme in the biosynthetic pathway of *P. aeruginosa* virulence factors. *Structure (Camb)* 10: 269-279.



98. Bellon S, Fitzgibbon MJ, Fox T, Hsiao HM, Wilson KP (1999) The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. *Structure Fold Des* 7: 1057-1065.
99. Lahiri SD, Zhang G, Dunaway-Mariano D, Allen KN (2002) Caught in the act: the structure of phosphorylated beta-phosphoglucosyltransferase from *Lactococcus lactis*. *Biochemistry* 41: 8351-8359.
100. Yamaguchi H, Hendrickson WA (1996) Structural basis for activation of human lymphocyte kinase Lck upon tyrosine phosphorylation. *Nature* 384: 484-489.
101. Georgescu RE, Alexov EG, Gunner MR (2002) Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophysical Journal* 83: 1731-1748.
102. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235-242.
103. Sprang SR, Acharya KR, Goldsmith EJ, Stuart DI, Varvill K, et al. (1988) Structural changes in glycogen phosphorylase induced by phosphorylation. *Nature* 336: 215-221.
104. Hubbard SR (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *Embo J* 16: 5572-5581.
105. Hubbard SR, Wei L, Ellis L, Hendrickson WA (1994) Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* 372: 746-754.
106. Jorgensen WL, Maxwell DS, TiradoRives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* 118: 11225-11236.
107. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B* 105: 6474-6487.
108. Schrodinger LLC (1991-2003) Jaguar 5.0. Jaguar. 5.0 ed. Portland, OR: Schrodinger, L. L. C.
109. Tannor DJ, Marten B, Murphy R, Friesner Ra, Sitkoff D, et al. (1994) Accurate First Principles Calculation of Molecular Charge-Distributions and Solvation Energies

- from Ab-Initio Quantum-Mechanics and Continuum Dielectric Theory. *Journal of the American Chemical Society* 116: 11875-11882.
110. Marten B, Kim K, Cortis C, Friesner RA, Murphy RB, et al. (1996) New model for calculation of solvation free energies: Correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. *Journal of Physical Chemistry* 100: 11775-11788.
  111. Abagyan R, Totrov M (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235: 983-1002.
  112. Bruccoleri RE, Karplus M (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 26: 137-168.
  113. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins-Structure Function and Genetics* 51: 41-55.
  114. Nolen BJ, Littlefield RS, Pollard TD (2004) Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP. *Proc Natl Acad Sci U S A* 101: 15627--15632.
  115. Knappik A, Ge L, Honegger A, Pack P, Fischer M, et al. (2000) Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* 296: 57--86.
  116. Stanfield RL, Takimoto-Kamimura M, Rini JM, Profy AT, Wilson IA (1993) Major antigen-induced domain rearrangements in an antibody. *Structure* 1: 83-93.
  117. Vargas-Madrado E, Paz-García E (2003) An improved model of association for VH-VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues. *J Mol Recognit* 16: 113-120.
  118. Banfield MJ, King DJ, Mountain A, Brady RL (1997) VL:VH domain rotations in engineered antibodies: crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins* 29: 161-171.

119. Nakanishi T, Tsumoto K, Yokota A, Kondo H, Kumagai I (2008) Critical contribution of VH-VL interaction to reshaping of an antibody: the case of humanization of anti-lysozyme antibody, HyHEL-10. *Protein Sci* 17: 261--270.
120. Chatellier J, Van Regenmortel MH, Vernet T, Altschuh D (1996) Functional mapping of conserved residues located at the VL and VH domain interface of a Fab. *J Mol Biol* 264: 1-6.
121. Khalifa MB, Weidenhaupt M, Choulier L, Chatellier J, Rauffer-Bruyère N, et al. (2000) Effects on interaction kinetics of mutations at the VH-VL interface of Fabs depend on the structural context. *J Mol Recognit* 13: 127-139.
122. Glasner ME, Fayazmanesh N, Chiang RA, Sakai A, Jacobson MP, et al. (2006) Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. *J Mol Biol* 360: 228-250.
123. Bellon S, Fitzgibbon MJ, Fox T, Hsiao HM, Wilson KP (1999) The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. *Structure* 7: 1057-1065.
124. Frantz B, Klatt T, Pang M, Parsons J, Rolando A, et al. (1998) The activation state of p38 mitogen-activated protein kinase determines the efficiency of ATP competition for pyridinylimidazole inhibitor binding. *Biochemistry* 37: 13846-13853.
125. Sours KM, Kwok SC, Rachidi T, Lee T, Ring A, et al. (2008) Hydrogen-exchange mass spectrometry reveals activation-induced changes in the conformational mobility of p38alpha MAP kinase. *J Mol Biol* 379: 1075--1093.
126. Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* 273: 927-948.
127. Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196: 901-917.
128. Martin AC, Thornton JM (1996) Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* 263: 800-815.

129. Carter PJ (2006) Potent antibody therapeutics by design. *Nat Rev Immunol* 6: 343-357.
130. Jain M, Kamal N, Batra SK (2007) Engineering antibodies for clinical applications. *Trends Biotechnol* 25: 307-316.
131. Presta L (2003) Antibody engineering for therapeutics. *Curr Opin Struct Biol* 13: 519-525.
132. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, et al. (2006) Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci* 15: 949--960.
133. Ewert S, Honegger A, Pluckthun A (2003) Structure-based improvement of the biophysical properties of immunoglobulin VH domains with a generalizable approach. *Biochemistry* 42: 1517--1528.
134. Ewert S, Honegger A, Pluckthun A (2004) Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering. *Methods* 34: 184--199.
135. Sivasubramanian A, Maynard JA, Gray JJ (2008) Modeling the structure of mAb 14B7 bound to the anthrax protective antigen. *Proteins* 70: 218--230.
136. Barderas R, Desmet J, Timmerman P, Meloen R, Casal J (2008) Affinity maturation of antibodies assisted by in silico modeling. *Proc Natl Acad Sci U S A* 105: 9029-9034.
137. Fontayne A, De Maeyer B, De Maeyer M, Yamashita M, Matsushita T, et al. (2007) Paratope and epitope mapping of the antithrombotic antibody 6B4 in complex with platelet glycoprotein Ibalph. *J Biol Chem* 282: 23517--23524.
138. Fontayne A, Vanhoorelbeke K, Pareyn I, Van Rompaey I, Meiring M, et al. (2006) Rational humanization of the powerful antithrombotic anti-GPIbalph antibody: 6B4. *Thromb Haemost* 96: 671--684.
139. Kuroda D, Shirai H, Kobori M, Nakamura H (2008) Structural classification of CDR-H3 revisited: A lesson in antibody modeling. *Proteins* 75: 139-146.
140. Morea V, Lesk AM, Tramontano A (2000) Antibody modeling: implications for engineering and design. *Methods* 20: 267-279.

141. Whitelegg N, Rees AR (2004) Antibody variable regions: toward a unified modeling method. *Methods Mol Biol* 248: 51-91.
142. Whitelegg NR, Rees AR (2000) WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng* 13: 819-824.
143. Sivasubramanian A, Sircar A, Chaudhury S, J. GJ (2009) Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins* 74: 497-514.
144. Chothia C, Novotný J, Bruccoleri R, Karplus M (1985) Domain association in immunoglobulin molecules. The packing of variable domains. *J Mol Biol* 186: 651-663.
145. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP (2008) Toward better refinement of comparative models: Predicting loops in inexact environments. *Proteins* 72: 959-971.
146. Eigenbrot C, Randal M, Presta L, Carter P, Kossiakoff AA (1993) X-ray structures of the antigen-binding domains from three variants of humanized anti-p185HER2 antibody 4D5 and comparison with molecular modeling. *J Mol Biol* 229: 969-995.
147. Andrusier N, Nussinov R, Wolfson HJ (2007) FireDock: Fast interaction refinement in molecular docking. *Proteins* 69: 139-159.
148. May A, Zacharias M (2005) Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim Biophys Acta* 1754: 225-231.
149. McLean GR, Torres M, Elguezabal N, Nakouzi A, Casadevall A (2002) Isotype can affect the fine specificity of an antibody for a polysaccharide antigen. *J Immunol* 169: 1379-1386.
150. Torres M, Casadevall A (2008) The immunoglobulin constant region contributes to affinity and specificity. *Trends Immunol* 29: 91-97.
151. Torres M, Fernandez-Fuentes N, Fiser A, Casadevall A (2007) The immunoglobulin heavy chain constant region affects kinetic and thermodynamic parameters of antibody variable region interactions with antigen. *J Biol Chem* 282: 13917-27.

152. Torres M, Fernandez-Fuentes N, Fiser A, Casadevall A (2007) Exchanging murine and human immunoglobulin constant chains affects the kinetics and thermodynamics of antigen binding and chimeric antibody autoreactivity. *PLoS ONE* 2: e1310.
153. Torres M, May R, Scharff MD, Casadevall A (2005) Variable-region-identical antibodies differing in isotype demonstrate differences in fine specificity and idiotype. *J Immunol* 174: 2132-2142.
154. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673-4680.
155. Chapman B, Chang J (2000) Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter* 20: 15-19.
156. Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. *Proteins* 17: 49-61.
157. Gray JJ (2006) High-resolution protein-protein docking. *Curr Opin Struct Biol* 16: 183-193.
158. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47: 409-443.
159. Vajda S, Camacho CJ (2004) Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol* 22: 110-116.
160. Zhu K, Shirts MR, Friesner RA, Jacobson MP (2007) Multiscale optimization of a truncated Newton minimization algorithm and application to proteins and protein-ligand complexes. *Journal of Chemical Theory and Computation* 3: 640-648.
161. Pollard TD (2007) Regulation of actin filament assembly by Arp2/3 complex and formins. *Annu Rev Biophys Biomol Struct* 36: 451--477.
162. Sept D, McCammon JA (2001) Thermodynamics and kinetics of actin filament nucleation. *Biophys J* 81: 667-674.
163. Baum B, Kunda P (2005) Actin nucleation: spire - actin nucleator in a class of its own. *Curr Biol* 15: R305-308.

164. Kovar DR (2006) Molecular details of formin-mediated actin assembly. *Curr Opin Cell Biol* 18: 11-17.
165. Amann KJ, Pollard TD (2001) Direct real-time observation of actin filament branching mediated by Arp2/3 complex using total internal reflection fluorescence microscopy. *Proc Natl Acad Sci U S A* 98: 15009-15013.
166. Mullins RD, Heuser JA, Pollard TD (1998) The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proc Natl Acad Sci U S A* 95: 6181-6186.
167. Goley ED, Welch MD (2006) The ARP2/3 complex: an actin nucleator comes of age. *Nat Rev Mol Cell Biol* 7: 713--726.
168. Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. *Cell* 112: 453-465.
169. Wang W, Goswami S, Sahai E, Wyckoff JB, Segall JE, et al. (2005) Tumor cells caught in the act of invading: their strategy for enhanced cell motility. *Trends Cell Biol* 15: 138-145.
170. Goley ED, Rodenbusch SE, Martin AC, Welch MD (2004) Critical conformational changes in the Arp2/3 complex are induced by nucleotide and nucleation promoting factor. *Mol Cell* 16: 269--279.
171. Martin AC, Xu X-P, Rouiller I, Kaksonen M, Sun Y, et al. (2005) Effects of Arp2 and Arp3 nucleotide-binding pocket mutations on Arp2/3 complex function. *J Cell Biol* 168: 315--328.
172. Welch MD, Mullins RD (2002) Cellular control of actin nucleation. *Annu Rev Cell Dev Biol* 18: 247--288.
173. Winter D, Lechler T, Li R (1999) Activation of the yeast Arp2/3 complex by Bee1p, a WASP-family protein. *Curr Biol* 9: 501-504.
174. Yarar D, To W, Abo A, Welch MD (1999) The Wiskott-Aldrich syndrome protein directs actin-based motility by stimulating actin nucleation with the Arp2/3 complex. *Curr Biol* 9: 555-558.
175. Rohatgi R, Ho HY, Kirschner MW (2000) Mechanism of N-WASP activation by CDC42 and phosphatidylinositol 4, 5-bisphosphate. *J Cell Biol* 150: 1299-1310.

176. Machesky LM, Mullins RD, Higgs HN, Kaiser DA, Blanchoin L, et al. (1999) Scar, a WASP-related protein, activates nucleation of actin filaments by the Arp2/3 complex. *Proc Natl Acad Sci U S A* 96: 3739--3744.
177. Bompard G, Caron E (2004) Regulation of WASP/WAVE proteins: making a long story short. *J Cell Biol* 166: 957-962.
178. Stradal TEB, Scita G (2006) Protein complexes regulating Arp2/3-mediated actin assembly. *Curr Opin Cell Biol* 18: 4-10.
179. Welch MD, Rosenblatt J, Skoble J, Portnoy DA, Mitchison TJ (1998) Interaction of human Arp2/3 complex and the *Listeria monocytogenes* ActA protein in actin filament nucleation. *Science* 281: 105-108.
180. Gouin E, Egile C, Dehoux P, Villiers V, Adams J, et al. (2004) The RickA protein of *Rickettsia conorii* activates the Arp2/3 complex. *Nature* 427: 457-461.
181. Jeng RL, Goley ED, D'Alessio JA, Chaga OY, Svitkina TM, et al. (2004) A *Rickettsia* WASP-like protein activates the Arp2/3 complex and mediates actin-based motility. *Cell Microbiol* 6: 761-769.
182. LeClaire LLr, Baumgartner M, Iwasa JH, Mullins RD, Barber DL (2008) Phosphorylation of the Arp2/3 complex is necessary to nucleate actin filaments. *J Cell Biol* 182: 647--654.
183. Robinson RC, Turbedsky K, Kaiser DA, Marchand JB, Higgs HN, et al. (2001) Crystal structure of Arp2/3 complex. *Science* 294: 1679-1684.
184. Rodal AA, Sokolova O, Robins DB, Daugherty KM, Hippenmeyer S, et al. (2005) Conformational changes in the Arp2/3 complex leading to actin nucleation. *Nat Struct Mol Biol* 12: 26--31.
185. Rouiller I, Xu X, Amann K, Egile C, Nickell S, et al. (2008) The structural basis of actin filament branching by the Arp2/3 complex. *J Cell Biol* 180: 887-895.
186. García (1992) Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* 68: 2696-2699.
187. Hayward S, Kitao A, Go N (1995) Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins* 23: 177-186.



188. Barrett CP, Hall BA, Noble ME (2004) Dynamite: a simple way to gain insight into protein motions. *Acta Crystallogr D Biol Crystallogr* 60: 2280-2287.
189. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14: 33-38, 27-38.
190. Balsera M, Wriggers W, Oono Y, Schulten K (1996) Principal component analysis and long time protein dynamics. *Journal of Physical Chemistry* 100: 2567-2572.
191. Egile C, Rouiller I, Xu X-P, Volkmann N, Li R, et al. (2005) Mechanism of filament nucleation and branch stability revealed by the structure of the Arp2/3 complex at actin branch junctions. *PLoS Biol* 3: e383.
192. Dalhaimer P, Pollard TD, Nolen BJ (2007) Nucleotide-Mediated Conformational Changes of Monomeric Actin and Arp3 Studied by Molecular Dynamics Simulations. *J Mol Biol* 376: 166-183.
193. Pfaendtner J, Voth GA (2008) Molecular dynamics simulation and coarse-grained analysis of the Arp2/3 complex. *Biophys J* 95: 5324--5333.
194. Beltzner CC, Pollard TD (2004) Identification of functionally important residues of Arp2/3 complex by analysis of homology models from diverse species. *J Mol Biol* 336: 551--565.
195. Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC (1990) Atomic structure of the actin:DNase I complex. *Nature* 347: 37--44.
196. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters; 2006 November 11-17; Tampa, Florida.
197. Martyna GJ, Tobias DJ, Klein ML (1994) Constant-pressure molecular-dynamics algorithms. *Journal of Chemical Physics* 101: 4177-4189.
198. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22: 2695-2696.
199. R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <http://www.R-project.org> .

200. Liwo A, Czaplewski C, Oldziej S, Scheraga HA (2008) Computational techniques for efficient conformational sampling of proteins. *Curr Opin Struct Biol* 18: 134--139.
201. Jiang J, Suppiramaniam V, Wooten MW (2006) Posttranslational modifications and receptor-associated proteins in AMPA receptor trafficking and synaptic plasticity. *Neurosignals* 15: 266-282.
202. Chen BS, Roche KW (2007) Regulation of NMDA receptors by phosphorylation. *Neuropharmacology* 53: 362-368.
203. Marie N, Aguila B, Allouche S (2006) Tracking the opioid receptors on the way of desensitization. *Cell Signal* 18: 1815-1833.
204. Xu ZC, Yang Y, Hebert SC (1996) Phosphorylation of the ATP-sensitive, inwardly rectifying K<sup>+</sup> channel, ROMK, by cyclic AMP-dependent protein kinase. *J Biol Chem* 271: 9313--9319.
205. MacGregor GG, Xu JZ, McNicholas CM, Giebisch G, Hebert SC (1998) Partially active channels produced by PKA site mutation of the cloned renal K<sup>+</sup> channel, ROMK2 (kir1.2). *Am J Physiol* 275: F415-422.
206. Liou HH, Zhou SS, Huang CL (1999) Regulation of ROMK1 channel by protein kinase A via a phosphatidylinositol 4,5-bisphosphate-dependent mechanism. *Proc Natl Acad Sci U S A* 96: 5820--5825.
207. Kuo A, Gulbis JM, Antcliff JF, Rahman T, Lowe ED, et al. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* 300: 1922--1926.
208. Nishida M, Cadene M, Chait B, Mackinnon R (2007) Crystal structure of a Kir3.1-prokaryotic Kir channel chimera. *EMBO J* 26: 4005-4015.
209. Antcliff JF, Haider S, Proks P, Sansom MSP, Ashcroft FM (2005) Functional analysis of a structural model of the ATP-binding site of the KATP channel Kir6.2 subunit. *EMBO J* 24: 229--239.
210. Haider S, Tarasov AI, Craig TJ, Sansom MSP, Ashcroft FM (2007) Identification of the PIP(2)-binding site on Kir6.2 by molecular modelling and functional analysis. *EMBO J* 26: 3749--3759.

211. Haider S, Khalid S, Tucker SJ, Ashcroft FM, Sansom MSP (2007) Molecular dynamics simulations of inwardly rectifying (Kir) potassium channels: a comparative study. *Biochemistry* 46: 3643--3652.
212. Lee C, Huang P, Lou K, Liou H (2008) Functional and structural characterization of PKA-mediated pH(i) gating of ROMK1 channels. *J Mol Graph Model* 27: 332-341.
213. Fakler B, Schultz JH, Yang J, Schulte U, Brandle U, et al. (1996) Identification of a titratable lysine residue that determines sensitivity of kidney potassium channels (ROMK) to intracellular pH. *EMBO J* 15: 4093-4099.
214. Rapedius M, Paynter JJ, Fowler PW, Shang L, Sansom MSP, et al. (2007) Control of pH and PIP2 gating in heteromeric Kir4.1/Kir5.1 channels by H-Bonding at the helix-bundle crossing. *Channels (Austin)* 1: 327--330.
215. Rapedius M, Fowler PW, Shang L, Sansom MSP, Tucker SJ, et al. (2007) H bonding at the helix-bundle crossing controls gating in Kir potassium channels. *Neuron* 55: 602--614.
216. Rapedius M, Haider S, Browne KF, Shang L, Sansom MSP, et al. (2006) Structural and functional analysis of the putative pH sensor in the Kir1.1 (ROMK) potassium channel. *EMBO Rep* 7: 611--616.
217. Seibert FS, Chang XB, Aleksandrov AA, Clarke DM, Hanrahan JW, et al. (1999) Influence of phosphorylation by protein kinase A on CFTR at the cell surface and endoplasmic reticulum. *Biochim Biophys Acta* 1461: 275-283.
218. Chappe V, Irvine T, Liao J, Evagelidis A, Hanrahan JW (2005) Phosphorylation of CFTR by PKA promotes binding of the regulatory domain. *EMBO J* 24: 2730-2740.
219. Hegedus T, Serohijos AWR, Dokholyan NV, He L, Riordan JR (2008) Computational studies reveal phosphorylation-dependent changes in the unstructured R domain of CFTR. *J Mol Biol* 378: 1052--1063.
220. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI (1998) Discrete molecular dynamics studies of the folding of a protein-like model. *Fold Des* 3: 577-587.

221. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI (2000) Identifying the protein folding nucleus using molecular dynamics. *J Mol Biol* 296: 1183-1188.
222. Bao X, Reuss L, Altenberg GA (2004) Regulation of purified and reconstituted connexin 43 hemichannels by protein kinase C-mediated phosphorylation of Serine 368. *J Biol Chem* 279: 20058-20066.
223. Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, et al. (2004) Ca(V)<sub>1</sub>.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* 119: 19-31.
224. Medzihradszky KF, Darula Z, Perlson E, Fainzilber M, Chalkley RJ, et al. (2004) O-sulfonation of serine and threonine: mass spectrometric detection and characterization of a new posttranslational modification in diverse proteins throughout the eukaryotes. *Mol Cell Proteomics* 3: 429-440.
225. Sims RJ, 3rd, Reinberg D (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat Rev Mol Cell Biol* 9: 815-820.
226. Burlingame AL, Zhang X, Chalkley RJ (2005) Mass spectrometric analysis of histone posttranslational modifications. *Methods* 36: 383-394.
227. Zhang Y, Reinberg D (2001) Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* 15: 2343-2360.
228. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol* 14: 1025--1040.
229. Ruthenburg AJ, Li H, Patel DJ, Allis CD (2007) Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* 8: 983--994.
230. Botuyan MV, Lee J, Ward IM, Kim JE, Thompson JR, et al. (2006) Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. *Cell* 127: 1361-1373.
231. Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, et al. (2003) Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev* 17: 1870-1881.

232. Flanagan JF, Mi LZ, Chruszcz M, Cymborowski M, Clines KL, et al. (2005) Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* 438: 1181-1185.
233. Huang Y, Fang J, Bedford MT, Zhang Y, Xu RM (2006) Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* 312: 748-751.
234. Jacobs SA, Khorasanizadeh S (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* 295: 2080-2083.
235. Jacobs SA, Taverna SD, Zhang Y, Briggs SD, Li J, et al. (2001) Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J* 20: 5232-5241.
236. Li H, Fischle W, Wang W, Duncan EM, Liang L, et al. (2007) Structural basis for lower lysine methylation state-specific readout by MBT repeats of L3MBTL1 and an engineered PHD finger. *Mol Cell* 28: 677-691.
237. Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, et al. (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416: 103-107.
238. Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV, et al. (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* 442: 100-103.
239. Hughes RM, Wiggins KR, Khorasanizadeh S, Waters ML (2007) Recognition of trimethyllysine by a chromodomain is not driven by the hydrophobic effect. *Proc Natl Acad Sci U S A* 104: 11184-11188.

# Appendix A. Supplementary Tables and Figures for Chapter 2

Phosphorylated Threonine		
Atom ID	Charge in -2 State	Charge in -1 State
CB	0.280	0.280
HB	-0.030	-0.030
CG2	-0.210	-0.210
1HG2	0.060	0.060
2HG2	0.060	0.060
3HG2	0.060	0.060
OG1	-0.700	-0.500
P	1.570	1.530
O1P	-1.030	-1.000
O2P	-1.030	-1.000
O3P	-1.030	-0.750
HO	NA	0.500

**Supplementary Table A.1 Partial atomic charges employed for phosphothreonine in its unprotonated (-2 charge) and protonated (-1 charge) states.**

Phosphorylated Serine		
Atom ID	Charge in -2 State	Charge in -1 State
CB	0.280	0.280
1HB	-0.030	0.035
2HB	-0.030	0.035
OG	-0.700	-0.600
P	1.570	1.500
O1P	-1.030	-1.000
O2P	-1.030	-1.000
O3P	-1.030	-0.750
HO	NA	0.500

**Supplementary Table A2 Partial atomic charges employed for phosphoserine in its unprotonated (-2 charge) and protonated (-1 charge) states.**

Phosphorylated Tyrosine		
Atom ID	Charge in -2 State	Charge in -1 State
CB	-0.005	-0.005
1HB	0.060	0.060
2HB	0.060	0.060
CG	-0.115	-0.115
CD1	-0.115	-0.115
HD1	0.115	0.115
CE1	-0.115	-0.115
HE1	0.115	0.115
CZ	0.220	0.240
CE2	-0.115	-0.115
HE2	0.115	0.115
CD2	-0.115	-0.115
HD2	0.115	0.115
OH	-0.700	-0.600
P	1.570	1.550
O1P	-1.030	-0.970
O2P	-1.030	-0.970
O3P	-1.030	-0.750
HO	NA	0.500

**Supplementary Table A3 Partial atomic charges employed for phosphotyrosine in its unprotonated (-2 charge) and protonated (-1 charge) states.**

Phosphorylated Aspartate		
Atom ID	Charge in -2 State	Charge in -1 State
CB	-0.470	-0.440
1HB	0.220	0.230
2HB	0.220	0.230
CG	0.940	0.870
OD1	-0.590	-0.540
OD2	-0.730	-0.670
P	1.470	1.500
O1P	-1.020	-0.950
O2P	-1.020	-0.950
O3P	-1.020	-0.790
HO	NA	0.510

**Supplementary Table A4 Partial atomic charges employed for phosphoaspartate in its unprotonated (-2 charge) and protonated (-1 charge) states.**



## Appendix B. Supplementary Tables and Figures for Chapter 3

15c8	1bfv	1clz	1ehl	1fve	1iai	1jrh	1m7d	1nbz	1oak	1qkz	25c8
1a14	1bgx	1cr9	1ejo	1g7h	1ic4	1k4c	1m7i	1nc2	1oag	1qyg	2ap2
1a2y	1bj1	1ct8	1emt	1g7i	1ic5	1k4d	1mam	1nc4	1obl	1r3i	2bfv
1a3l	1bln	1cu4	1eo8	1g7m	1ic7	1k6q	1mf2	1nea	1oew	1r3j	2cgr
1a3r	1bm3	1cz8	1etz	1g9m	1ifh	1kc5	1mfb	1ncb	1om3	1r3k	2ck0
1a4j	1bog	1d5b	1ezv	1g9n	1igc	1ker	1mfc	1ncc	1op3	1r3l	2dbl
1a5f	1bql	1d5i	1f3d	1gc1	1igf	1kcs	1mfd	1ncd	1opg	1rfd	2dlf
1a6t	1bvk	1d6v	1f4w	1ggb	1igj	1kcu	1mh5	1ncw	1ors	1rih	2f19
1a6u	1bvl	1dba	1f4x	1ggc	1igm	1kev	1mhp	1nd0	1p2c	1riu	2f58
1a7n	1bz7	1dbb	1f4y	1ggi	1ikf	1keg	1mim	1ndg	1pg7	1riv	2f5a
1a7o	1c08	1dbj	1f58	1ghf	1il1	1kel	1mj8	1ndm	1plg	1rmf	2f5b
1a7p	1cle	1dbm	1f8t	1gig	1iqd	1kem	1mjj	1nfd	1pz5	1rz8	2h1p
1a7q	1c5b	1dee	1f90	1gpo	1iqw	1kfa	1mju	1ngp	1q1j	1rzf	2hrp
1a7r	1c5c	1dfb	1fai	1h3p	1it9	1kip	1mlb	1ngq	1q72	1rzj	2iff
1ad0	1c5d	1dl7	1fdl	1hh6	1j05	1kiq	1mlc	1ngw	1q9k	1rzk	2igf
1ad9	1cbv	1dlf	1fe8	1hi6	1j1o	1kir	1mnu	1ngx	1q9l	1tet	2jel
1ai1	1ce1	1dn0	1fgn	1hil	1j1p	1kn2	1mpa	1ngy	1q9o	1vfa	2mpa
1aif	1cfn	1dqj	1fgv	1him	1j1x	1kn4	1mqk	1ngz	1q9q	1vfb	2pcp
1ap2	1cfq	1dql	1fh5	1hyx	1jfq	1l7i	1mvu	1nj9	1q9r	1vge	35c8
1axt	1cft	1dqm	1fj1	1hyy	1jgu	1l7t	1n4x	1nl0	1q9t	1wej	3f58
1ay1	1cfv	1dqq	1fl3	1i7z	1jgv	1lk3	1n64	1nlb	1q9v	1yec	3fct
1b2w	1cgs	1dsf	1fns	1i8i	1jhl	1lo0	1n7m	1nld	1q9w	1yee	43c9
1b4j	1cic	1e4x	1frg	1i8k	1jnh	1lo2	1n8z	1nmb	1qbl	1yef	43ca
1bbd	1ck0	1e6o	1fvc	1i8m	1jps	1lo3	1nbv	1nmc	1qbm	1yeg	6fab
1bfo	1clo	1eap	1fvd	1i9j	1jpt	1lo4	1nby	1nsn	1qfu	1yeh	7fab

Supplementary Table B1 PDB IDs of antibodies in the full 300-member dataset

15c8	1bfv	1clz	1ehl	1fve	1iai	1jrh	1m7d	1nbz	1oak	1qkz	25c8
1a14	1bgx	1cr9	1ejo	1g7h	1ic4	1k4c	1m7i	1nc2	1oaq	1qyg	2ap2
1a2y	1bj1	1ct8	1emt	1g7i	1ic5	1k4d	1mam	1nc4	1ob1	1r3i	2bfv
1a3l	1bln	1cu4	1eo8	1g7m	1ic7	1k6q	1mf2	1nca	1ocw	1r3j	2cgr
1a3r	1bm3	1cz8	1etz	1g9m	1ifh	1kc5	1mfb	1ncb	1om3	1r3k	2ck0
1a4j	1bog	1d5b	1ezv	1g9n	1igc	1kcr	1mfc	1ncc	1op3	1r3l	2dbl
1a5f	1bql	1d5i	1f3d	1gc1	1igf	1kes	1mfd	1ncd	1opg	1rfd	2dlf
1a6t	1bvk	1d6v	1f4w	1ggg	1igj	1kcu	1mh5	1ncw	1ors	1rih	2f19
1a6u	1bvl	1dba	1f4x	1ggc	1igm	1kcv	1mhp	1nd0	1p2c	1riu	2f58
1a7n	1bz7	1dbb	1f4y	1ggi	1ikf	1keg	1mim	1ndg	1pg7	1riv	2f5a
1a7o	1c08	1dbj	1f58	1ghf	1il1	1kel	1mj8	1ndm	1plg	1rmf	2f5b
1a7p	1cle	1dbm	1f8t	1gig	1iqd	1kem	1mjj	1nfd	1pz5	1rz8	2h1p
1a7q	1c5b	1dee	1f90	1gpo	1iqw	1kfa	1mju	1ngp	1q1j	1rzf	2hrp
1a7r	1c5c	1dfb	1fai	1h3p	1it9	1kip	1mlb	1ngq	1q72	1rzj	2iff
1ad0	1c5d	1dl7	1fdl	1hh6	1j05	1kiq	1mlc	1ngw	1q9k	1rzk	2igf
1ad9	1cbv	1dlf	1fe8	1hi6	1j1o	1kir	1mnu	1ngx	1q9l	1tet	2jel
1ai1	1ce1	1dn0	1fgn	1hil	1j1p	1kn2	1mpa	1ngy	1q9o	1vfa	2mpa
1aif	1cfn	1dqj	1fgv	1him	1j1x	1kn4	1mqk	1ngz	1q9q	1vfb	2pcp
1ap2	1cfq	1dql	1fh5	1hyx	1jfq	1l7i	1mvu	1nj9	1q9r	1vge	35c8
1axt	1cft	1dqm	1fj1	1hyy	1jgu	1l7t	1n4x	1nl0	1q9t	1wej	3f58
1ay1	1cfv	1dqq	1fl3	1i7z	1jgv	1lk3	1n64	1nlb	1q9v	1yec	3fct
1b2w	1cgs	1dsf	1fns	1i8i	1jhl	1lo0	1n7m	1nld	1q9w	1yee	43c9
1b4j	1cic	1e4x	1frg	1i8k	1jnh	1lo2	1n8z	1nmb	1qbl	1yef	43ca
1bbd	1ck0	1e6o	1fvc	1i8m	1jps	1lo3	1nbv	1nmc	1qbm	1yeg	6fab
1bfo	1clo	1eap	1fvd	1i9j	1jpt	1lo4	1nby	1nsn	1qfu	1yeh	7fab

**Supplementary Table B2 PDB IDs of antibodies in the non-redundant dataset.**

The first nine antibodies (highlighted in red) in the first column were chosen as part of the test set for energy-based calculations by virtue of: a) having <90% sequence identity over the full Fv region and over the interface residues (see Methods); b) being crystallized as Fv fragments in the absence of antigen. These were used to choose structures from the full dataset such that any two antibodies in this set have <90% sequence identity over both the Fv region and the interface residues. Eleven more antibodies were chosen randomly for energy-based calculations and are highlighted in blue.

PDB1	PDB2	RMSD	PDB1	PDB2	RMSD
12e8(L,H)	12e8(M,P)	3.1	1nj9(L,H)	1nj9(A,B)	0.5
1ruq(L,H)	1rur(L,H)	0.1	1oaq(L,H)	1ocw(L,H)	1.7
1a4j(L,H)	1a4j(A,B)	0.3	2q76(A,B)	2q76(C,D)	0.6
1a6t(A,B)	1a6t(C,D)	0.3	1q9o(A,B)	1q9o(C,D)	0.8
1ad0(A,B)	1ad0(C,D)	0.2	1qbl(L,H)	1qbm(L,H)	0.6
1aif(L,H)	1aif(A,B)	0.2	1qlr(A,B)	1qlr(C,D)	0.2
1bfo(C,D)	1bfo(G,H)	0.3	1rhh(A,B)	1rhh(C,D)	0.6
1bln(A,B)	1bln(C,D)	0.4	1rz8(A,B)	1rz8(C,D)	0.7
1bm3(L,H)	1opg(L,H)	0.5	1rzi(E,F)	1rzi(I,J)	0.9
1bvl(B,A)	1bvl(D,C)	1.0	1t4k(A,B)	1t4k(C,D)	0.3
1c5d(L,H)	1c5d(A,B)	0.4	1ub6(L,H)	1ub6(B,A)	0.9
1d5b(A,B)	1d5b(L,H)	1.0	1uwe(U,V)	1uwe(X,Y)	0.6
1dlf(L,H)	2dlf(L,H)	0.4	1uyw(L,H)	1uyw(N,M)	0.4
1dn0(A,B)	1dn0(C,D)	0.4	1wt5(C,A)	1wt5(D,B)	1.2
1dqq(A,B)	1dqm(L,H)	0.6	3cfj(E,F)	3cfk(G,I)	1.2
1f8t(L,H)	1s5i(L,H)	0.8	1yy8(A,B)	1yy8(C,D)	0.1
1fai(L,H)	2f19(L,H)	0.2	1za6(E,F)	1za6(G,H)	0.9
1ggb(L,H)	1ggc(L,H)	0.3	2aj3(A,B)	2aj3(C,D)	0.5
1gpo(L,H)	1gpo(M,I)	1.0	2g75(B,A)	2g75(D,C)	0.0
1hil(A,B)	1hil(C,D)	0.4	2fjf(C,D)	2fjf(S,T)	1.1
1hzh(L,H)	1hzh(M,K)	0.2	2gjz(A,B)	2gk0(L,H)	0.9
1xf3(L,H)	1xf4(A,B)	1.7	2gsg(A,B)	2gsg(C,D)	1.1
1igf(L,H)	1igf(M,J)	0.1	2iqa(L,H)	2iqa(A,B)	0.9
1igt(A,B)	1igt(C,D)	0.5	2orb(L,H)	2orb(M,I)	1.0
1j05(L,H)	1j05(A,B)	0.8	2z91(B,A)	2z91(D,C)	0.1
1mf2(L,H)	1mf2(M,N)	0.3	3cfe(L,H)	3cfe(A,B)	0.1
1n4x(L,H)	1n4x(M,I)	1.4	43c9(C,D)	43c9(G,H)	0.6
1ruk(L,H)	1rum(L,H)	0.1	8fab(A,B)	8fab(C,D)	1.3

**Supplementary Table B3 Difference in orientation comparing unbound structures of identical antibodies**

PDB1	PDB2	RMSD	PDB1	PDB2	RMSD
15c8(L,H)	25c8(L,H)	1.0	1m71(A,B)	1pz5(A,B)	1.0
1a2y(A,B)	1vfa(A,B)	1.2	1mf2(L,H)	2hrp(L,H)	1.6
1a3l(L,H)	1rur(L,H)	0.8	1mhh(C,D)	1nlb(L,H)	0.7
1a3r(L,H)	1bbd(L,H)	0.5	1mju(L,H)	1mjj(A,B)	0.7
1aj7(L,H)	2res(L,H)	1.4	1mlb(A,B)	1mlc(A,B)	1.5
2ap2(A,B)	1ap2(A,B)	1.5	1mnu(L,H)	2mpa(L,H)	1.0
1ar1(D,C)	1mqk(L,H)	0.9	1nd0(G,H)	1ruk(L,H)	0.7
1ay1(L,H)	1bgx(L,H)	1.0	1ngx(L,H)	1ngz(A,B)	2.6
1cfq(A,B)	1cfs(A,B)	1.7	2bjm(L,H)	1oaq(L,H)	3.6
1bvk(A,B)	1bvl(B,A)	1.3	1om3(L,H)	1zlu(L,H)	0.9
1c5b(L,H)	1c5c(L,H)	0.4	1q0x(L,H)	1q0y(L,H)	1.0
1cbv(L,H)	1nbv(L,H)	2.0	1rfd(L,H)	1riv(L,H)	0.3
1cgs(L,H)	1ynk(L,H)	0.7	1q9l(C,D)	1q9q(A,B)	1.1
1ck0(L,H)	2ck0(L,H)	0.2	1qbm(L,H)	1wej(L,H)	0.9
1ucb(L,H)	1clz(L,H)	0.6	1yy1(L,H)	1rz8(A,B)	1.2
1cr9(L,H)	1cu4(L,H)	1.1	1um6(L,H)	1um4(L,H)	1.1
1d5i(L,H)	1d6v(L,H)	0.4	1y0l(E,F)	3cfj(E,F)	1.3
1dba(L,H)	1dbj(L,H)	0.4	1yy8(A,B)	1yy9(C,D)	1.1
1dqj(A,B)	1dqq(A,B)	1.1	2adg(A,B)	2adi(A,B)	0.8
1f4w(L,H)	1f4y(L,H)	0.6	2ajz(L,H)	2aju(L,H)	1.0
1hil(A,B)	1him(H,L)	1.3	2dd8(L,H)	2g75(B,A)	0.8
1hzh(M,K)	1n0x(M,K)	0.4	2dqt(L,H)	2dtm(L,H)	1.0
1xf4(A,B)	2fr4(A,B)	2.8	2eh8(L,H)	2eh7(L,H)	1.0
1igf(L,H)	2igf(L,H)	0.7	2fat(L,H)	2fd6(L,H)	1.6
1igi(L,H)	1igj(C,D)	0.7	2fjf(S,T)	2fjg(A,B)	2.3
1jgu(L,H)	1jgv(L,H)	0.4	2iqa(A,B)	2r0w(L,H)	1.2
1jn6(A,B)	1jnh(E,F)	0.3	2o5x(L,H)	2o5z(L,H)	0.5
1jpt(L,H)	1pg7(M,I)	0.9	2orb(L,H)	2or9(M,I)	1.6
1kc5(L,H)	1kcu(L,H)	0.2	2z92(B,A)	2z91(D,C)	0.3
1kcs(L,H)	1kcv(L,H)	1.0	3cfb(L,H)	3cfc(L,H)	1.4
1kel(L,H)	1kem(L,H)	0.2	3cfd(L,H)	3cfe(L,H)	0.4
1yeh(L,H)	1yek(L,H)	0.2	43c9(G,H)	43ca(C,D)	0.6
1l7t(L,H)	1vpo(L,H)	0.9			

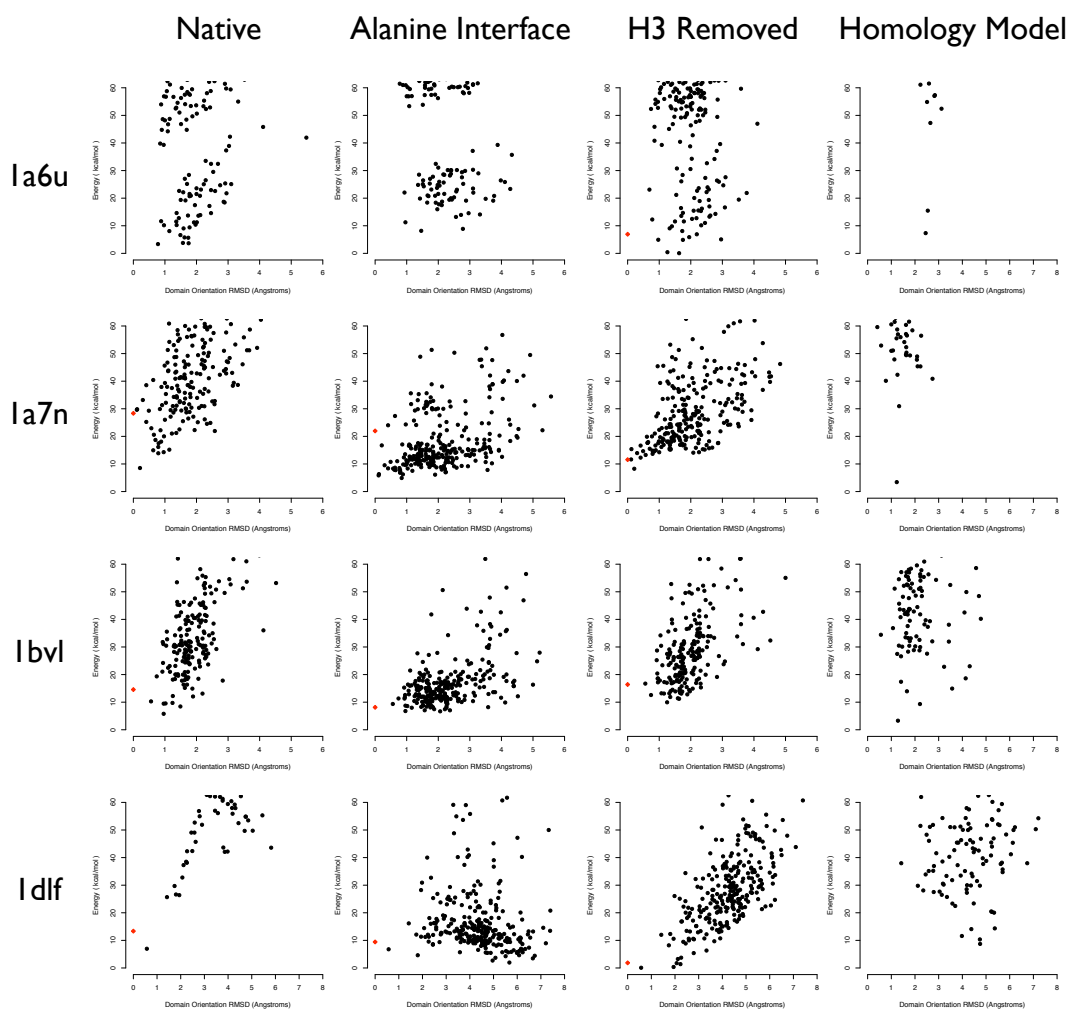
**Supplementary Table B4 Difference in orientation comparing unbound and antigen bound structures of identical antibodies**

PDB1	PDB2	RMSD	PDB1	PDB2	RMSD
12e8(L,H)	12e8(M,P)	3.1	1nj9(L,H)	1nj9(A,B)	0.5
1a7n(L,H)	1a7r(L,H)	0.1	1oaq(L,H)	1ocw(L,H)	1.7
1ruq(L,H)	1rur(L,H)	0.1	2q76(A,B)	2q76(C,D)	0.6
1a4j(L,H)	1a4j(A,B)	0.3	1q9o(A,B)	1q9o(C,D)	0.8
1a6t(A,B)	1a6t(C,D)	0.3	1qbl(L,H)	1qbm(L,H)	0.6
1ad0(A,B)	1ad0(C,D)	0.2	1qlr(A,B)	1qlr(C,D)	0.2
1aif(L,H)	1aif(A,B)	0.2	1rhh(A,B)	1rhh(C,D)	0.6
1bfo(C,D)	1bfo(G,H)	0.3	1rz8(A,B)	1rz8(C,D)	0.7
1bln(A,B)	1bln(C,D)	0.4	1rzi(E,F)	1rzi(I,J)	0.9
1bm3(L,H)	1opg(L,H)	0.5	1t4k(A,B)	1t4k(C,D)	0.3
1bvl(B,A)	1bvl(D,C)	1.0	1ub6(L,H)	1ub6(B,A)	0.9
1c5d(L,H)	1c5d(A,B)	0.4	1uwe(U,V)	1uwe(X,Y)	0.6
1d5b(A,B)	1d5b(L,H)	1.0	1uyw(L,H)	1uyw(N,M)	0.4
1dlf(L,H)	2dlf(L,H)	0.4	1wt5(C,A)	1wt5(D,B)	1.2
1dn0(A,B)	1dn0(C,D)	0.4	3cfj(E,F)	3cfk(G,I)	1.2
1dqq(A,B)	1dqm(L,H)	0.6	1yy8(A,B)	1yy8(C,D)	0.1
1f8t(L,H)	1s5i(L,H)	0.8	1za6(E,F)	1za6(G,H)	0.9
1fai(L,H)	2f19(L,H)	0.2	2aj3(A,B)	2aj3(C,D)	0.5
1ggb(L,H)	1ggc(L,H)	0.3	2g75(B,A)	2g75(D,C)	0.0
1gpo(L,H)	1gpo(M,I)	1.0	2fjf(C,D)	2fjf(S,T)	1.1
1hil(A,B)	1hil(C,D)	0.4	2gjz(A,B)	2gk0(L,H)	0.9
1hzh(L,H)	1hzh(M,K)	0.2	2gsg(A,B)	2gsg(C,D)	1.1
1xf3(L,H)	1xf4(A,B)	1.7	2iqa(L,H)	2iqa(A,B)	0.9
1igf(L,H)	1igf(M,J)	0.1	2orb(L,H)	2orb(M,I)	1.0
1igt(A,B)	1igt(C,D)	0.5	2z91(B,A)	2z91(D,C)	0.1
1j05(L,H)	1j05(A,B)	0.8	3cfe(L,H)	3cfe(A,B)	0.1
1mf2(L,H)	1mf2(M,N)	0.3	3cle(L,H)	3clf(L,H)	0.4
1n4x(L,H)	1n4x(M,I)	1.4	43c9(C,D)	43c9(G,H)	0.6
1ruk(L,H)	1rum(L,H)	0.1	8fab(A,B)	8fab(C,D)	1.3

**Supplementary Table B5 Difference in orientation comparing unbound structures of antibodies with identical interface sequence**

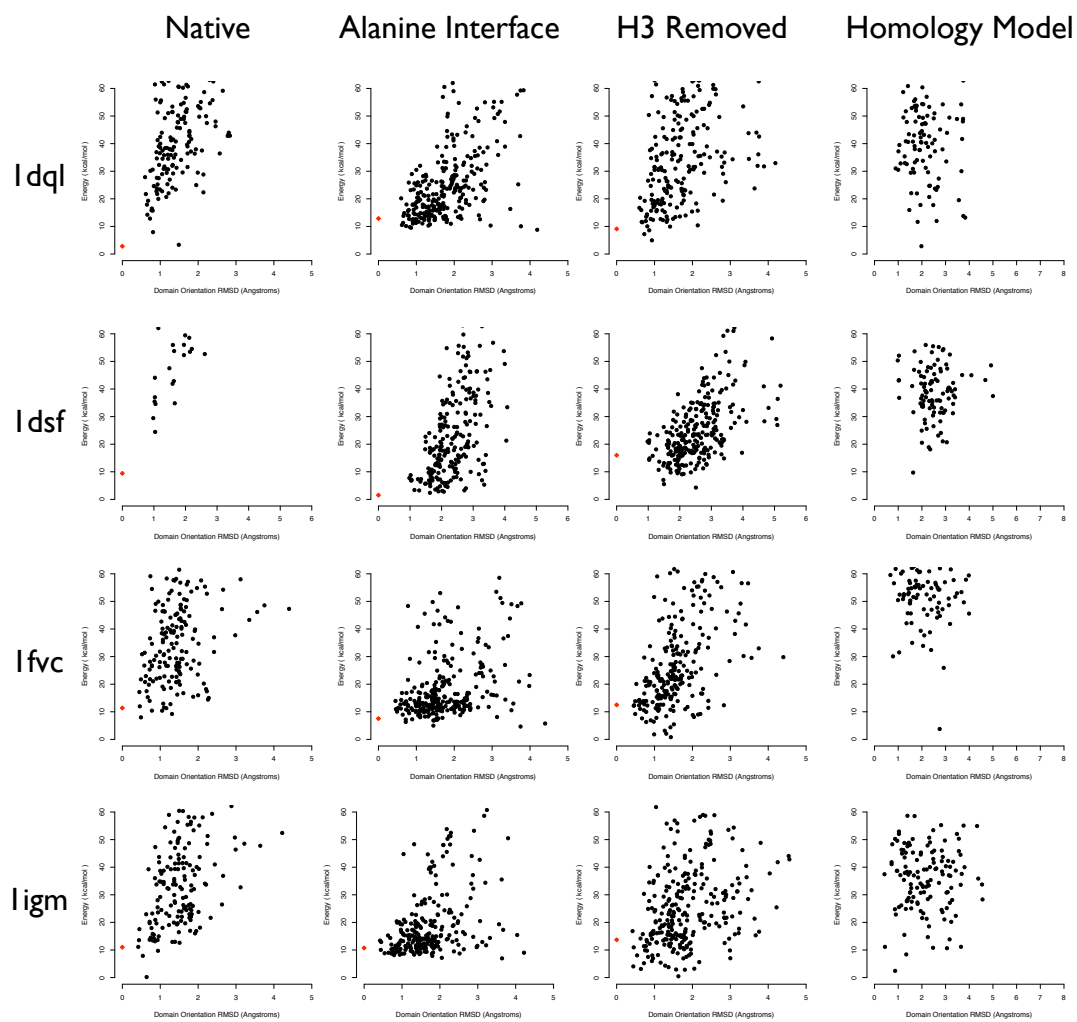
PDB1	PDB2	RMSD	PDB1	PDB2	RMSD
15c8(L,H)	25c8(L,H)	1.0	1l7t(L,H)	1vpo(L,H)	0.9
1a2y(A,B)	1a7n(L,H)	1.2	1m71(A,B)	1pz5(A,B)	1.0
1a3l(L,H)	1rur(L,H)	0.8	1mf2(L,H)	2hrp(L,H)	1.6
1a3r(L,H)	1bbd(L,H)	0.5	1mhh(C,D)	1nlb(L,H)	0.7
1aj7(L,H)	2res(L,H)	1.4	1mju(L,H)	1mjj(A,B)	0.7
2ap2(A,B)	1ap2(A,B)	1.5	1mnu(L,H)	2mpa(L,H)	1.0
1ar1(D,C)	1mqk(L,H)	0.9	1nd0(G,H)	1ruk(L,H)	0.7
1ay1(L,H)	1bgx(L,H)	1.0	1ngx(L,H)	1ngz(A,B)	2.6
1cfq(A,B)	1cfs(A,B)	1.7	2bjm(L,H)	1oaq(L,H)	3.6
1bvk(A,B)	1bvl(B,A)	1.3	1om3(L,H)	1zlu(L,H)	0.9
1c5b(L,H)	1c5c(L,H)	0.4	1p2c(A,B)	2q76(A,B)	1.6
1cbv(L,H)	1nbv(L,H)	2.0	1q0x(L,H)	1q0y(L,H)	1.0
1cgs(L,H)	1ynl(L,H)	1.0	1rfd(L,H)	1riv(L,H)	0.3
1ck0(L,H)	2ck0(L,H)	0.2	1q9l(C,D)	1q9q(A,B)	1.1
1ucb(L,H)	1clz(L,H)	0.6	1qbm(L,H)	1wej(L,H)	0.9
1cr9(L,H)	1cu4(L,H)	1.1	1yyl(L,H)	1rz8(A,B)	1.2
1d5i(L,H)	1d6v(L,H)	0.4	1um6(L,H)	1um4(L,H)	1.1
1dba(L,H)	1dbj(L,H)	0.4	1y0l(E,F)	3cfj(E,F)	1.3
1dqj(A,B)	1dqq(A,B)	1.1	1yy8(A,B)	1yy9(C,D)	1.1
1f4w(L,H)	1f4y(L,H)	0.6	2adg(A,B)	2adi(A,B)	0.8
1ggc(L,H)	1ggi(L,H)	2.7	2ajz(L,H)	2aju(L,H)	1.0
1hil(A,B)	1him(H,L)	1.3	2dd8(L,H)	2g75(B,A)	0.8
1hzh(M,K)	1n0x(M,K)	0.4	2dqt(L,H)	2dtm(L,H)	1.0
1xf4(A,B)	2fr4(A,B)	2.8	2eh8(L,H)	2eh7(L,H)	1.0
1igf(L,H)	2igf(L,H)	0.7	2fat(L,H)	2fd6(L,H)	1.6
1igi(L,H)	1igj(C,D)	0.7	2fjf(S,T)	2fjg(A,B)	2.3
1jgu(L,H)	1jgv(L,H)	0.4	2iqa(A,B)	2r0w(L,H)	1.2
1jn6(A,B)	1jnh(E,F)	0.3	2o5x(L,H)	2o5z(L,H)	0.5
1jpt(L,H)	1pg7(M,I)	0.9	2orb(L,H)	2or9(M,I)	1.6
1kc5(L,H)	1kcu(L,H)	0.2	2z92(B,A)	2z91(D,C)	0.3
1kcs(L,H)	1kcv(L,H)	1.0	3cfb(L,H)	3cfc(L,H)	1.4
1kel(L,H)	1kem(L,H)	0.2	3cfd(L,H)	3cfe(L,H)	0.4
1yeh(L,H)	1yek(L,H)	0.2	43c9(G,H)	43ca(C,D)	0.6

**Supplementary Table B6 Difference in orientation comparing unbound and antigen bound structures of antibodies with identical interface sequence**



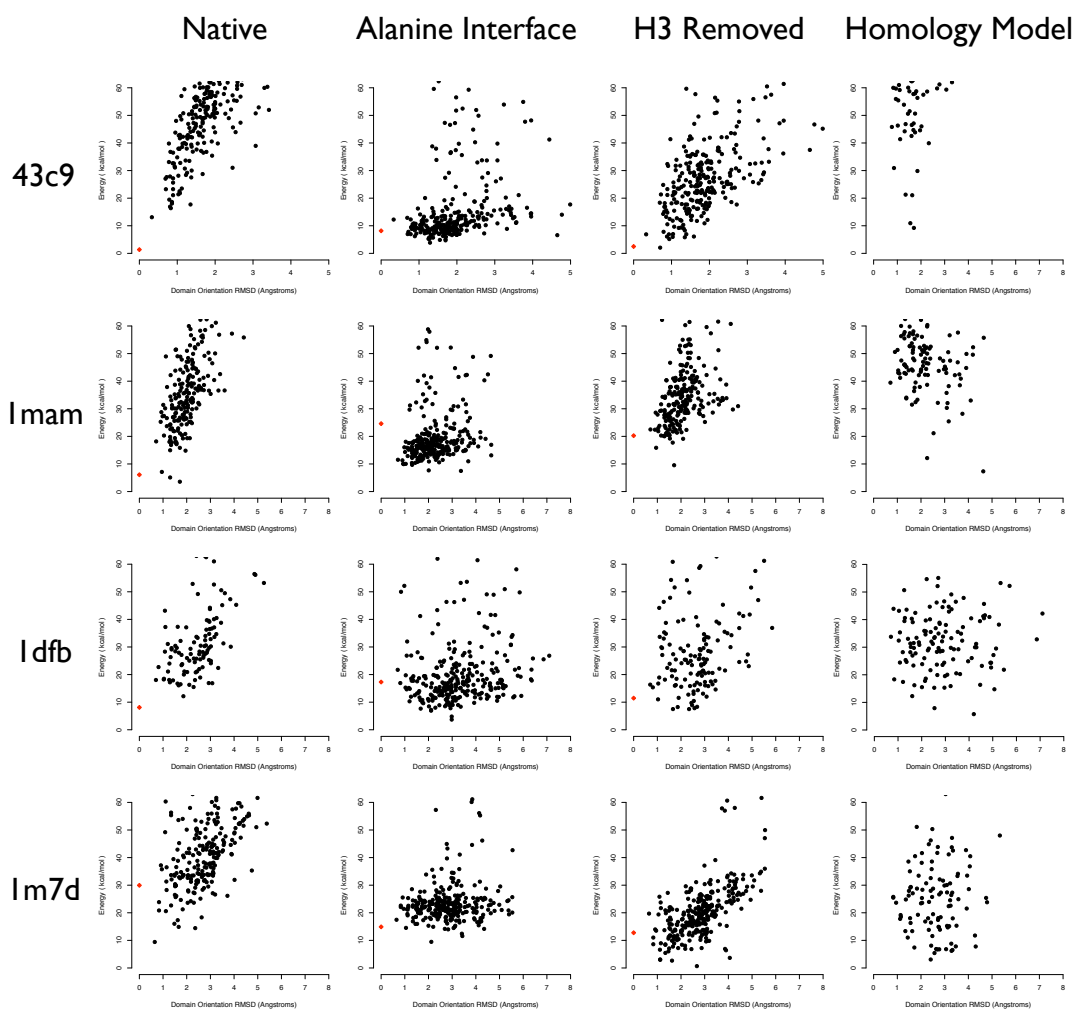
**Supplementary Figure B1 Predicted energy vs. difference in domain orientation relative to the crystal structure for each of the twenty test set antibodies**

Predictions were performed using native backbone conformations (Native), with interface residues mutated to alanine prior to prediction (Alanine Interface), with H3 CDR loop removed prior to prediction (H3 Removed), and comparative models of each domain (with the H3 CDR loop removed). In all cases involving the crystallographic domains, the energy calculated was the total energy of the variable region while for the predictions involving the modeled domains, the energy being plotted is the “interface energy” between the heavy and light chain variable domains. Figure continued on the next 4 pages.

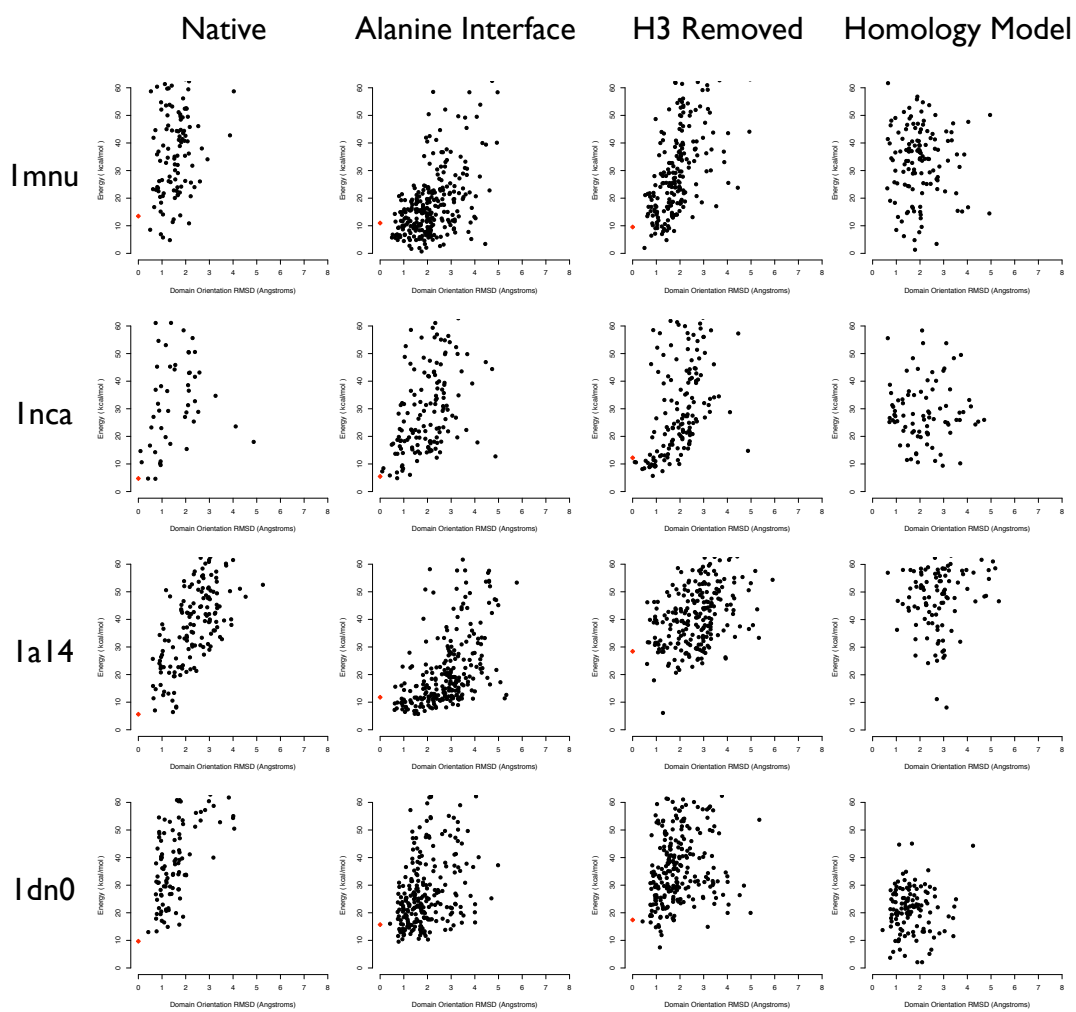


Supplementary Figure B1 continued

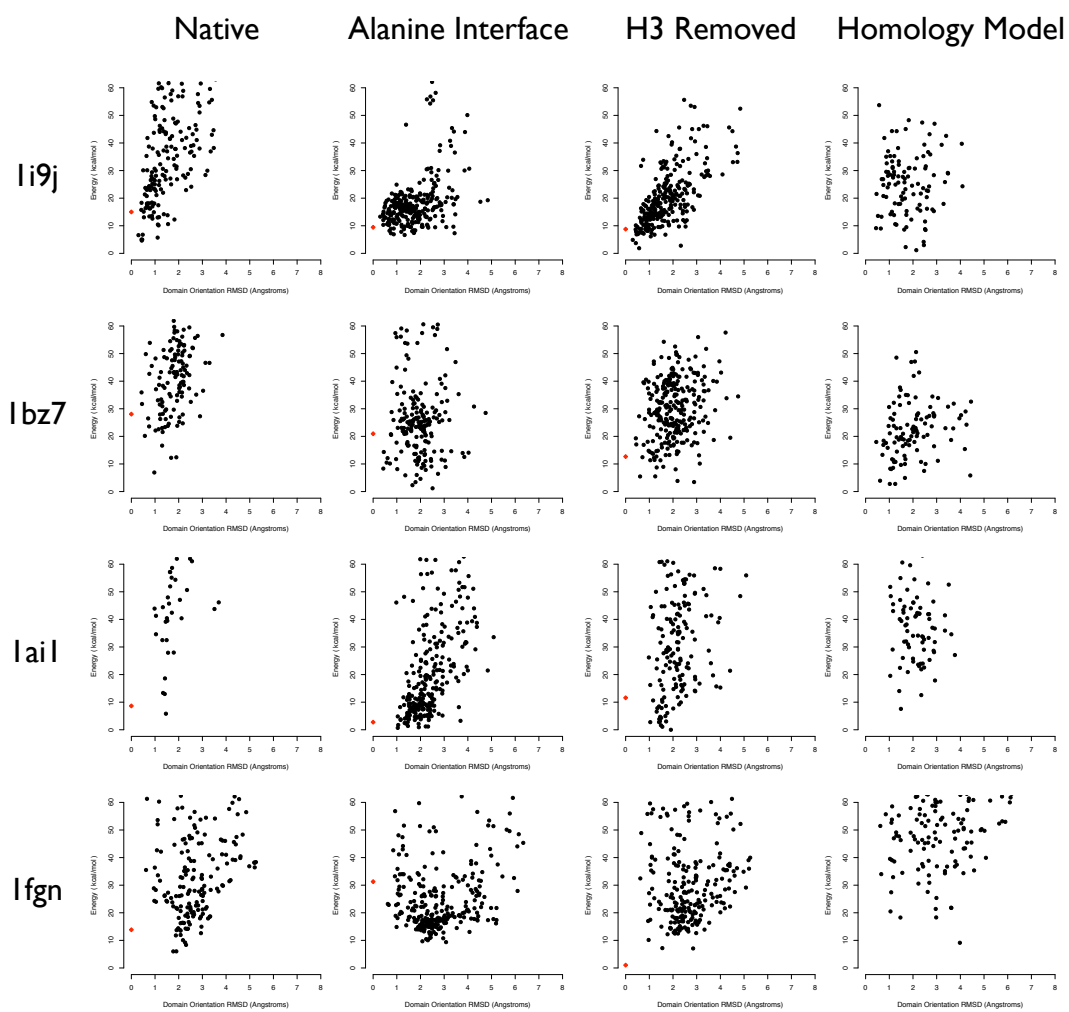




Supplementary Figure B1 continued



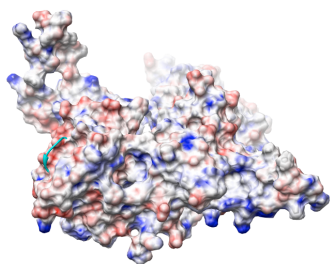
Supplementary Figure B1 continued



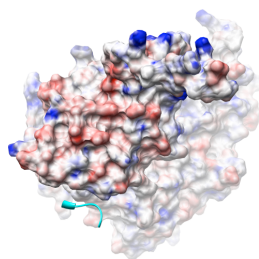
Supplementary Figure B1 continued

# Appendix C. Supplementary Figures for Chapter 4

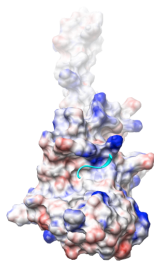
Arp2 binding site on Arp3



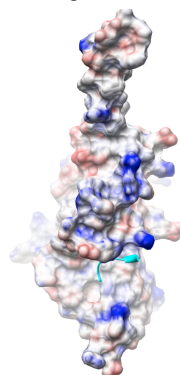
ARPC4 binding site on arp3



Arp2 binding site on ARPC4



Arp3 binding site on ARPC4



## Supplementary Figure C1 Charged nature of Arp3 and ARPC4 interaction surfaces


Surface maps of Arp3 and ARPC4 from the starting crystal structure (PDB 1K8K) colored by electrostatic potential calculated by Delphi [16]. The Arp3 surface interacting with Arp2 (A) and ARPC4 (B), and the ARPC4 surface interacting with Arp2 (C) and Arp3 (D) are shown. Arp2 residues Thr237 and Thr238 are shown in ribbon form in cyan. Delphi calculations were performed on the Arp2, Arp3, and ARPC4 subunits from the starting model for our simulations based on the unphosphorylated, apo crystal structure. Calculations were performed with a 2.0 grid unit / Å scale and with each subunit occupying 40% of the total volume. The internal dielectric used was 4.0, and the external dielectric was set to 80.0, with a probe radius of 1.4 Å used to delineate the molecular surface.

**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

  
\_\_\_\_\_  
Author Signature

10/30/2009  
\_\_\_\_\_  
Date