# Understanding Stability of Protein-Protein Complexes

Rudi Agius

September 2014

Biomolecular Modelling Laboratory,
Cancer Research UK London Research Institute
and
Faculty of Life Sciences University College London

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Life Sciences at the University College London.

I, Rudi Agius, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

# Abstract

For all living organisms, macromolecular interactions facilitate most of their natural functions. Alterations to macromolecular structures through mutations, can affect the stability of their interactions, which may lead to unfavourable phenotypes and disease. Presented here, are a number of computational methods aimed at uncovering the principles behind complex stability - as described by binding affinity and dissociation rate constants. Several factors are known to govern the stability of protein-protein interactions, however, no one factor dominates, and it is the synergistic effect of a number of contributions, which amount to the affinity, and stability of a complex. The characterization of complex stability can thus be presented as a two-fold problem; modelling the individual factors and modelling the synergistic effect of the combination of such individual factors. Using machine learning as a central framework, empirical functions are designed for estimating affinity, dissociation rates and the effects of mutations on these properties. The performance of all models is in turn benchmarked on experimental data available from the literature and carefully curated datasets. Firstly, a *wild-type* binding free energy prediction model is designed, composed of a diverse set of stability descriptors, which account for flexibility and conformational changes undergone by the complex in question. Similarly, models for estimating the effects of mutations on binding affinity are also designed and benchmarked in a community-wide blind trial. Emphasis here is on the detection of a small subset of mutations that are able to enhance the stability of two *de novo* protein drugs targeting the flu virus hemagglutinin. Probing further the determinants of stability, a set of descriptors that link hotspot residues with the off-rate of a complex are designed, and applied to models predicting changes in off-rate upon mutation. Finally, the relationship between the distribution of hotspots at protein interfaces, and the rate of dissociation of such interfaces, is investigated.

# Acknowledgements

I would like to begin by thanking my family for encouraging me in all my endeavours and to follow my aspirations. For this I am forever grateful.

While it is not possible for me to mention all the names, I would like to extend my thanks to all the teachers who have been part of my academic journey and inspired me to follow this path.

This PhD would not have been possible, not only without the funding, but also without the constant support provided by Cancer Research UK. I express my appreciation for all staff and supporters of the charity. In particular, I would like to thank Dr. Sally Leevers, Sabina Ebbols and Emma Rainbow. Furthermore, I would also like to thank my thesis committee, Dr. Caroline Hill and Dr. Martin Singleton for their objective advice and care they've shown during this PhD.

A special thanks to all members of the Biomolecular modelling lab. Firstly, Dr. Raphael Chaleil, the scientific officer, for never losing his patience, on even the most mundane of questions. Dr. Iain Moal, whom his ideas and discussions have been an invaluable source of knowledge, and for the productive collaborations we've had. Dr. Mieczyslaw Torchala for his instrumental role in the projects we collaborated on in the latter years of my PhD. Dr. Tammy Cheng, Dr. Melda Tozluoglu, Sakshi Gulati and Erick Pfeiffenberger who have made these past years scientifically stimulating and were supportive through all the ups and downs of these years.

Last but not least, I owe my deep felt gratitude to Dr. Paul Bates for a number of things. To begin with, for believing in my potential as a researcher, for introducing me to the world of protein interactions and for guiding me, but also giving me the space to pursue my own ideas. He has made these four years a pleasant experience, be it through his thought provoking ideas, our sometimes endless discussions, and his understanding in my less than stellar moments.

Finally I would like to thank my examiners, Prof. David T. Jones and Prof. Michael J. Sternberg for accepting to review my thesis and allowing me to defend it *viva voce*.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ACE | Analytic Continuum Electrostatics |
| ANN | Artificial Neural Network |
| BAP | Binding Affinity Prediction |
| GA | Genetic Algorithm |
| GA-FS | Genetic Algorithm Feature Selection |
| HS | Hotspot |
| $k_{off}$ | Dissociation Rate |
| $k_{on}$ | Association Rate |
| LR | Linear Regression |
| MARS | Multivariate Adaptive Regression Splines |
| MCC | Matthew's Correlation Coefficient |
| MD | Molecular Dynamics |
| MHC | Major |
| ML | Machine Learning |
| PCC | Pearson's Correlation Coefficient |
| PPI | Protein-Protein Interaction |
| RBF | Radial Basis Function Interpolation |
| RF | Random Forest |
| SNPs | Single Nucleotides polymorphisms |
| SVM | Support Vector Machine |
| ΔG | Binding Free Energy |
| $\Delta k_{off}$ | Change in Dissociation Rate upon Mutation |
| ΔΔG | Change in Binding Free Energy upon Mutation |

# Peer Reviewed Publications

List of publications published in peer-reviewed journals during my time as a PhD student:

Moal, I.H*., Agius, R*. & Bates, P.A. (2011). Protein-protein binding affinity prediction on a diverse set of structures. **Bioinformatics,** 27, 3002-3009.

Cheng, T.M.K*, Gulati, S*., Agius, R*. & Bates, P.A. (2012). Understanding Cancer Mechanisms through Network Dynamics. **Brief. Funct. Genomics** 11(6), 543-560.

Moretti, R., Fleishman, S.J., Agius, R., Torchala, M., Bates, P.A., Kastritis, P.L., Rodrigues, J.P.G.L.M., Trellet, M., Bonvin, A.M.J.J., Cui, M., Rooman, M., Gillis, D., Dehouck, Y., Moal, I., Fernandez-Recio, J., Flores, S., Pacella, M., Kilambi, K.P., Gray, J.J., Popov, P., Grudinin, S., Esquivel-Rodriguez, J., Kihara, D., Zhao, N., Korkin, D., Zhu, X., Demerdash, O.N.A., Mitchell, J.C., Kanamori, E., Tsuchiya, Y., Nakamura, H., Lee, H., Park, H., Seok, C., Sarmiento, J., Liang, S., Teraguchi, S., Standley, D.M., Shimoyama, H., Terashi, G., Takeda-Shitaka, M., Iwadate, M., Umeyama, H., Beglov, D., Hall, D.R., Kozakov, D., Vajda, S., Pierce, B.G., Hwang, H., Vreven, T., Weng, Z., Huang, Y., Li, H., Yang, X., Ji, X., Liu, S., Xiao, Y., Zacharias, M., Qin, S., Zhou, H.X., Huang, S.Y., Zou, X., Velankar, S., Janin, J., Wodak, S.J. & Baker, D. (2013). Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. **Proteins** 81(11), 1980-1987.

Torchala, M., Moal, I.H., Chaleil, R.A.G., Agius, R. & Bates, P.A. (2013). A Markov-chain model description of binding funnels to enhance the ranking of docked solutions. **Proteins** 81(12), 2143-2149.

Agius, R., Torchala, M., Moal, I.H., Fernandez-Recio, J. & Bates, P.A. (2013). Characterizing changes in the rate of protein-protein dissociation upon interface mutation using hotspot energy and organization. **PLoS Comput. Biol.** 9(9): e1003216

* These authors contributed equally to this work.

# Chapter 1

# 1 Introduction

These days more than ever, we live in a world of networks. At its roots, a network is defined as a set of 'nodes' and a related set of 'links'. A link between two nodes indicates some connection or relationship between these nodes. More interestingly, a link between two nodes may indicate a transfer of information. Be it a transfer of information as a result of a simple conversation between two friends on a social network, a flip of polarity at the output of a logic gate in an electronic circuit network, or, and what concerns this thesis mostly, the binding of two molecules in a biological network. Such binding events are at the core of all cellular processes, and networks of molecular interactions enable each cell to sense its external environment, propagate the necessary information inwards, and make decisions concerning its cellular state or even the states of its neighbouring cells. With this, it then becomes clear that, not only do we live in a world of networks, but our health too is the result of numerous intercommunicating biological networks.

This thesis is concerned with the link between two nodes, that very interaction between two molecules; in this case that between two proteins. The emphasis is placed on understanding what constitutes a stable interaction between them.

The stability of such interactions plays an important role in both our understanding of disease and that of designing better drugs; these aspects are detailed below in the thesis justification section, 1.4. First, as an introduction to this work, the signalling behaviour of T cells will be described. This case study is a prime example of both the centrality of protein-protein interactions, and how the change in stability of one of these interactions can affect the activity and response of the T cell itself.

## 1.1 Information Processing in the Cell: A Key Example, T Cell Receptor Signalling

The importance of understanding the factors controlling the binding affinities of proteins within a complex cellular information processing system can be well exemplified by the T cell receptor-signalling network. T cells (T lymphocytes) are a subset of white blood cells which form an integral part of our immune system fighting against virus infected or malignant cells. These include, T-Helper Cells, T-Suppressor and T-Killer Cells (cytotoxic T cells). Effectively, their function is to elicit a distinct and specific response depending on the foreign antigen detected. T cells work by a cascade of signalling events initiated from the T-cell receptor (TCR). The TCR recognizes peptides presented by Major Histocompatibility Complex (MHC) molecules from antigen presenting cells (APC). The peptides themselves are usually cleaved parts of cellular proteins. If the cell is infected with a virus, then some of these peptides will be from foreign proteins (See Figure 1.1a). The ability for T cells to make this distinction is therefore critical and defects in the normal T cell response lead to several autoimmune (Dejaco et al., 2006) or immunodeficiency related diseases (Edgar, 2008), some of which may have severe health consequences. Besides the binding of the TCR and pep-MHC, (MHC with antigenic peptide) simultaneous binding of specific co-receptors, CD4 on T helper cells, and CD8 on cytotoxic T cells, with the MHC molecule initiates a myriad of signalling events. Some of these interactions are depicted in Figure 1.1b.

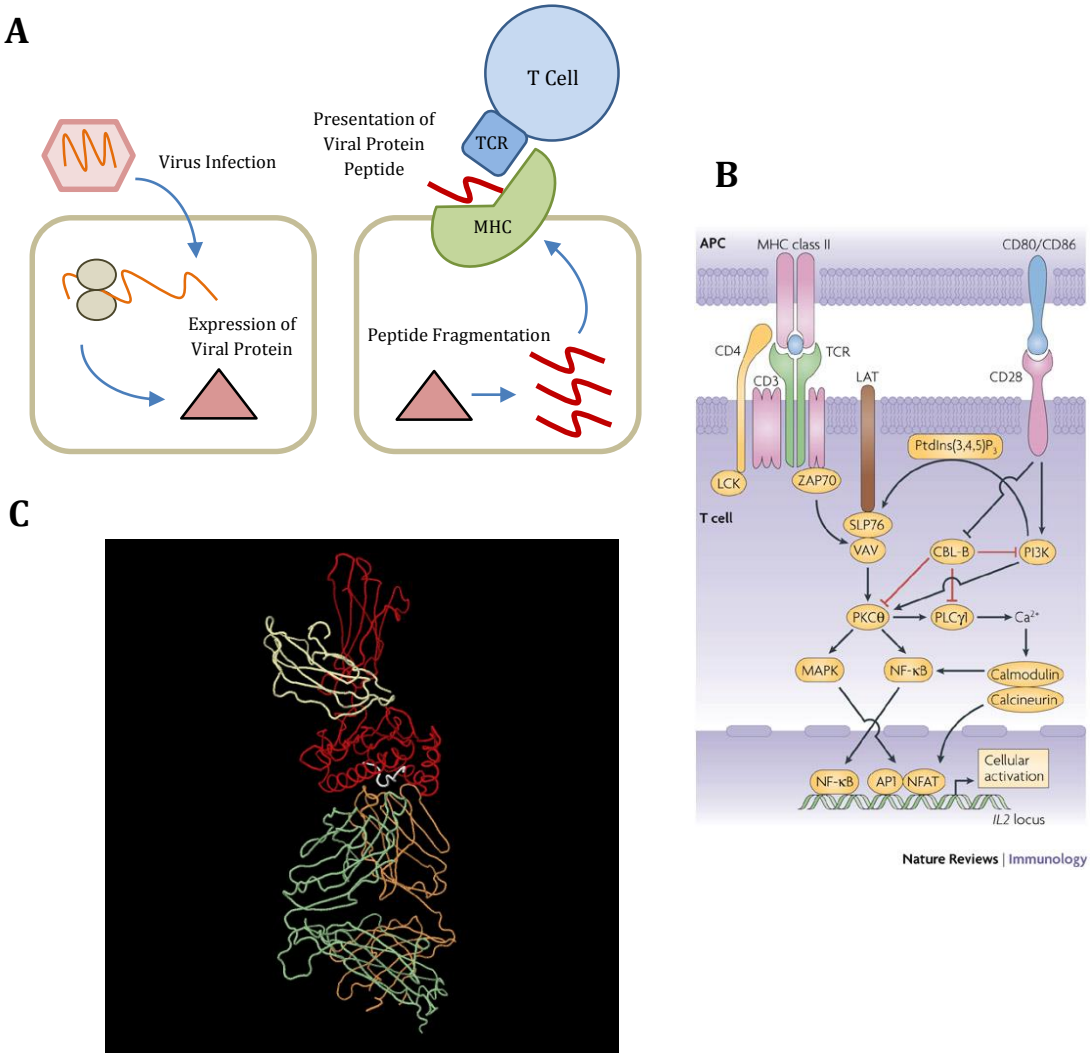



**Figure 1.1: T cell Receptor Signalling.**
(A) Pictorial depiction of viral infection, viral protein expression, peptide fragmentation and the presentation of the viral protein peptide on the Major Histocompatibility Complex (MHC). A neighbouring T Cell detects the foreign peptide using its T cell receptor (TCR). (B) Some of the interactions and signalling triggered by the formation of the TCR/pepMHC complex. Figure taken from Miller et al. (2007). (C) the structure the complex between human TCR b7, viral peptide (TAX) and MHC Class I molecular HLA-A 0201. (PDBid: 1BD2). This structure includes the extracellular portions of a T-cell receptor and class I MHC. TCR chains are in red and yellow. MHC chains are in green and orange. Peptide is shown in white. Of much debate is how the affinity and kinetic of this interaction affects T cell activity.

Upon TCR/pep-MHC binding, LCK (an Src family kinase) is recruited and phosphorylates the immune-receptor tyrosine-based activation motifs (ITAMS) which form part of the intracellular subunits of the TCR itself(Lin and Weiss, 2001). After phosphorylation of ITAMS, ZAP-70 is activated which binds to two adapter molecules LAT and SLP-76 and their subsequent phosphorylation of LAT and SLP-76 triggers the Ras pathway (Lin and Weiss, 2001). The signal continues further downstream until several transcription factors are activated. This in turn elicits a number of responses related to T cell activation which include cytokine release, proliferation and apoptosis amongst others.

*Sensitivity of T Cell response signalling to TCR/pep-MHC affinity and kinetics:*

The centrality of the interaction between TCR and pep-MHC (see Figure 1.1c), has led to many different models of T cell activation. Initial models propose the TCR as simple on-off switch where TCR/pep-MHC binding elicits a full T cell activation (Jameson, 1998). Experiments presenting different pep-MHC molecules however show that different TCR ligands trigger none or only some of the T cell activation responses (Kersh and Allen, 1996). These have been termed as TCR antagonists and partial agonists respectively. The fact that some but not all T cell activity responses may be activated led to development of the 'kinetic proofreading' model. In this model, the affinity (or off-rates) of the pep-MHC molecule with the TCR is proportional to the magnitude of the T cell response (!!! INVALID CITATION !!!). For low residence times (fast off-rates), early activation events, without the presence of late T cell activation events, are elicited. Slower off-rates on the other hand, enable a full T cell activation response. Evidence not supporting this model, such as the activation of late T cell signals with fast off-rates (Rosette et al., 2001), and the discovery that a small number of peptide-MHC can serially engage and trigger up to approximately 200 TCRs, instigated an alternative 'serial triggering ' hypothesis (Valitutti et al., 1995). In this case, the interaction's off-rate must be sufficiently low for initial signalling to be completed, but high enough to allow different TCRs to bind the same pep-MHC molecule. This suggest that there is an 'optimal dwell time' which elicits T cell activation and anything outside this optimal range results in reduced activity. A model of consensus is however still hindered by several challenges (Stone et al.,

2009). For example; outlier observations have been made that contradict both hypothesis; experimental binding measurements are generally made at lower temperatures than those of *in vivo* activity; and the effects of co-receptors CD4 and CD8 should complicate the story even further (Stone et al., 2009).

The overview of TCR signalling presented above is a crude one at most, and can only be refined once our theoretical knowledge of just how binding affinities are controlled at the atomic level improves. Moreover, there are a vast amount of molecular interactions and interplays between multiple pathways (Huse, 2009). Therefore, this example serves as a reminder of the complexity of protein interactions in cellular networks, and how the response of such a system may be affected by the stability of just one of those interactions. The information processing mechanisms of the T cell receptor network, as with many other signalling networks, can only be truly appreciated and understood when considering the dynamics and stability of its molecular interactions.

## 1.2 Thesis Outline

In this thesis, a number of computational investigations are performed aimed at understanding the stability of protein-protein complexes. The investigations revolve around the design of a number of predictive models that correlate with experimental measurements for stability. Therefore, in the following section, 1.3, a brief overview is given of the different terms that relate to complex stability, and those that form part of this study; these include binding affinities, dissociation rates and hotspots. In section 0, justification for this thesis is further underlined by showing that the study of the stability of protein-protein interactions (PPIs), has a direct impact on recent trends in drug design. This includes the growing interest in PPIs as drug targets (section 1.4.1), protein engineering and protein drugs (section 1.4.2) and the importance of considering off-rates for the enhancement of *in vivo* drug activity. In section 1.4.4 it will also be described how the functional interpretation of missense SNPs is dependent on our ability to characterise the changes in stability resulting from these mutations. In section 1.5 the equations governing the kinetics and thermodynamics of binding are presented the energetic terms used for

modelling of binding free energies detailed. In section 1.6.7 an overview of the current models for the prediction of binding affinities is given and their limitations highlighted. A number of machine learning algorithms are employed throughout this thesis of which an overview is given in section 1.6.

Finally, in section 1.7, I present some of my personal motivations and themes that drive the studies presented in this thesis. This chapter then concludes with an overview of each of the remaining thesis chapters. In summary, models for the prediction of binding affinities and their change upon mutation are presented in Chapters 3 and 4. In Chapter 5, models for the prediction of hotspots are presented and benchmarked. In Chapter 6, a set of descriptors that link hotspot residues with the off-rate of a complex are designed. Chapter 7 extends on this idea by building prediction models for off-rate changes upon mutation. Finally in Chapter 8, the relationship between the distribution of hotspots on an interface, and the rate of dissociation, is investigated.

## 1.3 Facets of Complex Stability in a Nutshell: Binding Affinities, Off-Rates and Hotspots

The pathways shown in diagrams similar to Figure 1.1b, provide a very sparse and static picture of the nature of the environment of protein interactions. In reality, proteins exist in a highly dense 'soup-like' environment in the cell (Lewitzky et al., 2012). For example, the intracellular concentration of proteins for mammalian cells is estimated at 200-300mg/ml (Luby-Phelps, 2000) and macromolecules themselves occupy 40% of the total cell volume (Fulton, 1982). For an interaction to take place, proteins must therefore rummage through this crowded environment, i) find their partner, ii) find the binding site and iii) form a complex for an indefinite amount of time. For those protein-protein interactions that are sufficiently long-lived, the strength of the interaction can be determined by the binding affinity. This means that for a pair of proteins, being able to predict their binding affinity, should in theory determine whether two such proteins make a biologically significant interaction. In kinetic terms (as derived in section 1.5.1), the binding affinity of an interaction

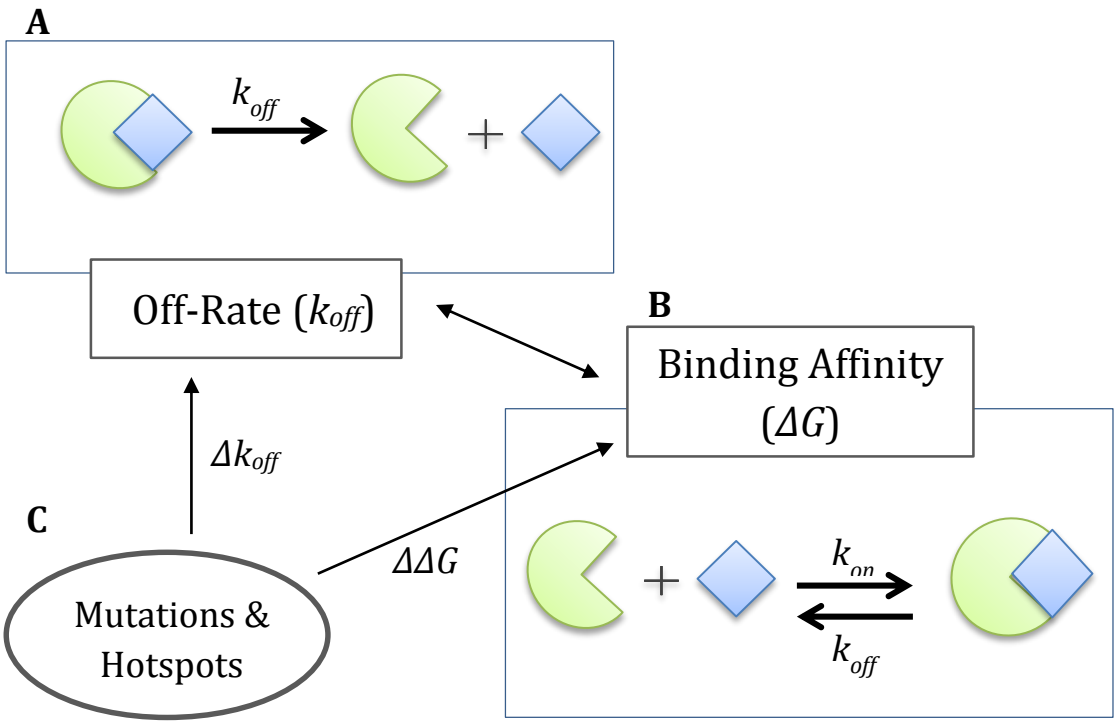



**Figure 1.2: The relationship between the different facets at which complex stability may be characterised and those which are studied in this thesis.** (A) The off-rate ($k_{off}$) represents the intrinsic disposition of a complex to dissociate once it has formed. The higher the time for which the complex is bound, the lower the off-rate. (B) Adding to this, knowledge of the intrinsic disposition for the complex to associate ($k_{on}$), the binding affinity ($\Delta G$) may also be calculated. (C) Characterisation of the effects of mutations on both the off-rate, $k_{off}$, and on the binding affinity, $\Delta\Delta G$, is central to the functional interpretation of disease and for computational drug design. Alanine scanning experiments have shown that only a few mutations cause significant disruption to complex stability. These are known as hotspots and are the residues responsible for most of the binding affinity of a protein-protein interaction.

is related to how easy it is for the two partners to reach the bound state ($k_{on}$), and how easy it is for the two partners to unbind back into separate protein conformations ($k_{off}$). Prediction of the $k_{off}$ of a complex effectively determines the length of time (residence time = $1/k_{off}$) for which the complex is bound. From alanine scanning experiments on protein-protein interfaces, only a small subset of interface 'hotspot' residues is found to be responsible for the binding affinity

of the complex (Bogan and Thorn, 1998, Clackson et al., 1998). These hotspot residues may in turn affect binding through a change in $k_{off}$ independently of $k_{on}$ and vice versa (Moal and Fernandez-Recio, 2012). In this view, complex stability, as in fact explored in this thesis, may be approached at different yet related levels (See Figure 1.2).

## 1.4 A Thesis Justified

In this section it will be described how the characterisation of protein-protein binding affinities and off-rates has direct relevance to the current trends and difficulties of drug-design. In a similar vein, the functional interpretation of mutations involved in disease necessitates that we are able to accurately predict changes in affinities upon mutation.

### 1.4.1 Protein-Protein Interactions as Drug Targets

Despite their therapeutic relevance and major involvement in cellular signalling, PPIs have traditionally received less attention as drug targets, or attempts to target them have shown few success stories. For example, Bcl-2 family proteins are key regulators of programmed cell death and Bcl-$X_L$ and Bcl-2 are overexpressed in many cancers. Bcl-$X_L$ expression is correlated with chemo-resistance and reduction in Bcl-2 expression increases sensitivity to anticancer drugs and *in vivo* survival. Several drugs targeting these proteins have been explored but resultant affinities have not been found to be sufficiently high (Oltersdorf et al., 2005). The main difficulty in achieving high-affinity binding is that the structural properties of PPIs do not have common drug-like site properties. The large surface area of the PPI binding site is typically much larger than that covered by the small-molecule drug. In addition, PPIs have characteristically flat interfaces and no well-defined binding pockets; this limits the contact surface area the small-molecule drug can make with the protein (Mullard, 2012, Jin et al., 2014). The 'undruggable' view started to change in the 1990s after studies on protein-protein interactions identified certain hotspot residues responsible for most of the binding free energy (Bogan and Thorn,

1998, Clackson et al., 1998). This shows that, even though large in surface area, binding energy is not distributed homogenously across the interface and it is therefore potentially sufficient to design drugs which target only these hotspot residues (Hajduk et al., 2005). The recent interest in inhibiting PPIs is reflected by several pharmaceutical firms which are now in the process of extending drug discovery programs aimed at identifying PPI inhibitors and expanding their libraries to account for this class of targets (Mullard, 2012). In the light of this new interest for PPI inhibitors, a number of companies have also moved past the preclinical stage. Lifitegrast (SAR1118), a small molecular inhibitor for treatment of dry eye is in phase III trials. It works by reducing T cell-mediated inflammation, blocking the PPI between ICAM-1 and LFA-1. Two anti-cancer agents blocking the PPI of p53 and MDM2 are under phase 1b trials (Vassilev et al., 2004, Mullard, 2012) and key PPIs inhibiting the function of the pro-survival BCL-2 family proteins are in phase-II development as anticancer agents (Mullard, 2012, Oltersdorf et al., 2005).

Two main challenges are therefore present in targeting PPIs with small-molecule drugs; knowing where to target on the protein interface, and doing so with high affinity. For competitive drug binding, the affinity of the protein-drug complex on its own gives no indication to its inhibitory effect. Rather, this protein-drug affinity becomes relevant only when higher than the affinity of *wild-type* protein-protein interaction i.e. that which it is competing against. Therefore, knowledge of the *wild-type* protein-protein binding affinities, as presented in Chapter 2, is a critical piece of information in competitive inhibitor design. As mentioned above, for small-molecule drugs targeting PPIs, only a small portion of the protein-protein interface can be targeted; therefore, knowing where at the interface to do so is imperative. Although hotspots are indeed good targets, unappreciated is the fact that hotspots can occur at disjointed parts of an interface or within clusters called hotregions (Keskin et al., 2005). Therefore, whereas the presence of hotspots greatly reduces the druggable search space of an interface, multiple potentially druggable sites are still present. In Chapter 8, an investigation is reported on which hotspot sites are contributing the most towards stability. Such

investigations should further guide the design of small-molecule drugs targeting PPIs.

### 1.4.2 Protein Engineering and Protein Drugs

In the previous section it was described how knowledge of the affinity and determinants of complex stability for protein-protein interactions, is important to the design of small-molecule drugs inhibiting PPIs. Here a more direct application of the methods developed in this thesis, that of protein engineering and protein drug design is presented. Protein engineering refers to the reengineering of proteins to enhance the affinity of existing interactions or develop new ones. In theory, the applications are numerous and include the rewiring of cellular networks by redesigning specificities; the design of proteins mimicking antigenic epitopes for potent vaccines and the design of protein probes for dissection of cellular protein networks and protein drug inhibitors (Mandell and Kortemme, 2009). Though applications are still exploratory in nature, proofs of concept have already started to surface. Recent work in the computational design of protein interactions includes the redesign of specificity at a protein-protein interface which was applied to model novel interacting DNase-inhibitor protein pairs (Kortemme et al., 2004); the use of positive (affinity increasing) and negative (affinity decreasing) design strategies to convert a homodimer into a heterodimer (Bolon et al., 2005); the redesign of a micromolar affinity human hyperplastic disc protein binding the kinase domain of PAK1 (Jha et al., 2010); the design of a high affinity interaction by grafting known key residues onto an unrelated protein scaffold (Liu et al., 2007) and more recently, (Fleishman et al., 2011) designed two proteins that bind a conserved surface patch on the stem of the influenza hemagglutinin (HA) from the 1918 H1N1 pandemic virus with low nanomolar affinity.

The methods mentioned above employ a variety of computational approaches, including conformational sampling mechanisms, docking algorithms and scoring functions. The latter function should be capable of identifying designs (generally through interface mutations), which increase the affinity of the desired interaction. In Chapter 4, the design of computational models capable of rank

ordering mutations on a protein-protein interface according to their change in affinity (ΔΔG) is reported upon. The models are benchmarked on two protein drugs where affinity-increasing mutations formed less than 5% of all the mutations to be tested.

### 1.4.3 Off-Rates in Drug design

Traditionally, early stage drug development is characterised by the optimization of the binding affinity or its other forms, IC50, or EC50 that calculate the drug concentration needed to achieve half-maximal inhibition. This is based on the assumption that binding affinity in closed *in vitro* systems is a good indicator of *in vivo* drug efficacy (Pan et al., 2013). *In vivo* systems, where the concentration of a drug-like ligand exposed to its target receptor is not constant, the drug efficacy is no longer well described by the *in vitro* measured dissociation constant. Rather, it depends on the association ($k_{on}$) and dissociation ($k_{off}$) rate constants (Copeland et al., 2006). The enhancement of the on-rate is limited in several ways, which highlights the reduction of the off-rate as the more favoured route. For example, the diffusion-rate remains an upper-bound restricting further optimisation of the on-rate. Modulating receptor desolvation and molecular orientation in a systematic way, is not trivial. Also, the rate of association depends not only on the $k_{on}$, but also on the concentration of ligand, which in turn is affected by multiple steps *in vivo*; as absorption, distribution and clearance all have an effect on ligand concentration (Copeland et al., 2006). Off-rate optimization on the other hand, is independent of such factors and entirely dependent on the short-range interactions between the bound monomers in question. Swinney (2004) hypothesizes that the most effective drugs utilize non-equilibrium transitions to enhance activity, and therefore methodologies that measure kinetics (most notably off-rates), non-equilibrium binding events and conformational diversity might have more potential than previously thought. Similar recent opinions can be found in (Holdgate and Gill, 2011), where surrogates of the off-rate, i.e. residence time ($1/k_{off}$) and kinetic efficiency are proposed as additional optimization targets to improve drug potency. A case in point is the management of Chronic Obstructive Pulmonary Disease (COPD). COPD encompasses a number of pulmonary diseases including

chronic bronchitis, emphysema and chronic obstructive airways disease. *Ipratropium bromide* (Baigelman and Chodosh, 1977) the drug commonly administered for the treatment of COPD has now been replaced by *Tiotropium bromide (Kato et al., 2006)* as the drug of choice. Both of the drugs have similar drug mechanism of action; namely by binding to the $M_3$ muscarinic receptor, leading to a reduction in smooth muscle contraction which in turn opens up the airways. Both drugs also have similar structures and pharmokinetic profiles; however, the duration of action of *Tiotropium* (24hrs) is four times that of *Ipratropium*, which can be administered daily. Studies (Disse et al., 1999) show that the difference in the duration of action between the drugs lies in their rates of dissociation from the $M_3$ muscarinic receptor. Namely *Tiotropium* has a residence time of 34.7 hours compared to 0.26 hours for *Ipratropium*.

In contrast to studies on binding affinities and on-rates, work on off-rates is still very limited (Moal and Bates, 2012). Up until this work, no models for the prediction of changes in off-rate upon mutation were reported. The release of the SKEMPI dataset (Moal and Fernandez-Recio, 2012) which contained a set of 713 off-rate mutations, enabled for the first time the modelling of off-rates on a diverse set of PPIs. In chapter 5-8 work is presented on the design of descriptors and models for characterising changes in off-rate upon mutation using SKEMPI.

### 1.4.4 Changes in Protein-Protein Stability and Disease

In the previous sections, it is argued that understanding and predicting the stability of protein-protein complexes is at the core of applications related to drug design. Presented in this section is, the other side of the spectrum, namely that predicting the change in stability of mutations on protein-protein interactions, is central to the understanding of disease mutations, such as those driving cancer.

Single Nucleotides polymorphisms (SNPs) are variations in the DNA sequence that have a direct effect on our susceptibility to disease and response to treatment. For those SNPs that occur in the coding regions, the SNPs can either be synonymous (not affecting protein amino-acid sequence), or non-synonymous

(affecting the protein amino-acid sequence). For the latter category, the SNPs can either by nonsense, where the protein amino-acid sequence is truncated, or missense where amino-acid substitutions take place. Nonsense non-synonymous (nsSNPs) generally result in a non-functional protein as a result of the truncation (Gregersen et al., 2000), missense nsSNPS are however more diverse and depending on where the variation occurs, effect on protein function can be anything from disease related to indiscernible (Haber and Settleman, 2007).

A major goal is therefore linking nsSNPs to phenotype through structure and function. For example, missense nsSNPs which translate to a mutation at the core of a protein generally destabilizes the protein-fold (Yue et al., 2005). Consequently all of the protein's interactions are lost. A study on nsSNPs on a number of protein-protein interactions show that disease causing nsSNPs not found at the core of a protein, tend to frequent the interface more than the non-interacting surface (David et al., 2012). Missense nsSNPs resulting in surface mutations may affect PPIs in a number of ways; they may destabilise existing interactions by disrupting favourable intermolecular contacts at the interface, affect post-translational-modifications, or even modulate the intrinsic disorder of the protein. In some cases it may also lead to the creation of new interactions consequently re-wiring the PPI network (Yates and Sternberg, 2013).

Being able to predict the consequence of a mutation at a protein-protein interface is therefore vital to uncovering the mechanism of action of disease causing nsSNPs. For example, depending on its sign and magnitude, the prediction of the ΔΔG may tell us whether the mutation has no effect on the given interaction, whether it leads to its loss or whether it helps stabilise a potential novel interaction. Models for the prediction of ΔΔGs are designed and presented in Chapter 4.

## 1.5 Modelling the Binding Free Energy

### 1.5.1 The Kinetics of Binding

The derivation of the binding free energy of an interaction may be approached from two perspectives; from a kinetic and from a thermodynamic standpoint.

Take a non-covalent interaction between a receptor R and a ligand L and their complex form RL, where [R] and [L] is the concentration of the free molecules and [RL] is the concentration of their bound form. Then

$$R + L \rightleftharpoons RL \tag{1.1}$$

Two processes exist; an association process of the two molecules into their bound form RL; and a dissociation process back to free molecular R and L. The rate at which association or dissociation takes place, depends on the concentrations of each molecular species as:

$$\text{rate of association} = k_{on}[\text{R}][\text{L}] \tag{1.2}$$

$$\text{rate of dissociation} = k_{off}[\text{RL}] \tag{1.3}$$

$k_{on}$ and $k_{off}$ represent the intrinsic disposition of R and L to associate or RL to dissociate respectively. The rate of change of concentration of R, L and RL is as follows:

$$\frac{d[R]}{dt} = \frac{d[\text{L}]}{dt} = k_{off}[RL] - k_{on}[R][L] \tag{1.4}$$

$$\frac{d[RL]}{dt} = k_{on}[R][L] - k_{off}[RL] \tag{1.5}$$

For this system to be in equilibrium (i.e. constant concentrations of R, L and RL), the rate of association must equal the rate of dissociation and using equation 1.4 and 1.5:

$$k_{on}[R][L] = k_{off}[RL] \tag{1.6}$$

$$\frac{k_{off}}{k_{on}} = \frac{[R][L]}{[RL]} = K_D \tag{1.7}$$

where $K_D$ is the dissociation constant that is related to the binding affinity of the interaction.

### 1.5.2 The Thermodynamics of Binding

A second route towards characterising binding affinity is that based on the standard free energy of binding (Gilson et al., 1997). The free energy of binding is the change in free energy when one receptor and one ligand react to form a complex. The free energy of binding can therefore be expressed as

$$\Delta G = U_{RL} - U_L - U_R \qquad 1.8$$

Where $U_{RL}$ is the change in free energy of a solution when the complex RL is added to the system, and $-U_L$ and $-U_R$ are the change in free energy of a solution when one ligand L, and one receptor R, are removed from the system, respectively (Gilson and Zhou, 2007). The chemical potential $U_P$ of a protein can be expressed as

$$u_p = -RT \ln\left(\frac{8\pi^2}{C_p} \int e^{-(U_{(r_p)} + W_{(r_p)})/RT} dr_p\right) \qquad 1.9$$

where R is the gas constant and T the absolute temperature, $C_p$ is the concentration of the protein p, $U(r_p)$ is the potential energy of the protein at the conformation $r_p$ and $W(r_p)$ is the solvation energy at the conformation $r_p$ (Gilson and Zhou, 2007). Substituting equation 1.8 into 1.9 for each species, the free energy of binding can be obtained as:

$$\Delta G = -RT \ln\left(\frac{1}{8\pi^2} \frac{C_R C_L}{C_{RL}} \frac{\int e^{-(U_{(r_{RL})} + W_{(r_{RL})})/RT} dr_{RL}}{\int e^{-(U_{(r_R)} + W_{(r_R)})/RT} dr_R \int e^{-(U_{(r_L)} + W_{(r_L)})/RT} dr_L}\right) \qquad 1.10$$

The system is in equilibrium when the free energy of binding $\Delta G=0$. Therefore equation 1.10 becomes

$$\left(\frac{C_{RL}}{C_R C_L}\right)_{equilibrium} = \frac{1}{8\pi^2} \frac{\int e^{-(U_{(r_{RL})} + W_{(r_{RL})})/RT} dr_{RL}}{\int e^{-(U_{(r_R)} + W_{(r_R)})/RT} dr_R \int e^{-(U_{(r_L)} + W_{(r_L)})/RT} dr_L}\Bigg) \qquad 1.11$$

Multiplying both sides of equation 1.11 by the standard concentration $C^o$ gives

$$(\frac{C_{RL}C^o}{C_RC_L})_{equilibrium} = \frac{C^o}{8\pi^2}\frac{\int e^{-(U_{(r_{RL})}+W_{(r_{RL})})/RT}dr_{RL}}{\int e^{-(U_{(r_R)}+W_{(r_R)})/RT}dr_R\int e^{-(U_{(r_L)}+W_{(r_L)})/RT}dr_L)} \quad 1.12$$

and replacing the concentration in equation 1.10 by the standard concentration $C^o$, the standard free energy of binding is

$$\Delta G^o = -RT\ln(\frac{C^o}{8\pi^2}\frac{\int e^{-(U_{(r_{RL})}+W_{(r_{RL})})/RT}dr_{RL}}{\int e^{-(U_{(r_R)}+W_{(r_R)})/RT}dr_R\int e^{-(U_{(r_L)}+W_{(r_L)})/RT}dr_L)} \quad 1.13$$

Substituting equation 1.12 in 1.13 gives

$$\Delta G^o = -RT\ln(\frac{C_{RL}C^o}{C_RC_L})_{\text{eq}} \quad 1.14$$

and from the kinetics approach and equation 1.7

$$\Delta G^o = -RT\ln(\frac{k_{off}}{k_{on}})_{\text{eq}} \quad 1.15$$

$$K_D = e^{\Delta G^o/-RT} = (\frac{k_{off}}{k_{on}})_{\text{eq}} \quad 1.16$$

This equation links both the kinetics and the thermodynamics of the binding process.

Equation 1.13 can be decomposed into

$$\Delta G^o = \langle U_{RL}\rangle - \langle U_R\rangle - \langle U_L\rangle + \langle W_{RL}\rangle - \langle W_R\rangle - \langle W_L\rangle - T\Delta S^o_{config} \quad 1.17$$

The standard free energy of binding can be decomposed into the enthalpic contribution to binding $\Delta H^o$ and the entropic contribution to binding $T\Delta S^o$ as

$$\Delta G^o = \Delta H^o - T\Delta S^o \quad 1.18$$

In 1.17, $\langle U_{RL}\rangle$ and $\langle W_{RL}\rangle$ are Boltzmann-average potentials for the potential energy and solvation energy respectively. As seen in equation 1.13, in this form, though the integral is taken over all conformations of the species in question, only the low energy contributions contribute significantly to the potential. $T\Delta S^o_{config}$ represents the change in entropy when the receptor and ligand move from the unbound to the bound form, this includes, a loss in translational, and rotational entropy, and change in side-chain entropies. The solvation energies

$\langle W_{RL} \rangle - \langle W_r \rangle - \langle W_L \rangle$ also include an entropic component related to the freedom of water molecules. Equation 1.17 forms the basis of many binding affinity predictive models (see section 1.5.7), where a number of physics-based descriptors representative of the energetic terms in equation 1.17, are calculated. The modelling of the three main components of equation 1.17, the potential energy, solvation energy and entropy are discussed below in sections 1.5.3, 1.5.4, and 1.5.5 respectively. Further, in section 1.5.6, the use of statistical potentials and the role of miscellaneous descriptors for affinity are also mentioned.

### 1.5.3 Potential Energy

The potential energy of a macromolecule can be thought of as an energy surface, which is a function of the atomic, nuclear, and electron positions in space. The parameter space covering the positions and motions of electrons for large macromolecules is still too large to be dealt with using quantum mechanical methods. A more accessible alternative is the use of empirical force-fields, where the energy of a system is a function of the nuclear positions only (Leach, 2009). In general most of the molecular modelling force-fields describe both the intra- and intermolecular forces within a system. An example of which is the potential energy function $U(r^N)$ shown in Figure 1.3.

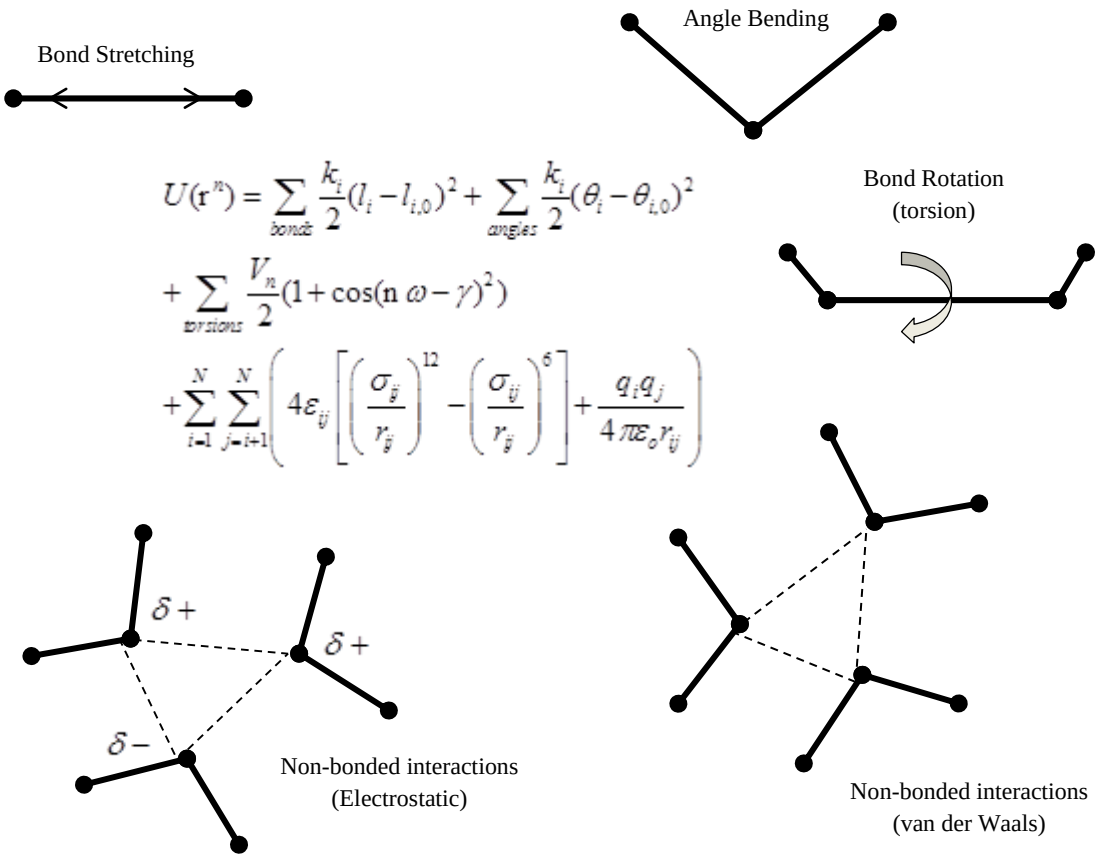



**Figure 1.3: Representation of the main energetic terms involved in a molecular mechanics force field describing the potential energy of a molecule or system.**
These are bond stretching, angle bending, torsional terms and non-bonded interactions. Figure derived from (Leach, 2009).

The intramolecular forces are described by terms which represent an energetic penalty associated with some deviation of bond lengths, angles or rotations from a reference state (Leach, 2009). The intermolecular forces may include energetic terms such as the Lennard-Jones 12-6 Van der Waals potential and the Columbic energy. The $r_{ij}^{-12}$ term in the Lennard-Jones potential is based on the Pauli exclusion principle, which states that no two particles can occupy the same region of space. Computationally, this prevents the generation of clashes that may arise from two interacting molecules. The $r_{ij}^{-6}$ term is related to correlated motions of electrons known as London dispersion forces, which give rise to spontaneous dipoles or higher multipoles. In turn these dipoles may induce

electrostatic complementarity, which decreases the potential energy of the system. The Coulombic energy represents the favourable electrostatic complementarity arising from charged particles within an electric field. Charged particles arise when electrons concentrate around atoms with large electronegativities, and deplete elsewhere. This leads to partial atomic charges, leading to polar atoms for those atomic charges that are large enough in magnitude. The potential describing this non-bonded interaction between partial atomic charges is represented by the product of the two-point charges $q_i$ and $q_j$, separated by a distance $r_{ij}$, where $\varepsilon_o$ represents the permittivity of free space. All terms in the empirical force-field shown in Figure 1.3 are a function of N atoms and their positions in space (**r**). Each term can be computed separately and therefore varying levels of sophistication can be added as required.

### 1.5.4 Solvation Energy

Protein interactions are surrounded by salt-water, which in turn has a significant effect on binding. The solvation energy represents the proteins' interactions with water and its effect can be summarized into the dielectric screening of water and the hydrophobic effect. Dielectric screening results from the different permittivities of different mediums. Water has a high dielectric constant, which makes the interaction between charged, and polar atoms in water favourable. Atoms in areas of low solvent accessibility, those forming part of the interface when a complex is bound, have a lower effective dielectric constant. There may therefore be an energetic penalty associated with moving polar atoms out of water and into a binding site (Gilson and Zhou, 2007). In simple solvation screening models, the dielectric constant is directly proportional to the inter-atomic distance of two particles. Methods such as the Poisson-Boltzmann (PB) (Honig et al., 1993), apart from other considerations, account for the fact that the solvent accessibility surrounding an atom, is also a function of the atoms surrounding it. A second effect of water on the formation of protein interactions is the tendency of non-polar atoms to be brought together and away from water (Kauzmann, 1959, Hildebrand, 1979). This is known as the hydrophobic effect, and is a major driving force in protein folding (Lins and Brasseur, 1995, Dill,

1990) and also an important aspect of protein binding (Tsai et al., 1997). The non-polar parts of the protein exposed to water, restricts the movement of water molecules resulting in the formation of ordered 'water cages'. Bringing non-polar atoms from a solvent exposed site, to a solvent inaccessible site such as the binding interface, results in an increase in the system entropy that decreases the binding free energy (equation 1.18). A common method employed to model the hydrophobic effect implicitly, is to calculate the change in the solvent accessible area of non-polar atoms upon going from the unbound to the bound state (Chen et al., 2004). The addition of surface area terms accounting for the hydrophobic effect in Poisson-Boltzmann implicit solvation methods are known as PBSA (Sitkoff et al., 1994). Faster approximations to the PBSA also exist such as the Generalized Born model with Surface Area (GBSA) (Qiu et al., 1997).

### 1.5.5 Configurational and Side-Chain Entropy

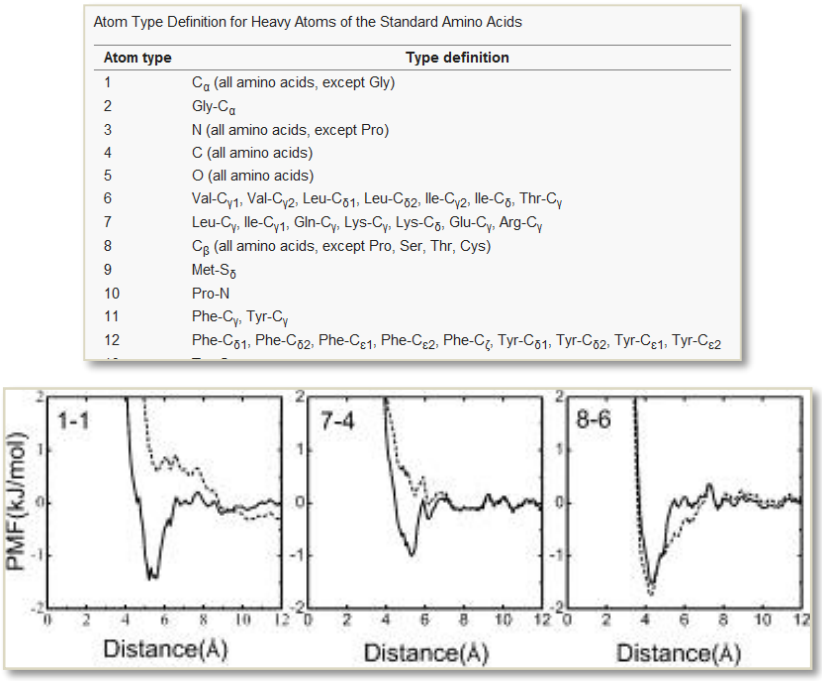
The binding free energy of complex formation, as presented in equation 1.18, shows that complex formation may be either enthalpy or entropy driven. Therefore, the correct modelling of the potential and solvent energies involved in the binding process (ΔH in equation 1.18) still does not give an accurate estimation of the binding free energy. To do so, the change in entropy (ΔS) of the system also has to be characterised. One entropic aspect important for binding is the change in entropy experienced by water molecules described by the hydrophobic effect. This is generally accounted for in solvation energy models such as those described in the previous section. The formation of a complex also involves changes in the configurational (rotational and translational) entropy of the receptor and ligand. In general it is widely assumed that the changes in rotational and translational entropy have negligible contribution to the binding free energy in aqueous solutions at 1 M standard state (Yu et al., 2001) or that they are constant across different interactions. However, it has been shown that for complexes, which are not tightly bound, the change in configurational entropy is not the same as that of a tightly bound complex (Chen et al., 2004). Upon binding, the side-chains of the receptor and ligand become conformationally restricted if they form part of the binding interface. This results

in a reduction in entropy upon binding. Traditional methods of accounting for the change in side-chain entropies make use of rotamer libraries (Dunbrack and Cohen, 1997, Chandrasekaran and Ramachandran, 1970, Dunbrack and Karplus, 1993, Dunbrack and Karplus, 1994) or simply the number of rotatable bonds affected upon binding (Finkelstein and Janin, 1989).

### 1.5.6 Knowledge-Based-Potentials and Miscellaneous Descriptors

The approaches discussed in previous sections, 1.5.3, 1.5.4 and 1.5.5, in modelling the terms of the binding free energy function presented in equation 1.17, are derived from the underlying physical processes driving complex formation. An alternative method is to use knowledge-based potentials. In this approach, rather than enumerating all potential physical processes responsible for complex formation and affinity, the relative positions of atoms or residues are used as an indication of the validity (in the case of protein folding or docking) or strength (in the case of binding affinity prediction) of the complex in question.

The central hypothesis made by knowledge-based potentials (also referred to as statistical potentials throughout this work), is that the frequency of two specific atoms/residues at a specific distance is an indication of how favourable the contact between the pair of atoms/residues is. More frequently occurring contacts are considered to be favourable and likely the result of capturing some underlying physical process.



Atom Type Definition for Heavy Atoms of the Standard Amino Acids

| Atom type | Type definition |
|---|---|
| 1 | $C_\alpha$ (all amino acids, except Gly) |
| 2 | Gly-$C_\alpha$ |
| 3 | N (all amino acids, except Pro) |
| 4 | C (all amino acids) |
| 5 | O (all amino acids) |
| 6 | Val-$C_{\gamma1}$, Val-$C_{\gamma2}$, Leu-$C_{\delta1}$, Leu-$C_{\delta2}$, Ile-$C_{\gamma2}$, Ile-$C_\delta$, Thr-$C_\gamma$ |
| 7 | Leu-$C_\gamma$, Ile-$C_{\gamma1}$, Gln-$C_\gamma$, Lys-$C_\gamma$, Lys-$C_\delta$, Glu-$C_\gamma$, Arg-$C_\gamma$ |
| 8 | $C_\beta$ (all amino acids, except Pro, Ser, Thr, Cys) |
| 9 | Met-$S_\delta$ |
| 10 | Pro-N |
| 11 | Phe-$C_\gamma$, Tyr-$C_\gamma$ |
| 12 | Phe-$C_{\delta1}$, Phe-$C_{\delta2}$, Phe-$C_{\varepsilon1}$, Phe-$C_{\varepsilon2}$, Phe-$C_\zeta$, Tyr-$C_{\delta1}$, Tyr-$C_{\delta2}$, Tyr-$C_{\varepsilon1}$, Tyr-$C_{\varepsilon2}$ |



**Figure 1.4: Example of Atom-Types and Contact Frequency-Distance plots of a typical knowledge-based potential**.
Top of figure shows the atoms considered in the statistical potential, and bottom of figure shows the potentials (frequency-distance plots) generated for three of the contacts. These include a backbone-backbone contact potential (1-1), a backbone-side-chain contact potential (7-4) and a side-chain-side-chain contact potential (8-6). The figures are taken from the work of (Su et al., 2009).

For example, the contact frequency-distance plots of the Potential-Mean-Force (PMF) potential from Su et al. (2009), show functions very similar to the Lennard-Jones potential (See Figure 1.4). This is characterised by strong repulsions at very short distances, followed by a global minimum on increasing distance, which approaches zero at larger distances. An important aspect which affects the success of statistical potentials, is the reference state taken. Namely, the reference state must account for frequency and volume, and many potentials do in fact differ by their reference state (Zhang et al., 2004, Su et al., 2009, Shen and Sali, 2006). Besides differences in the reference state, different statistical potentials include; atom-based and coarse-grained (Lu et al., 2008, Rykunov and Fiser, 2010) (residue level through centroid or $C_\alpha$, $C_\beta$ distances) potentials; pair potentials and multi-body potentials (Feng et al., 2010); those derived on protein structures for protein folding and stability (Zhou and Skolnick, 2011), and those

derived on protein-protein complexes for docking and binding affinity prediction (Liu and Vakser, 2011). One drawback of knowledge-based potentials is that they do not account for solvation or entropy terms and only recently for protein-ligand interactions has this consideration been attempted (Huang and Zou, 2010). Therefore, one way of thinking about statistical potentials in binding affinity prediction is as an alternative or addition to terms related to the potential energy.

Another class of descriptors termed as 'miscellaneous' descriptors, are again those that do not model a particular physical process, but their presence may capture some underlying physical property that favours complex stability. These include; secondary structure elements, such as the proportion of interface resides which are in alpha helices, or beta sheets; geometrical properties such as interface planarity, volume of empty space at the interface and interface surface complementarity. With this in mind, the inclusion of such descriptors in binding affinity models is primarily exploratory in nature and must be interpreted with caution.

### 1.5.7 Binding Affinity Prediction (BAP) Methods

Between 1989 and 2011, 19 publications have specifically dealt with the prediction of binding affinities for protein-protein complex formation. Most of these Binding Affinity Prediction (BAP) models contain empirical functions where the terms include relevant enthalpic and entropic contributions to binding (as the terms described in sections 1.5.3, 1.5.4 and 1.5.5); most commonly, terms for the contribution of electrostatics, hydrophobic burial, hydrogen bonding, side-chain entropy etc. (Novotny et al., 1989, Horton and Lewis, 1992, Krystek et al., 1993, Vajda et al., 1994, Nauchitel et al., 1995, Xu et al., 1997, Weng et al., 1997, Noskov and Lim, 2001, Ma et al., 2002, Jiang et al., 2005, Audie and Scarlata, 2007, Bougouffa and Warwicker, 2008, Bai et al., 2011). The second category of BAP models is model's that consist of statistical potentials (Zhang et al., 1997, Jiang et al., 2002, Liu et al., 2004, Su et al., 2009). Here the relative positions of atoms or residues observed in experimental structures are used to

infer a potential of mean force that is then correlated to binding affinity (see section 1.5.6).

On analysis of the aforementioned BAP models, the following limitations were identified:

I. Models restricted to complexes for which the component parts undergo little to no conformational changes upon complex formation.
II. Assumed that complexes and component parts exist as static structures (assumed all proteins are rigid entities)
III. Routine use of Linear Regression.

#### *1.5.7.1 Models Restricted to Proteins that Undergo Little to No Conformational Changes Upon Complex Formation.*

Most of the BAP models are designed under the assumption that minimal to no conformational changes take place upon complex formation. To satisfy this assumption, the complexes used to test the models are specifically selected to be rigid-body complexes. The descriptor calculations therefore generally take the form of:

$$\text{Complex} - (\text{Receptor}_{\text{Bound}} + \text{Ligand}_{\text{Bound}}) \quad 1.19$$

where the monomers are assumed to be pre-organised in their bound conformation when in their free state. Moreover, up until the work of Liu et al. (2004), careful analysis of the complexes used for training and testing were limited to protease-inhibitor pairs (Krystek et al., 1993, Nauchitel et al., 1995, Vajda et al., 1994, Wallqvist et al., 1995, Zhang et al., 1997) with the addition of a few other high-affinity rigid complexes such as such as Barnase-barstar, the insulin dimer, the α and β chains of deoxyhaemoglobin and lysozyme-antibody complexes (Ma et al., 2002, Horton and Lewis, 1992, Audie and Scarlata, 2007, Bougouffa and Warwicker, 2008, Jiang et al., 2002, Weng et al., 1997, Xu et al., 1997). For some of these models, the correlation with experimental binding affinities is exceptionally high. However, as seen from the restrictions on

conformational changes and on the diversity of structures used, the models are highly biased, and the final correlation coefficients should be treated with caution. This bias was confirmed in the work of (Kastritis and Bonvin, 2010; Kastritis et al., 2011) where the top performing BAP energy functions were tested on two recent benchmark datasets, with no restrictions on conformational changes. Correlations with experimental binding affinities were only as high as 0.53 and as low as 0.17 (Kastritis and Bonvin, 2010; Kastritis et al., 2011).

After the introduction of a larger (ranging from 52 to 86 complex structures) and a more diverse set structures by Liu et al. (2004), subsequent work on BAP was characterised by less accurate predictive models. Moreover, the bias was still towards rigid structures and conformational changes were never explicitly accounted for. It is also worth to note, that in a recent affinity benchmark dataset with 144 protein-protein complexes (Kastritis et al., 2011), when considering complexes with limited conformational changes (rmsd <1 Å), a ΔG prediction scheme that only uses the interface area achieves performance similar to more elaborate empirical models.

#### *1.5.7.2 Conformational Flexibility*

Proteins, and even protein complexes, do not exist as static structures but as an ensemble of conformations. As shown in equation 1.13, the binding free energy of a protein complex depends on the Boltzmann weighted average of the energies of the conformational states accessible by the complex, and those accessible by the free monomers. With this in mind, none of the BAP models mentioned above (with one exception (Vajda et al., 1994)) explicitly account for this. Rather all energetic calculations are calculated on a single static structure. In the case of the work of (Vajda et al., 1994), the static restriction is not employed. However, flexibility is still only accounted for the ligands, which in this case are flexible peptides binding an MHC receptor. For these cases, the authors also show that ligand flexibility contributes 30-50% of the free energy change. A recent study (Yang et al., 2009) shows how the inclusion of an ensemble of protein-ligand conformations, obtained from MD simulations, improves the prediction accuracy of affinity scoring functions. Though

promising, this work is again limited to ligand flexibility, which has a significantly lower conformational space than the two components of a binary protein complex.

#### *1.5.7.3 Routine use of Linear Regression*

The diversity in macromolecular interactions and their structural properties (Nooren and Thornton, 2003) suggests that an energetic contribution dominant in a given interaction is not necessarily the dominant contribution in another. For example it is known that protein-protein interfaces tend to be hydrophobic (Young et al., 1994, Chothia and Janin, 1975) and planar (Baker and Der, 2013). However, hydrophilic interfaces are also common (Ben-Naim, 2006) and interfaces can also be protruding (Yura and Hayward, 2009). Moreover, Cho et al. (2006) show that there are specific interaction types based on the functional category of the protein complex, and such interaction types are conserved through the common binding mechanism, rather than through sequence or structure conservation. Effectively, this indicates that generalizations concerning the determinants of protein-protein binding affinity may be limited in the context of a large and diverse dataset of protein complex families. Hence, a model such as Linear Regression (LR), which can only exploit globally well-rounded descriptors, might not be adept for a set of diverse complexes, such as the one used in this work.

All BAP models developed until the work reported in this thesis (those reported in section 1.5.7), that are not statistical potentials, use LR to combine the energetic factors deemed responsible for complex affinity. Effectively, LR seeks a set of descriptors which best describe the dataset as a whole, which means that certain intricacies of a dataset, perhaps represented by a particular set of descriptors that are each specific to different cases, are overridden by descriptors which achieve higher overall, but limited, correlations. For example, electrostatics is a major driving force for small interface formation whereas hydrophobic burial tends to be more significant in larger interface formation (Sheinerman and Honig, 2002). Hence, for a dataset where small interfaces are underrepresented the effect of electrostatics may be underestimated as opposed

to the hydrophobic burial effect. Namely, as datasets become diverse, LR is not a sufficient model to represent such diversity. Rather, feature space-partitioning methods, which can encompass logical reasoning such as: 'if interface is small use these descriptors, if the interface is larger use these other descriptors', is more appropriate; these models are able to subset the feature space so that different features can contribute in different situations. The topic of machine learning is detailed the following section 1.6.

### 1.5.8 Hotspot Prediction

The binding free energy of a complex may also be understood through alanine scanning of residues at its interface. From such scans, it is understood that not all interface residues have marked effects on binding. Rather, only a subset of residues termed 'hotspots' contribute significantly to the binding energy of the complex (Clackson and Wells, 1995, Bogan and Thorn, 1998). Traditionally, a residue is a hotspot, if upon its substitution into alanine, is causes a reduction in binding free energy of 2kcal/mol or higher. Analysis on protein-protein interfaces and hotspot residues has shown that: hotspots tend to occur in regions of low solvent accessibility (Bogan and Thorn, 1998); Tyr, Trp and Arg are the most frequent hotspots (Ma and Nussinov, 2007, Bogan and Thorn, 1998); and hotspot tend to cluster into densely packed regions known as hotregions (Keskin et al., 2005). As mentioned in section 1.4.1, the major attraction of hotspot residues is that they are crucial for targeting of protein-protein interfaces with small drug-like molecules (Fry, 2012, Thangudu et al., 2012, Arkin and Wells, 2004). This has led to the development of several computational hotspot prediction algorithms (Kortemme and Baker, 2002, Cho et al., 2009, Lise et al., 2009, Lise et al., 2011, Tuncbag et al., 2010, Tuncbag et al., 2009, Xia et al., 2010, Zhu and Mitchell, 2011, Grosdidier and Fernandez-Recio, 2012, Morrow and Zhang, 2012, Wang et al., 2012). The predictors generally use a combination of solvent accessibility and physiochemical descriptors, which are then fed into machine learning algorithms trained on experimental datasets such as *ASEdb* (Bogan and Thorn, 1998)and *BID* (Fischer et al., 2003). For example, *Robetta* (Kortemme and Baker, 2002) uses an empirical energy function using potential, solvation and entropic energy terms. These include, the Lennard-Jones potential,

orientation dependant hydrogen bonding, shape complimentarity and an implicit solvation model. *KFC2* (Zhu and Mitchell, 2011) consists of two support vector machine models (*KFC2a* and *KFC2b*). Besides standard energy terms such as van der Waals terms and hydrogen bonding, the solvent accessibility and local flexibility surrounding the target residue, were also included as features. Hotpoint (Tuncbag et al., 2010) takes a more efficient approach by basing the hotspot prediction only on solvent accessibility and a pair potential. The authors claim that even with such minimal features, the method still outperforms *Robetta* and *KFC2*. One major limitation of the aforementioned algorithms is that they have been trained and tested on very limited alanine scanning databases, namely the *ASEdb* (Thorn and Bogan 2001) and *BID* (Fischer, Arunachalam et al. 2003). The shortcoming of these datasets as benchmarks has been highlighted in (Xia, Zhao et al. 2010; Moal and Fernandez-Recio 2012).

In Chapter 5, two hotspot prediction algorithms (*RFSpot* and *RFSpot_KFC2*) are designed and their performance compared to a number of hotspot predictors. The hotspot predictors are then used in scheme which involves alanine scanning for the prediction of off-rate changes upon mutation, as described in Chapter 6.

## 1.6 Machine Learning

Machine Learning (ML) is a subfield of computer science that deals with frameworks for identifying and exploiting patterns in data (Bishop, 2007). Nowadays, in all its forms, ML has become an enabling technology in a number of fields and industries, and even if it is not immediately obvious, your first guess should be that at any moment, you are making use of something where ML has been implemented in. This includes machine vision algorithms for your camera's face recognition feature (Turk and Pentland, 1991); your e-mail spam filter and virus software on your computer (Bishop, 2007); or even your movie recommendations on Netflix (Ricci et al., 2011). The search of patterns from data is neither a novel idea nor limited to artificial systems. Rather, throughout history, most of what we know today about the world around us, is based on observers uncovering regularities and patterns in some physical phenomena. For example, Johannes Kepler only developed the empirical laws of planetary motion

by discovering consistencies in the astronomical observations of Tycho Brahe in the 16th century. Pattern recognition (not necessarily learnt recognition) is an inherent characteristic of even the simplest forms of living organisms (Bray, 2009). In addition, associative learning is one of the main characteristics of organisms with nervous systems and evidence also shows that organisms without such a dedicated system are capable of advanced learning behaviour, such as the anticipation of environmental stimuli (Mitchell et al., 2009). The first computational learning algorithms, most commonly, Artificial Neural Networks (ANNs), (the earliest example of which being the perceptron (Rosenblatt, 1958)), are in fact inspired from the human's central nervous systems. ANNs use a number of artificial neurons connected together to learn the appropriate response from a given input pattern. ANNs and other similar supervised machine learning algorithms are concerned with the automatic discovery of regularities in data using computer algorithms (Bishop, 2007). Their aim is to make predictions (apply the appropriate response) on some unseen data based on the regularities they have discovered and based on the comparison of these regularities to those observed in the new data.

Setting up a problem in a ML framework, for instance that of *supervised learning*, invariably requires three main elements; a training dataset of target output values, a set of input features and a learning algorithm. The aim is to make predictions on some unseen data after having learnt a model from the training dataset. The model effectively learns a mapping between the input features and the target output values. Once this mapping is learnt, the model can be invoked to make new predictions on input features calculated on data with unknown target output values. Apart from *supervised learning*, other ML frameworks, which are not necessarily distinct from each other, include *unsupervised learning* which involves the clustering of data into distinct regions without target values to learn on; *Anomaly detection* (both supervised and unsupervised) which involves the identification of irregularities which do not conform to the expected pattern of data and *Reinforcement Learning* where an optimal sequence of decisions are to be made in an environment which is largely unknown. All ML

methods implemented in this thesis are either supervised classification or regression methods.

### 1.6.1 Machine Learning in this Thesis

The ML framework is used consistently throughout this thesis for modelling the stability of protein-protein complexes at various levels. This includes *wild-type* binding affinity (ΔG) prediction, ΔΔG and hotspot prediction; and $\Delta k_{off}$ prediction. It is important to highlight that highly ML specific investigations are beyond the scope of this thesis. For example, there is no motivation to compare and contrast different learning algorithms for the same problem. The general belief employed is that the largest gains in prediction accuracy are best made with better feature design and careful analysis of the dataset. Consequently, in seeking better predictions, an exhaustive evaluation of a number of learning algorithms or parameters in hope of increasing accuracy, is not employed. With this in mind, a conscious effort is made throughout, firstly to validate the choice of learning algorithms in relation to the datasets and features available and secondly, not to use machine learning in a black-box fashion. Figure 1.5 shows the dependencies between different elements of a supervised machine-learning framework. These dependencies are not an exhaustive list, but rather highlight those dependencies that are given careful consideration in this work. These include dependencies considered prior to the learning phase, and those discovered subsequent to it upon analysis of the results.

### 1.6.2 Dependencies in a Supervised Machine Learning Framework

The dataset is the main source of training and benchmarking and its biases and diversity have a direct effect on both the choice of learning model and features employed. Bias in the context of this work is not limited to a class distribution bias (commonly referred to as dataset imbalance). The datasets used in this thesis (see Methods section 2.1) contain PPIs, which in turn come in many forms; different structural, physiochemical and conformational properties (Moal and Fernandez-Recio, 2012, Kastritis et al., 2011). A dataset which is not a representative sample of this diversity is therefore, also biased. For example, in

section 1.5.7.1 it is shown that previous binding affinity prediction (BAP) models use datasets without conformational and complex-family diversity. In this case, simple learning algorithms such as linear regression are sufficient to produce accurate results ('Dataset->Choice of Learner' in Figure 1.5).

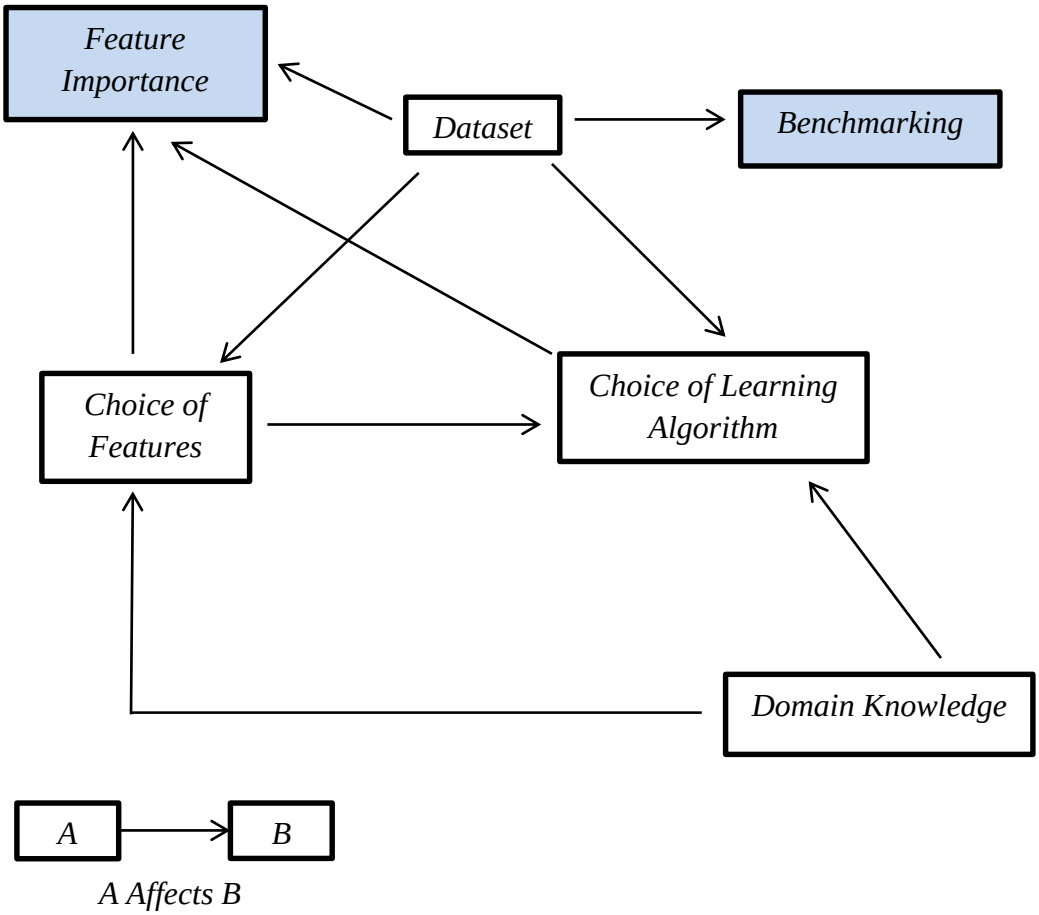



**Figure 1.5: Dependencies in a supervised machine learning framework.** White boxes highlight dependencies that must be considered prior to learning; blue boxes highlight dependencies affecting the interpretation of results subsequent to learning. Though not an exhaustive list of dependencies, the ones shown here are those that are discussed in this thesis. Some of these dependencies were considered pre-emptively; for example when selecting descriptors or learning algorithms, whereas others were discovered upon analysis of the results; for example when assessing the results of feature importance measures for different learning algorithms.

The BAP model developed in Chapter 3 made use of a larger and more diverse set of structures. To account for this diversity, new descriptors (for example

those accounting for unbound to bound transitions) and non-linear ML models were introduced ('Dataset->Choice of Learner/ Choice of Features' in Figure 1.5). The choice of the learning algorithm should be an informed one taking into consideration both the domain knowledge of the problem at hand and the nature of the features available for learning ('Choice of Features/Domain Knowledge->Choice of Learner' in Figure 1.5). For instance, features also come in many forms, particularly those for modelling complex stability (See sections 1.5 and 2.2). Some might be good global estimators, whereas others might only hold predictive value within certain ranges. For the latter, a learning algorithm like linear regression cannot exploit these locally predictive regions. Therefore, it is imperative that the choice of learner matches this information. Domain knowledge refers to the prior beliefs we have about the problem at hand. This in turn again affects the learner choice.

An informed interpretation of the results and analysis (through benchmarking and descriptor importance) must also consider the relevant dependencies (see Figure 1.5). For example in Chapter 3 it is observed that descriptors identified as being important for modelling affinity are not only a function of the dataset at hand, but also a function of the learning algorithm employed and other features available to the learning algorithm. Moreover, in in section 7.3.3 it is shown that certain descriptors are highly important to the characterization of certain off-rate mutations in the dataset but not for others. Therefore, global feature importance measures are not necessarily the appropriate choice, particularly in diverse datasets. In summary, this shows that any tentative conclusions on the importance of a particular descriptor must also be made in light of all of its dependencies.

## 1.7 Outline of Thesis

### 1.7.1 Motivations Behind this Work

This work deals directly with computational experiments investigating protein-protein complex stability at various levels. These include prediction models for binding affinity, dissociation rates and hotspots. Characterising the effect of mutations on complex stability forms a major part of this thesis, and emphasis is given on the detection of rare mutations that can further enhance the stability of protein-protein interactions. The understanding is that being able to do so, is central to future computational drug design algorithms (See section 1.4). Though ultimately, the central motivation is to design accurate predictive models, parallel to this, an equal goal is that of uncovering determinants of complex stability.

The intention is that this thesis asks the questions which have not been asked before, or for those which have been, improvements are made which directly address the deficiencies of current methods. In some instances, this work might take unconventional approaches, be it the use of the uncommon Radial Basis Function learner in Chapter 3; the position-specific models designed in Chapter 4; or even, the design of descriptors derived from hotspots for characterising changes in dissociation rates in Chapter 6. The thesis attempts to answer most of the questions that come to mind, but for those that remain answered, the hope is that the questions raised are worthy of further investigation. The aim for this thesis was also to be in line with what we do know about complex stability. For instance, building a large and diverse set of descriptors (not limited to biophysical descriptors) and using them for characterizing the stability of protein-protein interactions may seem as a naïve or ungrounded pursuit to some. On the other hand what best way forward than to make use of, otherwise forgotten, descriptors that have been carefully designed by other researchers in the field in these last years. With this in mind, this methodology is not to be confused with one where a 'bunch' of descriptors are thrown blindly, in hope of finding something that correlates with our target. Firstly, the works of authors

that do so are reported for their deficiencies, and all models developed in this work are analysed for their biophysical plausibility. Recent publications (Kastritis and Bonvin, 2013) also put forward the argument that a simple model of buried surface area or simple biophysical models can still achieve reasonable correlations for binding affinity. The implication being, usually explicitly mentioned, that more complex binding affinity models (where complexity refers to large sets of descriptors and machine learning models), improve very little. These arguments I feel, fail to make a distinction between good correlation and high correlation. For uncovering a relationship, good correlations are acceptable, as they are for proofs of concept. A predictive algorithm with reasonable correlation is however unusable in most practical situations. Any predictive algorithm designed in this work, and others', remain purely explorative in nature until significantly high accuracies are achieved. Only until then may such algorithms become standard protocol. Predictive performance is thus one of the major motivations behind this work. It should be noted that the algorithms designed here are part of an on-going pipeline of algorithms that came before and will come after. Attention is therefore given to highlighting clearly where the algorithms fail, which structures we still cannot characterise well, which mutations are harder to predict and how we might improve. In a similar vein of reasoning, all publications resulting from these investigations contain model prediction lists for direct comparison analysis by other researchers. Careful benchmarking is also employed, be it with the use of alternative cross-validation folds, or diverse and validated datasets.

### 1.7.2 Chapter Summaries and Themes

*Chapter 2*: The datasets, stability descriptors and machine learning models used throughout this thesis are summarized and described here. Following this, the hotspot prediction algorithm developed in this work (RFSpot), is described and benchmarked against other hotspot predictor algorithms. In addition a number of descriptors generated using the predicted hotspots are presented. These are termed as hotspot descriptors and subsequently used for the prediction of off-rate changes upon mutation in Chapters 6-8.

*Chapter 3*: In this chapter the prediction of *wild-type* binding affinities on a diverse set of protein-protein interactions is investigated. In contrast to previous related work, the structures cover a wide range of complex families and conformational changes upon complex formation; thereby addressing the limitations associated with the BAP methods up until this work. Moreover, non-linear machine learning algorithms were used for modelling and the use of unbound structures and conformational ensembles was also introduced into the descriptor calculations. A number of molecular descriptors were calculated which include, biophysical, statistical and miscellaneous descriptors. The prediction model (a consensus of four machine learning algorithms) achieves a cross-validated correlation coefficient with experimental affinities of R=0.77. Significant reduction in accuracy is observed for complexes undergoing conformational changes and those for which their experimental affinities have not been corroborated.

*Chapter 4:* In this chapter the prediction of changes in binding affinity upon mutation (*ΔΔG*) is studied. The models are benchmarked in CAPRI round 26 on a blind set of *circa* 1800 mutations on two different protein drugs each binding the stem of the flu virus hemagglutinin. For the first round, a *ΔΔG* predictor based on similar principles as those presented in Chapter 3 are used. For the second round, a *ΔΔG* model that exploits correlations between similar mutations at a given mutation site, is designed. For both rounds, the predictions compared favourable to other competing groups, and also ranked as the top predictor for one of the protein drug targets. The difficulty in such scenarios, is that datasets available for training are mostly dominated with alanine mutations, which tend to be neutral or destabilizing (affinity decreasing). On the other hand, stabilizing mutations (affinity increasing) are rarely alanine mutations and only form ~2-5% of all the 1800 mutations considered. In turn, the detection of these affinity increasing mutations is central for high affinity drug binding.

Chapter 5: In this chapter, two hotspot prediction algorithms (*RFSpot* and *RFSpot_KFC2*) are designed and benchmarked against a number of hotspots predictors. The results confirm the importance of having solvent-accessibility related descriptors and more comprehensive ΔΔG datasets.

*Chapter 6*: This chapter approaches complex stability from a more specific facet, that of the dissociation rate. The question here shifts to that of understanding what makes a complex remain bound once the complex has formed. A dataset of 50 complexes with 713 mutations for which their $\Delta k_{off}$ was measured experimentally was extracted from the SKEMPI database (Moal and Fernandez-Recio, 2012). Computational alanine scans, using a number of hotspot prediction algorithms, were performed on the *wild-type* and mutated interfaces. The hotspots predicted from these scans are used to design a number of hotspot related descriptors, which are correlated with $\Delta k_{off}$. When compared to molecular descriptors, the hotspot descriptors achieve consistently higher correlations. The 'averaging out effect' of energetics across an interface when using molecular descriptors and the synergy of hotspot residues are proposed as the two main contributors to the success of the hotspot descriptors.

*Chapter 7:* In the previous chapter, hotspot descriptors are introduced and benchmarked against molecular descriptors, as estimators of $\Delta k_{off}$. This chapter goes one step further and feeds both sets of hotspot and molecular descriptors into ML regression and classification algorithms. Besides the numerical prediction of $\Delta k_{off}$, emphasis is also put on the detection of the rare, residence time increasing ($k_{off}$ increasing) mutations which amount to < 5% of the off-rate dataset. ML models with hotspot descriptors show consistently better predictive performance both in the numerical prediction and for the detection of $k_{off}$ increasing mutations. In order to see whether certain classes of mutations are harder to characterise, the 713 off-rate mutation dataset is subset into data regions, and results analysed separately for each. Predictions for mutations occurring at the rim region of protein complex interfaces for example are less accurate to those at the core of region of interfaces. The relationships between different descriptors and different regions of the dataset are studied using descriptor-data region networks. These networks uncovered highly specific relationships between descriptors and certain classes of mutations, and conversely, descriptors that are broadly predictive over a number of mutation classes. The effects of conformational changes and alternative cross-validation routines, on predictive accuracy, are also reported.

*Chapter 8:* In chapters 7 and 8 it is shown how counting the energies of hotspot energies, pre- and post-mutation provides an accurate description of changes in $k_{off}$. Here, the focus shifts towards understanding to which extent the off-rate of a complex is affected by the distribution of hotspots. For example, studies have shown that hotspots are likely to occur at the core regions of an interface and tend to cluster into hotregions. Though these two properties are observed on protein-protein interfaces, their link to stability is only implicated. The main motivation behind this chapter is to uncover advantages, if any, of hotspot distributional properties, by assessing the effect they have on the dissociation rate. As a result of the investigations, it is found that hotspots in the core region are solely critical for the stability of large complexes. For small complexes, rim hotspots become as important and their role is no longer secondary. The intention of introducing distribution into the equation of stability is to be able to make more informed decisions on 'where' to mutate when designing computational interactions.

# Chapter 2

# 2 Materials & Methods

In this chapter, the datasets (section 2.1), stability descriptors (section 2.2) and machine learning models (section 0) used in this thesis are presented. The performance measures applied for the assessment of model predictions and descriptors are also detailed in section 2.4.

## 2.1 Datasets

### 2.1.1 Dataset for Binding Affinity ($\Delta G$)

The structures and experimental affinities for the recently published binding affinity benchmark (Kastritis et al., 2011) were used as the main source for training and testing the BAP described in Chapter 3. As listed in the appendices Table 10.2, this dataset consists of a total of 144 complex structures for which the crystal structures of each complex,

along with each of its unbound components at high resolution (< 3.25Å), are available. To avoid redundancy and the potential for over training, complexes with high sequence identity are not included; this facilitates the use of cross-validation routines such as leave-one-out for benchmarking test predictions. A key aspect of this dataset, which improves upon previous datasets used in BAP, is its diversity:

- Several receptor/ligand protein-binding partners undergo significant conformational change upon complex formation, of which some exhibit disorder to order transitions.
- Complexes, within different protein families, cover a wide range of functions; a total of 19 Antibody/Antigen, 40 Enzyme/Inhibitor, 21 Enzyme-regulatory/accessory chains, 17 G-protein binding proteins, 13 Receptor containing complexes and 34 Miscellaneous.
- Wide range of affinities. A total of 20 high affinity ($K_D < 10^{-10}M$), 90 medium affinity ($10^{-10}M < K_D < 10^{-6}M$) and 34 low affinity ($K_D > 10^{-6}M$).

#### *2.1.1.1 Validated Set*

The affinities available for the protein complexes in the binding affinity benchmark come from a number of experimental methods including isothermal titration calorimetry, surface plasmon resonance, stopped flow fluorimetry and other spectroscopic techniques. For a number of complexes, more than one group measured the $K_D$ values or an additional experimental technique was used. For such measures that are within 1 kcal $mol^{-1}$ of each other, the complexes were said to form part of the 'validated set'. This high-quality subset is used to assess to which extent experimental error in affinities affects the model predictions. One should note that in this validate set the diversity in affinity and complex families is still present. Affinities range between 13 kcal $mol^{-1}$ and complex families include; 3 antibody/antigen complexes, 16 enzyme/inhibitor complexes, 5 enzyme substrate complexes, 5 enzyme complexes, 8 G-protein binding complexes, 7 receptor-ligand complexes and a remaining 13 miscellaneous complexes.

### 2.1.2 Dataset for Off-Rate ($\Delta k_{off}$)

The structures and experimental off-rates from the SKEMPI database (Moal and Fernandez-Recio, 2012) were used as the main source of benchmarking descriptors, training and testing models for $\Delta k_{off}$ prediction in Chapter 6, 7 and 8. *Wild-type* and mutant $k_{off}$ values were transformed into $\Delta\log_{10}(k_{off})$ using

$$\Delta\log_{10}(k_{off}) = \log_{10}(k_{off})^{\mathrm{Mut}} - \log_{10}(k_{off})^{\mathrm{WT}} \quad 2.1$$

Where the value range is, $-8.6 < \Delta\log_{10}(k_{off}) < 6.5$ with a mean of 0.7 (destabilizing). The 713 off-rate mutations from SKEMPI are also subdivided into the following data regions for analysis: Single-Point (SP) alanine mutations, 361; SP non-alanine mutations, 155; SP mutations, 516; Multi-Point (MP) mutations, 197; SP mutations to polar (Q, N, H, S, T, Y, C, M, W) residues, 39; SP mutations to hydrophobic (A, I, L, F, V, P, G) residues, 309; SP mutations to charged (R, K, D, E) residues, 68; mutations exclusively on core regions, 272; rim regions, 79; support regions, 114; mutations on complexes of Large-Interface-Area (>1600 Å$^2$) , 355 and Small-Interface-Area (<1600 Å$^2$), 358. The off-rate dataset is listed in appendices Table 10.3.

An assessment of how severely variations in experimental temperature, ionic strength and pH can introduce noise into $\log_{10}(k_{off})$ and $\Delta\log_{10}(k_{off})$ was also performed. Firstly, 635 of the 713 values come from experiments reported to be performed in the 295–298K range, and 72 values either did not have their temperature reported, or were reported as 'room temperature' or 'standard conditions', corresponding to the 293–298K range. The remaining six experiments were performed at 323K. Thus, only 0.8% of the data lies outside of a 5K temperature range. Although not reported in the SKEMPI database, most of the rate constants were determined using surface plasmon resonance or stopped-flow fluorescence in a relatively narrow range of standard buffer conditions. Further, ionic strength and pH predominantly affect the rate of association rather than the rate of dissociation; electrostatic shielding and changes in protonation state influence the long-range forces which drive protein

association, rather than the short-range forces which keep the buried surfaces of the binding partners together. For instance, in the M3-XCL1 complex, in which ionic strengths in the 0.2 to 1.5 M NaCl range were investigated, the rate of association varied by over 70-fold, while the rate of dissociation varied by less than 3 fold (Figure 2C and Table III of Alexander-Brett and Fremont (2007)). Similarly, in a study of a VEGF-antibody interaction, varying pH in the 6.5–8.5 range resulted in around 30% variation in dissociation rate, while varying the ionic strength in the 10–1000 mM range produced a two-fold change in $k_{off}$ (Moore et al., 1999). Even assuming a large three-fold standard error in $k_{off}$, this would result in a standard error of $3/\ln 10 \approx 1.3$ in $\log k_{off}$ (Moore et al., 1999). Lastly and most importantly, the assumption was made that though reference states may change across experimental methods and studies, within a given experiment the reference state is constant for the experimental determination of the *wild-type* and its mutants, which tend to be generated within the same experimental work. Given that we training is performed on values for $\Delta\log_{10}(k_{off})$ as shown in equation 2.1, any systematic variations associated with experimental conditions are eliminated, this issue is less likely to be prominent for mutation prediction as it is for *wild-type*.

### 2.1.3 Off-rate Classification Data Sets (CDS1 and CDS2)

The 713 off-rate mutations in the previous section of 2.1.2 are partitioned into ($\Delta\log_{10}(k_{off})<-1$), representing the stabilizing portion of the dataset, and ($\Delta\log_{10}(k_{off})>0$), representing the neutral to destabilizing portion of the dataset (referred to as CDS1 –Classification Dataset 1). The motivations behind the thresholds of CDS1 are two-fold. Firstly, previous error estimates show that experimental noise in the data can be as high as 2kcal/mol (Moal et al., 2011, Moal and Fernandez-Recio, 2012). Experimental noise causes miscategorization errors when converting $\Delta\log_{10}(k_{off})$ from continuous values to categorical bins, and therefore, the exclusion of data-points within [−1, 0] should reduce sufficiently the number of miscategorization errors between stabilizing and neutral/de-stabilizing mutations. Secondly, being able to detect stabilizing mutations from neutral ones is an important aspect of interface design (see section 1.4.3). A total of 43% of the mutations lie within the range of [0, 1].

Therefore, the removal of $\Delta\log_{10}(k_{off})$ within the range [−1,0] still allows a sufficient amount of neutral mutations. This data subset, results in a dataset of 501 neutral to destabilizing mutations (referred to as non-stabilizing mutations) and 31 stabilizing mutations. To further investigate the discrimination ability of the descriptors, an additional threshold satisfying $|\Delta\log_{10}(k_{off})|$ >1 is also investigated. This dataset which removes most of the neutrals is referred to CDS2 – Classification Dataset 2.

### 2.1.4 Dataset for Hotspot ($\Delta\Delta G_{ALA}$)

All single-point alanine mutations, limited to the complex interfaces, were extracted from the SKEMPI database. This totals to a set of 635 non-redundant mutations with experimental $\Delta\Delta G$ in 59 different complexes and 154 hotspot residues with $\Delta\Delta G$ >= 2 kcal/mol. All hotspots represent the positive training examples and anything, which is not a hotspot ($\Delta\Delta G$ < 2 kcal/mol) as negative training examples. The hotspot dataset is listed in the appendices Table 10.4.

## 2.2 Stability Related Descriptors

A number of stability related descriptors are calculated and listed in Table 2.1. These include descriptors related to the potential and solvation energy, entropy related descriptors, statistical potentials and a number of miscellaneous descriptors. These different classes of descriptors have been described in the introductory section of 1.5.

**Table 2.1: Stability Related Descriptors.**

A list of stability related descriptors calculated in this thesis. The descriptors are categorized under four sections; *Potential / Total Energy*, *Solvation Energy*, *Entropy*, *Statistical Potentials* and *Miscellaneous Descriptors*. It should be noted that some of the descriptors are not exclusive to one type of category, but are only included one for ease of reference. The entries in the columns ΔG (Chapter 3), ΔΔG (Chapter 4), $\Delta k_{off}$ (Chapter 7) and HS - Hotspots (Chapter 5), indicate whether the descriptor was used in the respective predictive models. Note that this is only an indication of a descriptor being available to the learning models, and not necessarily the case that the descriptor formed part of the final prediction model. Those which do, are reported at the respective chapters. Not included in the table are all FoldX energy terms which are used for ΔΔG, $\Delta k_{off}$ and HS prediction models (Schymkowitz et al., 2005).

| Potential / Total Energy | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Descriptor Type** | **Description** | **Note / Package** | **Reference** | **ΔG** | **ΔΔG** | **$\Delta k_{off}$** | **HS** |
| ROS_HBOND | Directional H-Bonding Potential | PyRosetta | (Chaudhury et al., 2010) | Y | | | |
| H_BOND | 12_10 Hydrogen Bonding Potential | Firedock | (Andrusier et al., 2007) | Y | | | |
| PI_PI | Orientation Independent pi-pi | Firedock | (Misura et al., 2004) | Y | | | |
| CATION_PI | Orientation Independent catian-pi | Firedock | Misura, Morozov et al. 2004) | Y | | | |
| ALIPHATIC | Orientation Independent aliphatic-aliphatic | Firedock | Misura, Morozov et al. 2004) | Y | | | |
| ROS_TOTAL | Total Energy | PyRosetta | (Chaudhury et al., 2010) | Y | | | |
| ACE22_ALL | Total energy | CHARMM 22 Forcefield | (Schaefer and Karplus, 1996) | Y | | | |
| STC_H | STC Enthalpy | STC package | (Lavigne et al., 2000) | Y | | | |
| STC_G | STC free energy | STC package | (Lavigne et al., 2000) | Y | | | |
| ROS_FA_ATR / PY_fa_atr | Lennard-jones attractive | PyRosetta | (Chaudhury et al., 2010) | Y | Y | Y | Y |
| ROS_FA_REP PY_fa_rep | Lennard-jones repulsive | PyRosetta | (Chaudhury et al., 2010) | Y | Y | Y | Y |
| PY_fa_dun | Internal energy of side-chain rotamers as derived from Dunbrack's statistics based pair term | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_fa_pair | Favors salt bridges | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_hbond_lr_bb | Backbone-backbone H-bonds distant in primary sequence | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_hbond_sr_bb | Backbone-backbone H-bonds close in primary sequence | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_fa_Intra_rep | Lennard-jones repulsive between atoms in the same residue | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_hbond_bb_sc | H-bond energy sidechain-backbone | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_hbond_sc | H-bond energy sidechain-sidechain | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| PY_pro_close | Proline ring closure energy | PyRosetta | (Chaudhury et al., 2010) | | Y | Y | Y |
| ROS_CG_VDW | Coarse grained VDW | PyRosetta | (Chaudhury et al., 2010) | Y | | | |
| ACE22_COUL / ACE19_COUL | Coulombic Energy | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE22_ELEC / ACE19_ELEC | Total Electrostatic (ACE_INTE + SELF) | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE22_INTE / ACE19_INTE | COUL+SELF | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| CHARM_total | Total Energy | CHARMM 19 Forcefield | Schaefer and Karplus 1996) | | Y | Y | Y |
| CHARM_elec | Electrostatic Energy | CHARMM 19 Forcefield | Schaefer and Karplus 1996) | | Y | Y | Y |
| CHARM_vdwaals | VDW potential | CHARMM 19 Forcefield | Schaefer and Karplus 1996) | | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NUM_HB | Number of interfacial Hydrogen Bonds | HBPlus | (McDonald and Thornton, 1994) | Y | | | |
| NUM_SB | Number of interfacial Salt Bridges | HBPlus | (McDonald and Thornton, 1994) | Y | | | |
| NUM_WB | Number of interfacial Water Bridges | HBPlus | (McDonald and Thornton, 1994) | Y | | | |
| **Solvation Energy** | | | | | | | |
| **Descriptor Type** | **Description** | **Note** | **Reference** | **ΔG** | **ΔΔG** | **$\Delta k_{off}$** | **HS** |
| DELISI_SOLV | Atomic Desolvation Energies | ACE – Atomic Contact Energies | (Zhang et al., 1997) | Y | | | |
| LK_SOLV / PY_fa_sol | The Lazaridis-Karplus effective energy function | PyRosetta | (Lazaridis and Karplus, 1999) | Y | Y | Y | Y |
| SASA | SASA model | Ferrara et al. 2002 | Ferrara et al. 2002 | Y | | | |
| ROS_CG_ENV | Rossetta Cbeta Potential | PyRosetta | (Chaudhury et al., 2010) | Y | | | |
| ROS_CG_BETA | Rosetta Environment Potential | PyRosetta | (Chaudhury et al., 2010) | Y | | | |
| ACE22_SCRE / ACE19_SCRE | Electrostatic Screening | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE22_SELF / ACE19_SELF | Electrostatic Self Energy | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE22_SOLV / ACE19_SOLV | Sum of SELF and SCREEN | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE22_HYDR / ACE19_HYDR | Hydrophobic Burial | CHARMM 22/19 Forcefield | (Schaefer and Karplus, 1996) | Y | Y | Y | Y |
| ACE19_SASL | SASA Solvation Energy | CHARMM 19 Forcefield | (Schaefer and Karplus, 1996) | | Y | Y | Y |
| CHARM_gb | Generalized Born Implicit Solvation Energy | CHARMM 19 Forcefield | (Schaefer and Karplus, 1996) | | Y | Y | Y |
| CHARM_sasa | Hydrophobic Solvation Energy | CHARMM 19 Forcefield | (Schaefer and Karplus, 1996) | | Y | Y | Y |
| CHARM_gb+sasa | Generalized Born + Hydrophobic Solvation Energy | CHARMM 19 Forcefield | (Schaefer and Karplus, 1996) | | Y | Y | Y |
| STC_S_SOL | Hydrophobic Burial | STC package | (Lavigne et al., 2000) | Y | | | |
| **Entropy** | | | | | | | |
| **Descriptor Type** | **Description** | **Note** | **Reference** | **ΔG** | **ΔΔG** | **$\Delta k_{off}$** | **HS** |
| S_TR | Change in rotational+translational entropy upon complex formation | | | Y | | | |
| S_R | Change in rotational entropy upon complex formation rotational | | | Y | | | |
| S_T | Change in translational entropy upon complex formation | | | Y | | | |
| S_VIB | Change in vibrational entropy upon binding using normal modes via M1 scheme | | (Carrington and Mancera, 2004) | Y | | | |
| STC_S_SC | Entropy changes arising from restriction of side-chain conformation upon binding | STC Package | (Lavigne et al., 2000) | Y | | | |
| S_GP_ALL2 | Disorder to order transitions | | (Zhou, 2004) | Y | | | |
| S_GP_INT2 | Disorder to order transitions | | (Zhou, 2004) | Y | | | |
| S_WLC_ALL2 | Disorder to order transitions | | (Zhou, 2001) | Y | | | |
| S_WLC_INT2 | Disorder to order transitions | | (Zhou, 2001) | Y | | | |
| STC_S | Total Entropy Change | STC package | (Lavigne et al., 2000) | Y | | | |
| **Statistical Potentials** | | | | | | | |
| **Descriptor Type** | **Description** | **Note** | **Reference** | **ΔG** | **ΔΔG** | **$\Delta k_{off}$** | **HS** |
| ROS_FA_PP | Atomistic pair potential | Protein Folding | (Chaudhury et al., 2010, Simons et al., 1999) | Y | | | |
| ROS_CG_PP | Coarse-grained pair potential | Protein Folding | (Chaudhury et al., 2010, Simons et al., 1999) | Y | | | |
| AP_DARS | Atomic distance dependent | Protein Docking Inc. AP_URS/AP_MPS | (Chuang et al., 2008) | | Y | Y | Y |
| AP_DOPE | Atomic distance dependent | Protein Folding Inc. AP_DOPE_HR – High Res. | (Shen and Sali, 2006) | | Y | Y | Y |
| AP_T | Two-step atomic potential | Protein Docking Inc. AP_T1/2. | (Tobi, 2010) | | Y | Y | Y |
| CP_TSC | Two-step residue level contact potential | Protein Docking | (Tobi, 2010) | | Y | Y | Y |
| CP_TB | Residue level contact potential | Protein Folding | (Tobi and Bahar, 2006) | | Y | Y | Y |
| DFIRE | Atom based orientation dependent | Protein Folding | (Zhang et al., 2004) | Y | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DDFIRE | Atomic distance dependent level potential | Protein Folding | (Yang and Zhou, 2008) | Y | Y | Y | Y |
| DCOMPLEX | Atomic distance dependent Level potential | Protein Docking | (Liu et al., 2004) | Y | Y | Y | Y |
| OPUS_CA | Residue/C-Alpha distance dependent | Protein Folding | (Lu et al., 2008) | Y | Y | Y | Y |
| OPUS_PSP | Atom contact potential for Side-chain packing | Protein Folding<br>Inc. OPUS_PSP1/2/3. | (Lu et al., 2008) | Y | Y | Y | Y |
| RF_PP | Residue level potential | Protein Folding | (Rykunov and Fiser, 2010) | Y | | | |
| EMPIRE | Atomic level | Protein Docking | (Liang et al., 2007) | Y | | | |
| GEOMETRIC | Packing and distance dependent potential function | Protein Folding / Protein Interaction | Unpublished | Y | Y | Y | Y |
| CP_RMFCEN1 | Side-chain centroid distance dependent potential | Protein Folding | (Rajgaria et al., 2008) | | Y | Y | Y |
| CP_RMFCEN2 | Side-chain centroid distance dependent potential | Protein Folding | (Rajgaria et al., 2008) | | Y | Y | Y |
| CP_RMFCA | Calpha distance dependent | Protein Folding | (Rajgaria et al., 2006) | | Y | Y | Y |
| CP_SKOIP | Residue level interaction contact potential | Protein Docking | (Lu et al., 2003) | | Y | Y | Y |
| FOUR_BODY | Four-body coarse grain potential | Potentials'R'Us | (Feng et al., 2010) | Y | Y | Y | Y |
| GEN_4_BODY | Four-body coarse grain potential | Potentials'R'Us | (Feng et al., 2010) | Y | Y | Y | Y |
| SHORT_RANGE | Residue level pair potential | Potentials'R'Us | (Feng et al., 2010) | Y | Y | Y | Y |
| QA_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_Qa | (Feng et al., 2010) | Y | Y | Y | Y |
| QM_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_Qm | (Feng et al., 2010) | Y | Y | Y | Y |
| QP_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_Qp | (Feng et al., 2010) | Y | Y | Y | Y |
| HLPL_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_HLPL | (Feng et al., 2010) | Y | Y | Y | Y |
| SKOB_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_SKOb | (Feng et al., 2010) | Y | Y | Y | Y |
| SKOA_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_SKOa | (Feng et al., 2010) | Y | Y | Y | Y |
| SKJG_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_SKJG | (Feng et al., 2010) | Y | Y | Y | Y |
| MJPL_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MJPL | (Feng et al., 2010) | Y | Y | Y | Y |
| MJ3H_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MJ3h | (Feng et al., 2010) | Y | Y | Y | Y |
| MJ2H_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MJ2h | (Feng et al., 2010) | Y | Y | Y | Y |
| TS_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_Ts | (Feng et al., 2010) | Y | Y | Y | Y |
| BT_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_BT | (Feng et al., 2010) | Y | Y | Y | Y |
| BFKV_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_BFKV | (Feng et al., 2010) | Y | Y | Y | Y |
| TD_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_TD | (Feng et al., 2010) | Y | Y | Y | Y |
| TEL_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_Tel | (Feng et al., 2010) | Y | Y | Y | Y |
| TES_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_TEs | (Feng et al., 2010) | Y | Y | Y | Y |
| RO_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_RO | (Feng et al., 2010) | Y | Y | Y | Y |
| MS_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MS | (Feng et al., 2010) | Y | Y | Y | Y |
| MJ1_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MJ1 | (Feng et al., 2010) | Y | Y | Y | Y |
| MJ3_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MJ3 | (Feng et al., 2010) | Y | Y | Y | Y |
| GKS_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_GKS | (Feng et al., 2010) | Y | Y | Y | Y |
| VD_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_VD | (Feng et al., 2010) | Y | Y | Y | Y |
| MSBM_PP | Residue level pair potential | Potentials'R'Us<br>Also referred to as CP_MSBM | (Feng et al., 2010) | Y | Y | Y | Y |

| Miscellaneous | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Descriptor Type** | **Description** | **Note** | **Reference** | **ΔG** | **ΔΔG** | **$\Delta k_{off}$** | **HS** |
| DASA | Change in surface area upon binding | Naccess | | Y | | | |
| RES_P | % interface residues: polar | Naccess | | Y | | | |
| RES_NP | % interface residues: non-polar | Naccess | | Y | | | |
| RES_C | % interface residues: charged | Naccess | | Y | | | |
| ATOM_P | # interface atoms: polar | Naccess | | Y | | | |
| ATOM_NP | # interface atoms: non-polar | Naccess | | Y | | | |
| ATOM_N | # interface atoms: charged | Naccess | | Y | | | |
| PLANARITY | Interface planarity | SURFNET | (Laskowski, 1995) | Y | | | |
| ECCENTRIC | Numerical eccentricity | SURFNET | (Laskowski, 1995) | Y | | | |
| INT_ALPHA | Proportion of interface residues which are in alpha helices | DSSP | | Y | | | |
| INT_BETA | Proportion of interface residues which are in beta sheets | DSSP | | Y | | | |
| GAP_VOL | Volume of empty space at the interface | SURFNET | (Laskowski, 1995) | Y | | | |
| GAP_INDEX | Volume of empty space at the interface divided by interface Area | SURFNET | (Laskowski, 1995) | Y | | | |
| NIP | Interface packing score | | (Mitra and Pal, 2010) | Y | Y | Y | Y |
| NSC | Surface complimentarity score | | (Mitra and Pal, 2010) | Y | Y | Y | Y |
| STC_CP | Change in specific heat upon binding | STC package | (Lavigne et al., 2000) | Y | | | |
| BIOSIMZ_KON | Predicted log(kon) calculated using BioSimz | | Li,2011 | Y | | | |

## 2.3 Machine Learning Algorithms

### 2.3.1 Random Forest (RF)

The Random Forest (RF) (Breiman, 2001a) is the most commonly employed learning algorithm in this thesis. The RF is used both for problems of regression and classification and a Matlab implementation of the RF algorithm, as described by Breiman (2001a), is used. RF is an ensemble of decisions trees, where the final prediction is a majority vote (for classification) or an average (for regression) of all the trained decision trees. The 'Random' aspect of the RF algorithm is related to the way it builds each decision tree. For a training set of N samples, sampling with replacement is performed and two thirds of this sample is used as the training set for a given decision tree in the forest. The other one third (termed as the oob (out-of-bag) data, is used to get an unbiased estimate of the test error and for variable importance measures. The second randomization involved in the RF's decision trees, is that at each node, not all features are available for making a split. Rather a random sample of *mtry* features are chosen at each node and the best split is chosen amongst them. An important aspect of the RF is that the test error is reduced with more accurate and less correlated decision trees. Part of the randomization procedures employed in the tree building are in fact aimed at introducing variability in hope of achieving low correlation between decision trees. The *mtry* parameter is therefore central the RF. Given a powerful descriptor in the set of features, for high *mtry* values, it is more likely that this descriptor would be chosen in the random sample and subsequently used at the node split. Therefore this descriptor would dominate most of the trees, resulting in highly accurate trees but with low correlation. If the *mtry* parameter is set too low, then the powerful descriptor might be missed out from most of the trees. The RF would then consist of low correlation trees but with low accuracy. Though this parameter is the one for which the RF is most sensitive to, it has a broad range of optimal values (Breiman, 2001a). This was also found to be true in this work, and for most scenarios the *mtry* parameter was set to be withitn the range $\sqrt{M} < mtry < 3\sqrt{M}$ where M is the number of features available in the training set.

*RF Variable Importance Measure:* After the random forest has been built and the oob error estimate for each tree recorded, the importance of each feature to the prediction is measured as follows. For each feature *m*, all of its values are randomly permuted and the oob examples are fed through the trees with *m* randomly permuted. The importance score of feature *m* is the different between the original oob error estimates, and the new ones with *m* permuted. The importance score is then normalized by the standard deviation of these differences across all trees. Large values imply more important features. Another feature importance measure available to the RF, and invoked in this work, is the case-wise feature importance measure. Here, during permutation, the error of each oob example is recorded. In this way feature importance can also be quantified in relation to specific examples.

### 2.3.2 M5' Regression Tree (M5')

The M5 model tree is similar to standard regression trees with the additional possibility of having a linear regression model at the leaves (Quinlan, 1992). In this work, the M5′ algorithm, a modified version of the original M5 regression tree described by Wang and Witten (1996) was used. This version is able to achieve more interpretable trees through smaller trees which still have similar predictive performance. Two phases are used to build an M5' tree; the growing phase and the pruning phase. In the growing phase, a greedy algorithm is employed where at each node a split is made which minimizes the standard deviation of the examples falling on each side of the split. By the end of the growing phase, the tree is typically large and the samples partitioned by the latter splits are small in number. Therefore the error estimates for the latter splits become unreliable, and it is likely the tree overfits the data. To address this, a pruning stage is performed where a function which considers the tree size and the estimated test error is minimised. The M5' implementation in the M5PrimeLab toolbox in Matlab was used to construct one of the empirical binding free energy functions described in Chapter 3.

### 2.3.3 Multivariate-Adaptive-Regression-Splines (MARS)

MARS is a non-parametric regression method which uses a set of hinge functions to model non-linear relationships between the input variables and the target output (Friedman, 1991). The model is formed from a sum of weighted basis functions $B_i(x)$,

$$f(x) = \sum_{i=1}^{k} w_i B_i(x) \qquad 2.2$$

where each basis function contains a hinge function or a product of two or more hinge functions, if we seek to model higher order interactions between variables. The hinge function takes two forms; $max(0, m\text{-}const_{knot})$ or $max(0, const_{knot} - m)$, and is defined by some feature $m$ and a knot $const_{knot}$. Therefore in the training phase, MARS automatically assigns the weights for each basis function $w_i$, the variables for a given hinge function, and the values for the knot positions $const_{knot}$. Similar to the M5' regression tree, the MARS model also has two phases termed as the forward pass and the backward pass. In the forward pass, the basis functions are added in pairs until a stopping criterion is reached. This is usually set by the user, and can be some minimum error threshold or the maximum number of basis functions. Given that the forward phase my produce models that overfit the training data, in the backward phase, basis function are removed and model subsets are compared using a generalized cross-validation (GCV) routine. The GCV is a function of the residual sum-of-squares of the training data, the number of observations, the number of parameters and the number of knots. Therefore more flexible models, with the addition of more knots, are penalized in the backward phase. One notable aspect concerning hinge functions is that for the range in which the function is zero, the feature associated with it does not contribute to the prediction. Effectively this can be used as a mechanism to disregard noisy parts of descriptors and a higher weighting to more informative regions. The MARS implementation in the ARESLab toolbox in Matlab was used to construct one of the empirical binding free energy functions described in Chapter 3.

### 2.3.4 Radial-Basis-Function Interpolation (RBF)

RBFs are common in both artificial neural networks (ANNs) and support-vector-machine (SVM) learning algorithms. They are functions whose value depends on some distance from the origin or some point in space. The sum of a set of radial basis functions can in turn be used to approximate functions in the form of:

$$\mathrm{F}(x) = \sum_{i=1}^{N} w_i \phi(\| x - x_i \|) \qquad 2.3$$

Several distance functions may be used such as the multiquadric basis function:

$$\phi(\mathrm{d_i}) = \sqrt{d_i^2 + 1} \qquad 2.4$$

Where $d=||x-x_i||$. The fact that weights $w_i$ are learnt for examples rather than features, means that during training, uninformative examples are down-weighted and representative ones are up-weighted. The RBF implementation in Matlab was used to construct one of the empirical binding free energy functions described in Chapter 3.

### 2.3.5 Genetic Algorithm Feature Selection (GA-FS)

The GA-FS Algorithm runs feature selection on subsets of the off-rate mutation dataset defined as data regions. Two separate GA-FS runs are performed, one for Linear Regression models and another for Support Vector Machine (RBF) Regression Models (using *LIBSVM* package). Two separate 10-Fold Cross-Validation loops are used. One to assess prediction accuracy on the off-rate mutations for the given data region and the second to derive the optimal feature set. A 10-Fold inner-cross validation loop is used within the GA-FS fitness function to drive the feature selection process with Pearson's Correlation Coefficient. After the GA has converged, the LR/SVM model is tested for its accuracy on the outer-loop fold. This process is repeated 10 times such that all 10 outer loop folds are used as a test set validation for the final model. Therefore the accuracy of the final model is tested on data that is not used to derive the feature set. As an initial feature set available for selection, 110 molecular

descriptors (as shown in the $\Delta k_{off}$ column in Table 2.1) and 16 hotspot descriptors (as shown in Table 6.3) from the best performing off-rate prediction model *RFSpot_KFC2* are available. A fixed feature set size of 5 is chosen so as to avoid overfitting on smaller sized data regions. Therefore the genome size for the GS-FS (LR) is 5 whereas that for GA-FS (SVM) is 7 to also optimise the cost and gamma parameters of the RBF. Available Cost parameters values are quantized into 111 bins ranging from $2^{-5}$ to $2^{6}$. Gamma parameter values are quantized into 1300 bins ranging from $2^{-8}$ to $2^{5}$. The GA's initial population size was set at 1000 individuals, and generated such that the initial population included at least one instance of each of the 126 features. Tournament selection is employed with a size of 8 individuals. Uniform random crossover is used with a crossover fraction set to 50% and a mutation rate exponentially decreasing with the number of generations applied. Note that for each data region 50 separate GA-FS runs are performed.

### 2.3.6 Hotspot Descriptor Calculation and Dataset

As depicted in Figure 6.1, for any given complex, a computational alanine scanning is first performed on the *wild-type* interface using a hotspot prediction algorithm. This enables calculation of the set of hotspot descriptors summarized in Table 6.3 . The respective single-point or multi-point mutation is then applied using FoldX (Schymkowitz et al., 2005), and another computational alanine scan is performed on the mutated interface, again using the same hotspot prediction algorithm invoked for the *wild-type* scan, from which a new set of hotspot descriptors are calculated. The energetic value contributed by each hotspot descriptor is then the difference in its energetic value pre- and post-mutations:

$$\Delta E_{HS\_Desc} = \Delta E^{MUT}_{HS\_Desc} - \Delta E^{WT}_{HS\_Desc} \qquad 2.5$$

The hotspot descriptors are calculated for a set of 713 mutations from SKEMPI database (Moal and Fernandez-Recio, 2012) described in section 2.1.2. Therefore in total, for each hotspot prediction algorithm, 50 *wild-type* and 713 mutant computational alanine scans are made. To ensure that off-rate predictions are

not made via hotspots models trained on the same examples, all 713 computational alanine-scans made by *RFSpot, RFspot_KFC2, RFHotspoint1 and RFHotspoint2* are strictly 20-fold-test predictions for mutations common between the off-rate and hotspot datasets, and test predictions for the rest. Therefore all hotspot predictions on which the hotspot descriptors are calculated are unbiased and not susceptible to over-fitting.

### 2.3.7 Hotspot Descriptor Functional Forms and Design

The aim of the hotspot descriptors designed in this work is to capture both the energetics and distributional properties of hotspots. These in turn may affect complex destabilization to differing degrees. The relevance of each descriptor to off-rate variation is then assessed with different feature importance measures and the key determinants of the dissociation process reported.

#### *2.3.7.1 Interface Hotspot Descriptors*

*Int_Energy_1* is the difference in the sum of the single-point alanine *ΔΔG*s of all interface residues N, pre- and post-mutation.

$$Int\_Energy\_1 = \left( \sum_{n=1}^{N} \Delta\Delta G_{n \to Ala} \right)^{MUT} - \left( \sum_{n=1}^{N} \Delta\Delta G_{n \to Ala} \right)^{WT} \quad 2.6$$

*Int_HS_Energy* is the difference in the sum of the single-point alanine *ΔΔG*s of all hotspot residues $N_{HS}$, pre- and post-mutation.

$$Int\_HS\_Energy = \left( \sum_{n_{HS}=1}^{N_{HS}} \Delta\Delta G_{n_{HS} \to Ala} \right)^{MUT} - \left( \sum_{n_{HS}=1}^{N_{HS}} \Delta\Delta G_{n_{HS} \to Ala} \right)^{WT} \quad 2.7$$

*No_HS* is the change in number of hotspots predicted at the interface pre- and post-mutation. This can be considered to be a coarse-grained version of *Int_HS_Energy*.

#### *2.3.7.2 Solvent Accessible Region Hotspot Descriptors*

To account for the different solvent accessible regions where hotspots may occur at the interface, the following hotspot $\Delta\Delta G$s are summed separately for the core, rim and support regions and termed as *CoreHSEnergy*, *RimHSEnergy* and *SuppHSEnergy* respectively. Therefore these hotspot descriptors are similar to *Int_HS_Energy* but limited to counting $\Delta\Delta G$ for hotspots that fall in the given region. In addition, *CoreHS*, *RimHS* and *SuppHS* descriptors, count the hotspot changes within each region. Again these can be considered as coarse-grained versions of their respective counterparts. The core, rim and support regions of the complex interface are defined according to Levy (2010). Core residues are generally exposed in the unbound configuration but buried in the bound state. Rim regions are generally exposed in both the bound and unbound states whereas support residues are generally buried in both states. The thresholds chosen in defining these regions are such that each region has a similar number of residues (Levy, 2010).

#### *2.3.7.3 Hotregion Cooperativity Descriptors*

The cooperativity of a pair of residues m1 and $m_2$, can be calculated by comparing the gain of adding each residue separately from a neutral reference state of both *wild-type* residues mutated to alanine ($\Delta\Delta G_{A1,A2->A1,m2} + \Delta\Delta G_{A1,A2->m1,A2)}$ to that of adding both residues concurrently, given the same reference state ($\Delta\Delta G_{A1,A2->m1,m2}$) (Albeck et al., 2000). Namely, let $A_1$ and $A_2$ represent the alanine mutation of $m_1$ and $m_2$ respectively, then

$$\Delta\Delta\Delta G = (\Delta\Delta G_{A1,A2->A1,m2} + \Delta\Delta G_{A1,A2->m1,A2}) - \Delta\Delta G_{A1,A2->m1,m2} \quad 2.8$$

If $\Delta\Delta\Delta G$ is positive, this indicates positive cooperativity as the contribution of both residues together is more stabilizing than the sum of their parts. Conversely if the $\Delta\Delta\Delta G$ is negative, this indicates negative cooperativity, whereas if the $\Delta\Delta\Delta G$ is close to zero, then such pairs can be considered to be effectively independent of each other hence their contributions to be additive in relation to each other.

Expanding $\Delta\Delta G_{A1,A2->A1,m2}$ and $\Delta\Delta G_{A1,A2->m1,A2}$ we get

$$\Delta\Delta\Delta G = ([\Delta\Delta G_{m1,m2->A1,m2} - \Delta\Delta G_{m1,m2->A1,A2}] + [\Delta\Delta G_{m1,m2->m1,A2} - \Delta\Delta G_{m1,m2->A1,A2}]) - \Delta\Delta G_{A1,A2->m1,m2} \quad 2.9$$

$$\Delta\Delta\Delta G = (\Delta\Delta G_{m1,m2->A1,m2} + \Delta\Delta G_{m1,m2->m1,A2}) - \Delta\Delta G_{m1,m2->A1,A2} \quad 2.10$$

In this work, we only make single point-mutations during the alanine scan and calculate the energetics associated with such complex states as in equation 2.10: $\Delta\Delta G_{m1,m2->A1,m2}$ and $\Delta\Delta G_{m1,m2->m1,A2}$. The summation of these energies is then used as an estimate of the off-rate. If hotspots within a cluster are additive, then the summation of $\Delta\Delta G_{m1,m2->A1,m2} + \Delta\Delta G_{m1,m2->m1,A2}$ would be a sufficient estimate of the cluster's contribution to the off-rate. However if m1 and m2 are positively cooperative, then their contribution towards the off-rate using the summation $\Delta\Delta G_{m1,m2->A1,m2} + \Delta\Delta G_{m1,m2->m1,A2}$ would be an overestimate of the true contribution $\Delta\Delta G_{m1,m2->A1,A2}$, hence the positive value for $\Delta\Delta\Delta G$. Therefore in this case, to account for positive cooperativity we down-weight the summation of $\Delta\Delta G_{m1,m2->A1,m2} + \Delta\Delta G_{m1,m2->m1,A2}$. Conversely if m1 and m2 were negatively cooperative, then a positive weighting would be more suitable to account for the underestimation. Further, higher order cooperativity effects involving three or more residues are known (Albeck et al., 2000) and it is likely that many binding modules exhibit such complexity, where it is not possible to decouple the contributions from each individual residues. However, if we assume that cooperativity effects are taking place, the weighting applied should also reflect the number of residues suspected to be cooperative. With this in mind, the cooperativity hotspot descriptors are designed as follows; given a set of predicted hotspots at the interface, each hotspot is categorized according to the hotregion cluster size it is found in. As *Int_HS_Energy* assumes hotspot contribution is additive, the sum of the hotspot energies is independent of the hotspot locations (equation 2.7). On the other hand, *HSEner_PosCoop* and *HSEner_NegCoop* are the sum of the hotspot energies downweighted / upweighted using simple linearly decreasing / increasing functions related to the size of the hotregion the given respective hotspot is in:

$$HSEner\_PosCoop = \left( \sum_{n_{HS}=1}^{N} w_{HR}^{Dec} \times \Delta\Delta G_{n_{HS} \to Ala} \right)^{MUT} - \left( \sum_{n_{HS}=1}^{N} w_{HR}^{Dec} \times \Delta\Delta G_{n_{HS} \to Ala} \right)^{WT} \quad 2.11$$

$$HSEner\_NegCoop = \left( \sum_{n_{HS}=1}^{N} w_{HR}^{Inc} \times \Delta\Delta G_{n_{HS} \to Ala} \right)^{MUT} - \left( \sum_{n_{HS}=1}^{N} w_{HR}^{Inc} \times \Delta\Delta G_{n_{HS} \to Ala} \right)^{WT} \quad 2.12$$

where $w_{HR}^{Dec}$ = (0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1) and $w_{HR}^{Inc}$ = (1, 0.875, 0.75, 0.625, 0.5, 0.375, 0.25, 0.125) for hotspot $n_{HS}$ in a hotregion of sizes (HR=1, 2, 3, 4, 5, 6, 7, 8+) respectively. Though more complex non-linear weightings could be investigated, such as ones fitted to the off-rate data itself, this would require sacrificing parts of the data for fitting. With this in mind, all hotspot descriptors designed in this work were independent of any off-rate data. Coarse-grained versions *HS_PosCoop* and *HS_NegCoop*, which weight hotspot counts instead of energies, are also implemented in the model. One should note that since the energetic contribution of a hotregion taken as a whole is considered to be additive and independent of other hotregions (Keskin et al., 2005, Reichmann et al., 2005) we only aim to investigate and account for intra-hotregion cooperativity using these descriptors as opposed to inter-hotregion cooperativity.

#### *2.3.7.4 Hotspot Coverage Related Descriptors*

Other hotspot descriptors relate to the spread of hotspots across the interface. The intuition here is that a heterogeneous distribution of hotspots across the interface might be more beneficial to complex stability than if hotspots where concentrated onto a specific region of the interface only. *AVG_HS_PathLength* is the average path length between all possible pairs of hotspots at the interface, normalized to the average path length of all possible pairs of a random set of residues at the interface. The path length between two residues is calculated as the least number of contacting residues linking them together. Two residues are considered to be in contact if any of their atoms are at a distance smaller than the sum of their van der Waals radii + 0.5 Angstroms. *No_Clusters* counts the number of unique hot regions, where it is likely that more hotregions may span the

interface given that separate hotregions are not in contact. *MaxClusterSize* counts the change in the number of hotspots in the largest hotregion.

#### *2.3.7.5 Definition of a Hotregion*

Some of the hotspot descriptors use hotregion information within them (*No_Clusters*, *MaxClusterSize*, *HSEner_PosCoop/HS_PosCoop* and *HSEner_NegCoop/HS_NegCoop*). A hotregion is created whenever two or more hotspot residues are in contact. Two hotspot residues are considered to be in contact if any of their atoms are at a distance smaller than the sum of their van der Waals radii + 0.5Å. A hotspot residue is added to an existing hotregion, if any of its atoms makes contact with any of the hotspot residues already in the hotregion.

## 2.4 Performance Measures and Significance Tests

A number of performance measures are employed in this work to assess the fine-grained and coarse-grained ability of both descriptors and model predictions. For fine-grained assessment of how well a descriptor or model predictions describe experimental data, the Pearson's product-moment correlation coefficient (PCC) is used. This is calculated as the covariance of the two variables divided by the product of their standard deviation. This parametric measure of correlation assesses the strength of linear dependence between two variables and is a widely accepted metric. A second method employed is the Mann-Whitney U-test. This checks whether a set of two independent observations have smaller or larger values than the other. The test is used to assess the coarse-grain predictive power of our descriptors or predictors in discriminating between say stabilizing mutants from destabilizing mutations. Several other classification related measures are used for this same purpose also, namely:

*True-Positive-Rate (TPR) / Recall:*

$$\frac{TP}{TP + FN}$$

*False-Positive-Rate (FPR):*

$$\frac{FP}{FP+TN}$$

*Specificity:*

$$\frac{TN}{TN+FP}$$

*Precision:*

$$\frac{TP}{TP+FP}$$

*Accuracy:*

$$\frac{TP+TN}{TP+FP+FN+TN}$$

*Matthew's Correlation Coefficient (MCC):*

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

*F1-Score:*

$$\frac{2\times precision\times recall}{precision+recall}$$

where TP=True-Positive, FP=False-Positive, TN=True-Negative, FN=False-Negative.

For comparison of two PCCs, say for the comparison of two prediction algorithms, a significance of the difference between the two correlations is calculated using the Fisher r-to-z transformation. p-values less than 0.05 are considered to be significant.

# Chapter 3

# 3 A Model for Protein-Protein Binding Affinity Prediction

## 3.1 Introduction

In this chapter, the characterization and prediction of protein-protein affinities is studied. The computational prediction of binding affinities requires not only an understanding of the driving forces behind complex formation and stability, but also an accurate computational representation of such forces. Subsequently, a model is then employed to optimally combine the influence of each of these driving forces into one coherent prediction of affinity. Throughout this process, a benchmark set of protein-protein X-ray structures and their experimentally determined binding affinities, is used to validate the accuracy of the model predictions.

As detailed in section 1.5.7, up until this work, several attempts at the prediction of binding affinities were made. The limitations of these models are highlighted

in sections 1.5.7.1, 1.5.7.2 and 1.5.7.3 form the motivations behind the binding affinity prediction model devised in this work. First, the dataset of protein-protein structures and their experimental affinities used to benchmark the affinity model is described in section 3.3.1. A large set of molecular descriptors calculated on these structures is detailed section 3.3.2. This descriptor set significantly expands on what was used in previously published affinity models. Namely, besides a number of physical descriptors, it adds a broad range of statistical potentials, new solvation models and better entropic terms. The molecular descriptors are then fed into a number of machine learning models (described in 3.3.3) that are combined to make the final prediction. This is termed as the consensus model. The setup used is that of four base learning models: random forest (RF) regression, M5' regression, multivariate-adaptive-regression-splines (MARS) and radial-basis-function (RBF) interpolation, with the mean of their prediction constituting the final affinity prediction model. The motivations behind these learners are mostly based on the limitations surrounding linear regression in modelling, such as accounting for non-linearities and dealing with a large set of noisy descriptors.

The consensus model approach is successful in increasing upon the accuracy of its best base learners and more importantly, outperforms all other published methods tested. Two aspects that stood out from this work include firstly, the limitations in our ability to predict the affinities for complexes which undergo significant conformational changes and secondly, the reduction in accuracy observed when the errors in experimental binding affinities are not controlled for with a validated set of complexes. Finally, in section 3.5, the binding affinity methods developed post-publication of this work are discussed and suggestions for future work outlined.

This work was done in collaboration with my colleague Iain Moal. The selection and calculation of the molecular descriptor set was performed by Iain Moal. The machine learning algorithm selection and design was performed by myself. The analysis of the results was performed jointly.

## 3.2 Approach and Motivations

The limitations mentioned in section 1.5.7.1– bias towards complexes for which their component parts undergo little to no conformational change, section 1.5.7.2– static-structures and section 1.5.7.3– models not able to account for diversity, form the motivations behind the BAP model developed in this work. The following methods section 3.3 detailing the dataset, descriptors and learning models, shows how the limitations mentioned above are addressed. In section 3.3.1 the dataset of protein-protein complexes is described. This consists of a diverse set of protein-protein interactions with varying conformational changes. In section 3.3.2 the affinity descriptors used in this work are presented. These include energetic descriptors calculated on conformational ensembles of each complex and their unbound components. Finally, in section 3.3.3, the machine learning models used for training and prediction are detailed.

## 3.3 Methods

In section 3.3.1 the dataset of protein-protein complexes is described. This consists of a diverse set of protein-protein interactions with varying conformational changes. In section 3.3.2 the affinity descriptors used in this work are presented. These include energetic descriptors calculated on conformational ensembles of each complex and their unbound components. Finally, in section 3.3.3, the machine learning models used for training and prediction are detailed.

### 3.3.1 Binding Affinity Benchmark Dataset

The structures and experimental affinities for the recently published binding affinity benchmark (Kastritis et al., 2011)were used as the main source for training and testing the BAP model designed in this work. The dataset is described more thoroughly in section 2.1.1. In summary, the dataset consists of a total of 144 complex structures for which the crystal structures of each complex, along with each of its unbound components at high resolution (< 3.25Å), are

available. A key aspect of this dataset, which improves upon previous datasets used in BAP (see section 1.5.7), is its diversity:

- Several receptor/ligand protein-binding partners undergo significant conformational change upon complex formation, of which some exhibit disorder to order transitions.
- Complexes, within different protein families, cover a wide range of functions.
- Wide range of affinities.

From the total set of 144 complexes, 137 were used. The complexes with protein database codes, 1UUG, 1IQD and 1NSN, were removed, as affinities available were only denoted by upper limits; codes, 1DE4, 1M10, 1NCA and 1NB5 were removed, as certain features were difficult to calculate for them

A high-quality validated subset of the original affinity dataset is analysed separately. This validated set is used to assess to which extent experimental error in affinities affects the model predictions. One should note that in this validated set, the diversity in affinity and complex families is still present. More details on this validated set are presented in methods section 2.1.1.1.

### 3.3.2 You are what you eat.. Affinity Descriptors

In collaboration with Iain Moal, a large set of 200 molecular descriptors were calculated on the binding affinity benchmark and fed into the machine learning models described in section 3.3.3. A detailed list of the descriptors is provided in Table 2.1 of the methods section 2.2. The descriptor set covers a wide-range of known determinants of complex formation and affinity. The set contains different contributors to the free energy function described in section 1.5, and includes descriptors related to the potential energy, solvation energy and entropic contributions to binding. In addition to the biophysical descriptors, a number of statistical potentials are added, which vary from pair to multi-body potentials and contain both coarse-grain and atomistic potentials. Though some specific packages are used for a number of descriptors, most were calculated

using the ProtorP server (Reynolds et al., 2009), CHARMM forcefield (Brooks et al., 2009), PyRosetta (Chaudhury et al., 2010) and the Potentials 'R' Us server (Feng et al., 2010). For the assumption that binding is rigid-body, and structures are static, descriptors were calculated as:

$$E = E_{RL,b} - (E_{R,b} + E_{\mathrm{L,b}}) \qquad 3.1$$

The motivations behind the descriptors calculated here are several fold. Firstly, most of the descriptors are directly related to known physical contributors of affinity, including terms for Hydrogen Bonding, Van der Waals and Electrostatics. Emphasis was given to entropy related terms, as this effect is harder to characterise. Therefore entropic terms include rotational, translation and side-chain entropy terms, vibrational and disorder loop entropy terms along with terms for the hydrophobic effect. Solvation is another important aspect modelled at different levels of sophistication. Here, both simple terms related to buried surface area and more sophisticated continuum electrostatics models are included. Different to other BAP models, in this work we do not limit ourselves to physic-based descriptors only. A number of statistical potentials and miscellaneous descriptors are also added to the descriptor set. The advantage of statistical potentials is that they implicitly capture a number of effects that are otherwise only modelled individually using physics-based terms. As pH can have a significant effect on binding affinity, even over a narrow range, some descriptors were chosen for their ability to account for variable protonation states. PROPKA was used to determine the pH of the titratable amino acids (Bas et al., 2008). The most probable assignment of protonation states, at the experimental pH, was determined using PDB2PQR (Dolinsky et al., 2004). These assignments were used in all of the descriptors calculated using the CHARM22 forcefield, which are prefixed with ACE22.

The two major introductions in the BAP model of this work relate to structural ensembles and unbound structures. These are described in the following sections of 3.3.2.1 and 3.3.2.2 respectively.

#### 3.3.2.1 Unbound-Bound Descriptors

To account for potential conformational changes, descriptors were also calculated on the unbound receptor and ligand. Descriptor calculations ignored residues which were not in both the bound and unbound structures. In this way, any energetic differences in the two conformational states, is irrespective of additional residues in the bound. All descriptors calculated on the unbound structures have a suffix of '_UB' and are calculated as:

$$E_{UB} = (E_{R,b} - E_{R,\mathrm{u}}) + (E_{\mathrm{L},b} - E_{\mathrm{L,u}}) \qquad 3.2$$

#### 3.3.2.2 Descriptor Ensembles

Proteins both in their unbound and bound forms do not exist as static structures. Rather they exist in a number of conformations of varying energetic accessibility (See section 1.5.2). As seen in equation in 1.17, $\langle U_{RL} \rangle$ and $\langle W_{RL} \rangle$ are Boltzmann-average potentials for the potential energy and solvation energy respectively. Also, equation 1.13 shows that only the low energy conformations contribute significantly to the potential energy. Therefore sampling only the low energy conformations provides a sufficient approximation. To generate such conformational ensembles, the use of an approximate method CONCOORD (de Groot et al., 1997) was preferred to complex molecular dynamics simulations, mostly due to computational efficiency. Unlike MD trajectories, the conformations generated by CONCOORD have no dependencies on previous conformations; consequently, the conformational space is sampled more broadly. Comparisons of CONCOORD simulations against MD simulations on common structures show great overlap in both the accessible motions and their magnitude (de Groot et al., 1997).

For each example in the benchmark, an ensemble of 100 structures was generated using CONCOORD with dynamic tolerance setting. This, for each ligand, receptor and complex. Descriptors are then calculated on these

ensembles and given that CONCOORD generates structures of equal plausibility, a mean value is taken over all ensembles for each descriptor. To distinguish them from descriptors calculated on a single static structure, the ensemble calculations have a '_ENS' suffix and are calculated as:

$$E_{ENS} = \langle E_{RL,b} \rangle - (\langle E_{R,b} \rangle + \langle E_{\mathrm{L,b}} \rangle) \quad 3.3$$

For those descriptor calculations where the ensembles were calculated on the unbound ligand and receptor, a suffix of '_EBU' was used, and calculated as:

$$E_{EBU} = (\langle E_{R,b} \rangle - \langle E_{R,\mathrm{u}} \rangle) + (\langle E_{\mathrm{L},b} \rangle - \langle E_{\mathrm{L,u}} \rangle) \quad 3.4$$

### 3.3.3 Machine Learning Methods

As highlighted in section 1.5.7.3 models for BAP have previously been limited to a sum of terms with the weights of each optimised using linear regression. Here, a selection of four machine learning methods was combined to form a consensus prediction, with the consensus prediction being the mean prediction of the four base models. It should be noted that more complex forms of ensemble learning are indeed possible (Wolpert, 1992). For example one may have a meta-learner learn weights for each base-learner according to the input example at hand; however attractive, such methods would require a further validation set which is not available in this case. The four base models are the Random Forest (RF), the M5' Regression Tree, the Multivariate-Adaptive-Regression-Splines (MARS) and the Radial Basis Function Interpolation (RBF) each of which are describe in the methods section 0. The aim was not to use the learning models in a black-box fashion but rather the selection of models was guided by the following considerations:

- *Addressing limitations of linear regression.* LR has been routinely applied to protein-protein affinity prediction methods. The ML algorithms selected here aim to address some of the limitations LR would reach in our dataset and feature set. These include; inability to account for non-

linear relationships; inability to partition the input space and apply sub-models; degradation in performance in high-dimensions.

- *Differing conceptual attributes.* The prediction of each of the four ML models is combined to form a final prediction which is the mean of the four models. This is similar to a stacked learning methodology in its most basic form. Ideally for effective stacking, the base learning models should be accurate but show weak correlation between their predictions (Wolpert, 1992). The combination of all four base models would then work synergistically rather than redundantly. To try and achieve this, the learners were chosen on the basis of having different conceptual attributes in how they form their model. For instance, the RF is derived from the consensus of tree models trained on variable subsets of data and features. The M5' method on the other hand is built using one complex tree model with the added flexibility of applying further regression sub-models within the tree itself. Using its hinge functions, the MARS model works by allowing certain descriptors to contribute within certain ranges and not others. Therefore, it can exploit the 'predictive' parts of a descriptor and avoid the 'noisy' parts. Moreover, all of the methods above base their final prediction on a selection of features, rather than the whole available set. Therefore, depending on the final features selected by the model, this is likely to add some variability in their predictions. The variability in the features making it to each of the final models is confirmed in the results section 3.4.5. Finally, the RBF method works in a completely opposite fashion to the other three models. Here, the emphasis is placed on particular data-points that are furthest from the current data-point. Therefore, the RBF uses all descriptors but not all examples in its final model.

- *Overfitting avoidance.* Given the large set of descriptors available to the models and the limited size of the training set, overfitting can be an issue. To avoid this, the methods chosen either implicitly or explicitly avoid overfitting. RFs do not overfit as more trees are added. Rather, the test error converges to a limiting value (Breiman, 2001a). They are able to

achieve low bias predictions through trees built from different subsets of the data and descriptors, and low variance through averaging the output of all trees. The M5' and MARS learners have inbuilt backward elimination routines to reduce model complexity, by the removal of tree branches and basis functions respectively. Consequently, both of these operations reduce the number of features in the final model. In the RBF learner, the feature weights are not optimised. With this in mind, an outer-cross validation loop is still performed for benchmarking the predictions of each model.

- *Parameter optimization.* To avoid having to sacrifice data for parameter optimization, all methods chosen are known to work well under their default parameters settings. No tweaking of learning parameters was therefore performed.

- *Interpretability and visibility.* Understanding how the features are used in the final model was a key consideration for selecting the learners. Besides forming an accurate predictor of binding affinity, it is also important to ascertain the physical plausibility of the final models, by knowing which features are essential to the prediction and how they are employed by the model itself. With the RF model, both global feature importance and case-wise feature importance measures are available. The case-wise feature importance measure is particularly desirable. For example, one hopes that having descriptors which calculate energetics on the unbound structures of the complex would help the affinity prediction of complexes which undergo significant conformational changes. Invoking the case-wise feature importance measure one could verify this specifically by checking whether the features on unbound structures are shown as being important for those cases which undergo significant conformational change. The M5' regression trees used, lack an inbuilt feature importance measure; however, the trees can be easily visualized and feature importance was still evaluated. The nature of the MARS model basis functions, not only indicates which features form part of the final model, but also the functions applied to each of these features. Effectively the

function shows us the parts where the given feature has little influence, and where it positively contributes to the prediction.

#### *3.3.3.1 Random Forest (RF)*

A Matlab implementation of the RF algorithm, as described by (Breiman, 2001a), was used. The workings of the RF algorithm are detailed in section 2.3.1. In this implementation, the number of decision trees was set to 750 and, when building the decision trees, the *mtry* parameter was limited to 20 at each node; no maximum was set on the three depths and the final prediction is returned as the mean of all trees.

#### *3.3.3.2 M5' Regression Tree (M5')*

The M5' model tree is similar to standard regression trees with the additional possibility of having a linear regression model at the leaves (Quinlan, 1992). The workings of the M5' algorithm are detailed in section 2.3.2. Rather than applying one M5′ to the full feature set, an ensemble of M5′ regression trees was used. In total 16 M5' regression trees were divided into four tree sets of four. For each tree set, all features are divided randomly into four feature subsets. Each different random feature subset is then used to train each of the four trees within this tree set. Therefore, for a given tree set, all features are available for use, but for each tree within the tree-set, a random subset of features is available. For prediction, the mean output of all of the 16 trees is used.

#### *3.3.3.3 Multivariate-Adaptive-Regression-Splines (MARS)*

MARS is a non-parametric regression method which uses a set of hinge functions to model non-linear relationships between the input variables and the target output (Friedman, 1991). Default values were used without tuning, as follows: the maximum limit on the number of basis functions grown in the forward phase is 21, there was no limit on the number of basis functions used in the final model after pruning. Model complexity is also limited by setting the knot-cost to the recommended value of two. Piece-wise cubic modelling was used to model hinge regions for smoother transitions. To keep the model as interpretable as possible,

no self-interactions between input variables and no interactions between variables in the basis functions were allowed. The ARESLab toolbox implementation was used.

#### *3.3.3.4 Radial-Basis-Function Interpolation (RBF)*

A Matlab implementation of the RBF method, as section 2.3.4 was used. All descriptors values were normalized in the range [0, 1] before training. The key parameter in the RBF is the choice of the basis function. For this, the default multiquadric basis function was used. A unique characteristic for the RBF is that the model finds weights for examples as opposed to features. Therefore in this way, uninformative examples as opposed to uninformative features are weighted out of the model.

### 3.3.4 Model Evaluation

To assess our ability to model and predict binding affinities, leave-one-out cross-validation (LOO-CV) was employed and the predicted affinities were compared to the experimental affinities using Pearson's product-moment correlation coefficient. To establish significant differences in correlations achieved by different models, a Fisher r to z transformation of the correlation coefficients was used.

## 3.4 Results

### 3.4.1 Model Performance on the Binding Affinity Benchmark – Validated Set

Initially the four base learners (MARS, M5', RF and RBF) were trained and tested using leave-one-out cross-validation on the validated set. The performance of which is shown in Figure 3.1 alongside that of the Consensus model (Cons.), which combines the prediction of the four base learners by taking the mean of their predictions.

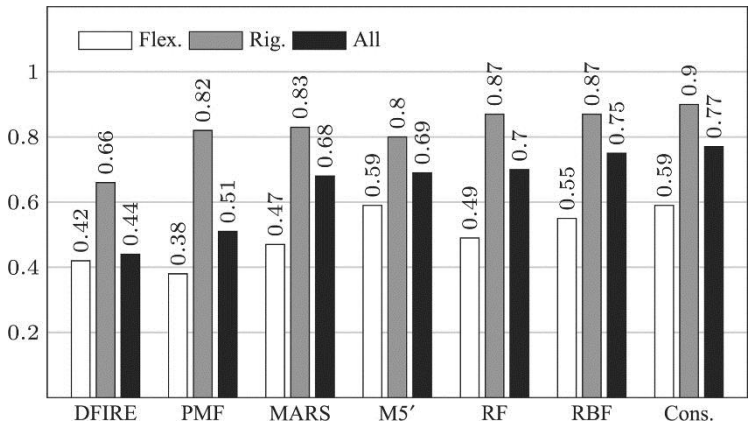



**Figure 3.1: Model performance for the 57 complexes in the validated set**. Correlation between the experimental and predicted binding affinities for the learners and their consensus, using leave-one-out cross-validation. The potentials of Liu et al. (2004) (DFIRE) and Su et al. (2009) (PMF) are also shown for comparison.

As a benchmark comparison, the performance of DFIRE (Su et al., 2009) and PMF (Liu et al., 2004) are also shown. To assess the effect of conformational changes on the prediction accuracy, performance is separately tested for cases which are rigid (Rig. with $C_{alpha}$ RMSD < 1.5 Å) and flexible (Flex. with $C_{alpha}$ RMSD > 1.5 Å). The consensus model achieves a correlation of $R_{VAL}$=0.77 with experimental affinity, which is significantly higher than that achieved by the potentials PMF ($R_{VAL}$=0.51 p=0.012) and DFIRE ($R_{VAL}$=0.44 p=0.003).

### 3.4.2 Model Performance on Binding Affinity Benchmark – Entire Dataset

The learners presented in Figure 3.1 were also evaluated on the remaining complexes that are not part of the validated set. To observe the performance over the complete dataset, the learners were trained on all 137 complexes, and the leave-one-out cross validated predictions of the non-validated complexes amalgamated with those of the validated set in Section 3.4.1. The correlations of the learners and experimental affinities, in a similar fashion to Figure 3.1, are presented in Figure 3.2.

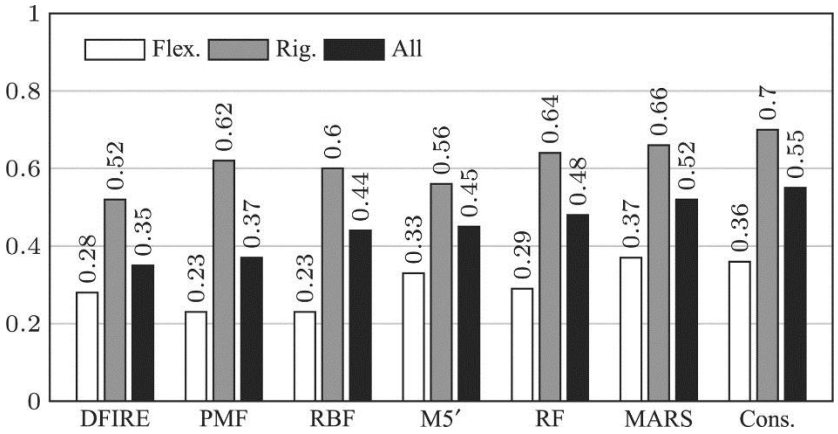



**Figure 3.2: Model performance for the 137 complexes in the whole benchmark.**

Correlation between the experimental and predicted binding affinities for the learners and their consensus. The potentials of Liu et al. (2004) (DFIRE) and Su et al. (2009) (PMF) are also shown for comparison.

Though, in comparison to the results on the validated set, the relative performance of the four base leaners changed, the consensus model still performs better than the most accurate base learner. In addition, the consensus model achieves significantly higher correlations ($R_{ALL}$=0.7) to that of DFIRE ($R_{ALL}$=0.52, p=0.02) and PMF ($R_{ALL}$=0.62,p=0.03).

Comparison of Figure 3.1 and Figure 3.2 indicates a clear drop in predictive power across all methods as experimental affinities that are not validated are used. One should note that, this drop is despite the fact that the validated set still has a proportion of non-rigid cases and interaction types similar to that of the entire set (see section 2.1.1). These results provide strong evidence to the importance of having affinity data that is corroborated using different experimental techniques. To remove any possibility that this drop in accuracy is model dependant, a number of methods are tested on the 37 complexes for which predictions are available for all methods, and presented in Figure 3.3.

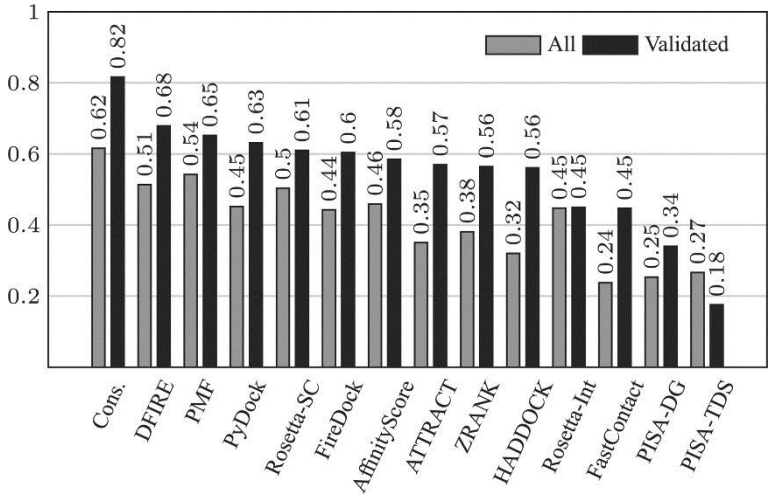



**Figure 3.3: Performance of the consensus model on the 37 complexes in the intersection between the dataset of (Kastritis and Bonvin, 2010) and the benchmark (All), and the 14 in the intersection with the validated set (Validated).**
Leave-one-out cross-validation is used for the interactions which intersect the validated set. Correlations for a number of other energy functions are also shown (see Section 3.4).

Comparing the performance of each method on all 37 complexes, and the 14 of which are validated, a consistent trend is observed were methods tend to perform better on the validated set. Once again, here it is shown how the consensus model is still the best performer, even on these specific test subsets.

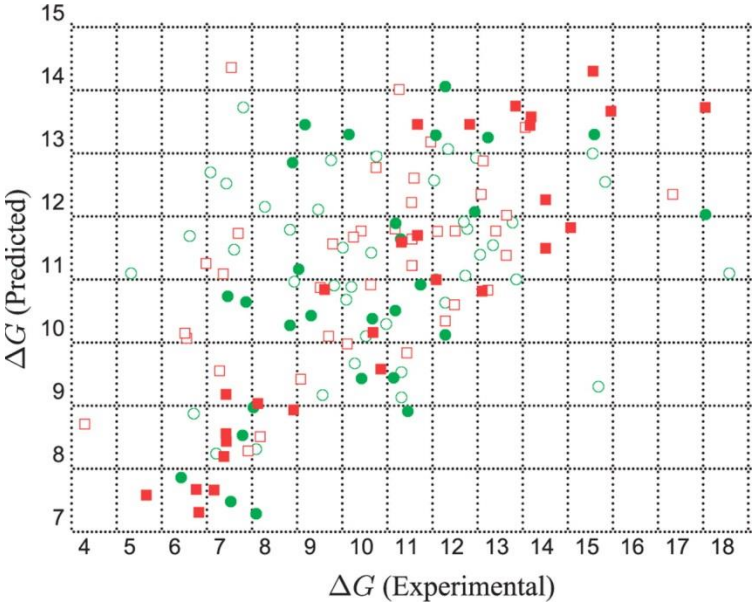



**Figure 3.4: Scatter plot for predicted and experimental affinities**.
Flexible (green circles) and rigid (red squares) proteins are shown. Leave-one-out cross-validated values for the validated set are highlighted in solid.

Figure 3.4 shows the scatter plot of predicted and experimental affinities were the validated complexes are marked in solid. There are two notable features, the first being the lower spread of points for the validated set. The second, that the dense upper left corner indicates that for those cases, the affinity is overestimated. Given that most of these data-points are flexible complexes, the entropy loss due to conformational change is not characterised well enough to balance out the enthalpic contribution towards binding affinity.

### 3.4.3 Consensus Model vs. a Single Learning Algorithm.

The consensus model in all three data types (All validated $R_{VAL}$=0.77, rigid $R_{VAL-RIG}$=0.9 and flexible $R_{VAL-FLEX}$=0.59) achieves a correlation, which is higher or as good as the best base learner in the set. For example, M5' achieves the highest correlation of $R_{VAL-RIG}$=0.59 on the flexible cases, but one of the poorest in predicting the rigid cases. In the latter case, the RBF achieves the highest correlation ($R_{VAL-RIG}$=0.87) of all base learners. The consensus model is able to take the best of both worlds by achieving the highest correlations in both of these situations. This confirms that the four base models are working

synergistically together and taking the mean of their predictions is a valid approach. The correlations between the predictions of each of the four base learners, is also evaluated. As expected, the tree-based learners (RF and M5') are highly correlated with R=0.95. The RBF method shows a correlation of R=0.87 and R=0.86 with the RF and M5' learners. The MARS model showed the least correlation with the other methods (R=0.65, R=0.69, R=0.68, respectively). Though this may suggest that the MARS is picking out features that the other learners are not, one must also keep in mind that the MARS model was the weakest learner of all.

### 3.4.4 Descriptors Derived from Unbound Structures, Improves Performance for Flexible Cases.

A key element of the BAP method developed in this work, as described in section 3.3.2, is the introduction of ensembles and unbound structures. To determine the gain in having energetics calculated on the unbound and non-static structures as part of the model, the consensus model is trained on specific feature subsets. These are: the UnBound (UB) subset; features calculated on the unbound structures, the ENSemble subset (ENS); features calculated using the CONCOORD ensembles of the bound components, the Ensemble Bound/Unbound EBU: features calculated using the ensembles of bound and unbound structures and BASIC in which neither ensembles nor unbound structures are considered. The LOO-CV correlation achieved by training the consensus model on just the BASIC descriptors is used as a reference point to assess gain or loss in predictive power by adding the UB, ENS, EBU features to this BASIC subset.

**Table 3.1: Performance of the consensus model trained on different feature subsets.**

The UB subset: features calculated on the unbound structures, ENS: features calculated using the CONCOORD ensembles of the bound components, EBU: features calculated using the ensembles of the bound and unbound structures and BASIC in which neither ensembles nor unbound structures are considered. The correlation achieved by training the consensus model on just the BASIC descriptors is used as a reference point to assess what is gained by adding the UB, ENS, EBU features to this BASIC subset. All correlations shown are those between the LOO-CV predictions with the experimental affinities.

| Feature Subset | All | Rigid | Flexible |
|---|---|---|---|
| **BASIC** | 0.67 | 0.91 | 0.44 |
| **BASIC+ENS** | 0.69 | 0.85 | 0.45 |
| **BASIC+UB** | 0.74 | 0.91 | 0.47 |
| **BASIC+EBU** | 0.73 | 0.90 | 0.54 |
| **ALL** | 0.77 | 0.90 | 0.59 |

The results are summarized in Table 3.1. First, it is noted that the addition of the unbound descriptors, both on the unbound static structures (UB) and on the unbound ensemble structures (EBU), increases the correlations over the BASIC model. BASIC $R_{VAL}$=0.67, whereas BASIC+UB and BASIC+EBU models achieve correlations of $R_{VAL}$=0.74 and $R_{VAL}$=0.73 respectively. This increase in the overall correlation results from the additional accuracy in predicting the flexible cases (from $R_{VAL\text{-}FLEX}$=0.44 to $R_{VAL\text{-}FLEX}$=0.47 and $R_{VAL\text{-}FLEX}$=0.54 for BASIC, BASIC+UB, BASIC+EBU respectively). In fact, the prediction of rigid cases remained constant at around $R_{VAL\text{-}RIG}$=0.9. One should note however that the increase in correlation with the addition of the unbound descriptors is mostly evident when ensembles were calculated on the unbound structures. Conversely, the addition of bound ensembles (with no consideration of the unbound structures) to the BASIC set has, as expected, no effect on the prediction of flexible cases. An interesting result is that addition of ensembles actually degrades the signal for rigid cases (BASIC $R_{VAL\text{-}RIG}$=0.91 and BASIC+ENS $R_{VAL\text{-}RIG}$=0.85). This may be explained by some conformational ensembles generated, not being representative of those accessible by the rigid complex in question. It may be the case that, for these rigid structures, more flexibility than is energetically accessible is being

generated by CONCOORD. This translates itself as noise added to the true signal to be captured.

From these results, it can be concluded that the inclusion of descriptors derived from the unbound structures improves the performance for the flexible complexes, without compromising the accuracy for the rigid cases. This improvement is enhanced when used in combination with structural ensembles, despite the ensembles not enhancing the consensus model when information derived from the unbound structures is omitted. These results should still be treated with caution, as the increases/decreases in correlation are not statistically significant with $p<0.05$ as the number of data-points is restricted to the 57 complexes in the validated set. Therefore, when the data allows, the same analysis must be performed again on a larger dataset to confirm the claims above. With this in mind, one complex, which for example shows clear improvement in the prediction of its affinity upon the inclusion of unbound descriptors, is the interaction between MK2 and p36 MAPK (PDB code, 2OZA). MK2 undergoes a significant disorder-order transition upon binding, and the strongest within the dataset. In this case, when training the consensus model on the BASIC set of features (i.e. not including unbound-bound transitions), the predicted affinity (17.4kcal $mol^{-1}$) overestimates the experimental affinity of (11.7kcal $mol^{-1}$). Once descriptors on the unbound were calculated, the learners are able to make use of available descriptors that calculate the entropy changes due to disorder-order transitions, and the predicted affinity (10.9kcal $mol^{-1}$) achieved was a closer approximation to the experimental affinity.

### 3.4.5 Learning from the Learners – Assessment of the Physical Plausibility of the Learning Models and the Key Determinants of Affinity.

One of the driving forces behind the selection of the base learners for the consensus model is the interpretability of the models. In this section, the learnt models from each of the base learners, is probed further for validation of their selected features.

*RF base learner*: Both the global features importance and the case-wise feature importance measures are invoked for the RF learner which was trained on the full set of descriptors and validated set of affinities.

**Table 3.2: Top 10 most important descriptors using for the RF base learner trained on the validated set.**
Feature importance in this case is the mean decrease in normalised mean square error generated from the RF learner.

| Rank | Descriptor | Descriptor Importance |
|---|---|---|
| **1** | ACE19_HYDR | 0.100 |
| **2** | ROS_FA_ATR | 0.094 |
| **3** | ACE22_VDW | 0.094 |
| **4** | ROS_HBOND_ENS | 0.078 |
| **5** | DDFIRE_ENS | 0.076 |
| **6** | S_VIB | 0.063 |
| **7** | MJ2H_PP | 0.049 |
| **8** | ROS_FA_ATR_ENS | 0.047 |
| **9** | MJ1_PP | 0.046 |
| **10** | H_BOND_ENS | 0.044 |

The top 10 most important features making up the RF model include a combination of thermodynamic terms, statistical potentials and miscellaneous descriptors. The most prominent being hydrophobic burial (ACE12_HYDR), London dispersion forces (ROS_FA_ATR), Van der Waals (ACE22_VDW) and hydrogen bonding (ROS_HBOND_ENS). Also ranked highly are the change in vibrational entropy (S_VIB) and a number of statistical potentials (DDFIRE_ENS, MJ2H_PP and MJ1_PP). This confirms the physical plausibility of the model as it includes terms related to the potential and solvation energy and also those related to entropic contributions (See section 1.5).

From the top 10 descriptors, four terms are calculated on structural ensembles, but no descriptors using the unbound structures are listed. In section 3.4.4 it was shown that the introduction of UB and EBU descriptors improves the prediction of the flexible cases, the case-wise feature importance measure of the RF was invoked in order to understand whether the UB/ EBU descriptors were at least being invoked for the flexible cases. Here, a feature calculated using the unbound structure appeared as one of these top 5 features for 16 of the 29 flexible

complexes (55%). This compares to only 3 of the 28 rigid complexes (11%); this indicates, that to some extent, the learnt model is making correct use of the UB and EBU descriptors for the complexes that should gain from it.

*M5' Base Learner*: The full descriptor set was assigned randomly to the four sub-trees within a tree set. A descriptor can therefore be in the final model of only one of the four sub-trees in a tree-set. This means that at most, a given descriptor can show up four times in the whole set of 16 M5' trees. Each of the M5' sub-tree models was analysed, its features extracted and their occurrence summed in Table 3.3.

**Table 3.3: Top 10 most important descriptors using for the M5' base learner trained on the validated set.**
Descriptor importance refers to the number of times a descriptor is part of a sub-tree. The maximum of which is four.

| Rank | Descriptor | Descriptor Importance |
|---|---|---|
| **1** | NSC | 4 |
| **2** | OPUS PSP ENS | 4 |
| **3** | ROS CG BETA | 4 |
| **4** | ROS FA ATR | 4 |
| **5** | BIOSIMZ KON | 3 |
| **6** | DDFIRE ENS | 3 |
| **7** | GEOMETRIC EBU | 3 |
| **8** | H BOND | 3 |
| **9** | INTERNAL UB | 3 |
| **10** | NUM HB | 3 |
| **11** | PLANARITY | 3 |
| **12** | ROS FA REP ENS | 3 |
| **13** | S R | 3 |
| **14** | SKJG PP | 3 |
| **15** | STC G ENS | 3 |

It is interesting to note that even though both RF and M5' are tree based algorithms, only a few descriptors such as DDFIRE_ENS, ROS_FA_ATR and H_BOND are common between them in the set of most important features. Similar to the RF, the most important descriptors in the case of the M5' trees are a combination of thermodynamic terms and statistical potentials. Even though a

number of entropic terms are available to the learning models, besides the change in vibrational entropy (S_VIB) and change in rotational entropy (S_R), entropic terms are not as common in the top features of the RF and M5' models.

*MARS base learner*: The MARS model trained on the validated set terminates with 14 basis functions using a total of 10 descriptors. The descriptors are ranked according to their global importance to the model and presented in Table 3.4.

**Table 3.4: Top 10 most important descriptors using for the MARS base learner trained on the validated set.**
Descriptor importance is ranked according to the standard deviation (STD). As stated by Friedman (1991), the STD gives an indication to the relative importance of the descriptors to the overall model, and is similar to a standardized regression coefficient in a linear model. Shown also are the generalized-cross-validation (GCV) scores. This represents the decrease in GCV upon removal of the descriptor. Lastly, #basis indicates the number of basis function the descriptor is part of.

| Rank | Descriptor | STD / GCV/ #basis |
|---|---|---|
| **1** | ROS FA ATR | 0.618 / 0.690 / 2 |
| **2** | S_VIB | 0.456 / 0.230 / 1 |
| **3** | ROS HBOND UB | 0.365 / 0.071 / 2 |
| **4** | OPUS CA UB | 0.364 / 0.065 / 2 |
| **5** | IRMSD | 0.259 / 0.097 / 1 |
| **6** | OPUS PSP | 0.191 / 0.048 / 1 |
| **7** | ROS HBOND | 0.176 / 0.069 / 1 |
| **8** | SKOA PP | 0.174 / 0.052 / 2 |
| **9** | ACE19 SOLV | 0.172 / 0.062 / 1 |
| **10** | ROS FA PP EBU | 0.169 / 0.060 / 1 |

Similar to the M5' and RF models, the London dispersion term ROS_FA_ATR is the most prominent descriptor followed by the vibrational entropy term S_VIB. Other descriptors include solvation terms, hydrogen bonding and statistical potentials. Most significant here is that the MARS model makes use of a number of descriptors on the unbound structures (ROS_HBOND_UB, OPUS_CA_UB, ROS_FA_PP_EBU, IRMSD). A key aspect of the MARS model is that it is able to assign a variable weight for each descriptor across its range. Effectively, it can choose to ignore the 'noisy' parts of a region of a descriptor by assigning a zero weight within that region. It then provides a weighting to more informative

regions of the descriptor. Such weights are presented and explained in Figure 3.5.

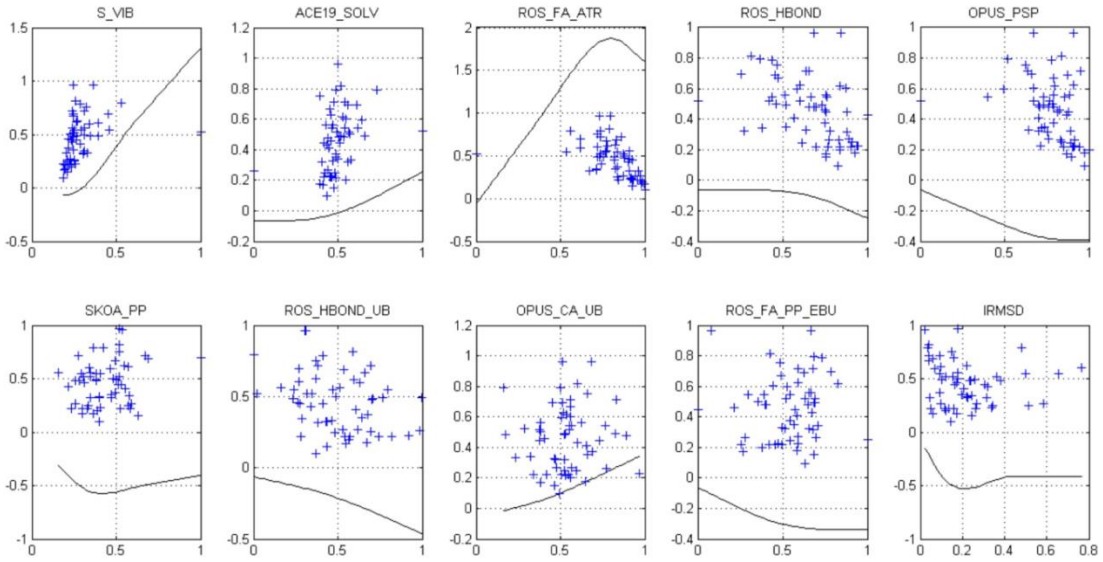



**Figure 3.5: Descriptor contribution profiles for the descriptors selected by MARS.**

Normalised descriptor values are on the (x-axis) and normalised affiniies are on the (y-axis). The normalisation is such that 0 is the lowest affinity (ΔG = -5.66kcal $mol^{-1}$) in the dataset and postively higher values indicate an increase in affinity (e.g at 0.53 the ΔG=-12.28kcal $mol^{-1}$ and at 1 the ΔG=-18.04 kcal $mol^{-1}$) The '+' plots show the experimental normalised affinities. The line graphs show the contribution towards affinity from the basis functions of the given descriptor.

For most of the data, ROS_FA_ATR's contribution to the binding affinity linearly increases with more favourable dispersion forces (the normalisation in ROS_FA_ATR is such that 0 is highly negative in energy). However, a hinge function models the outlier 2OZA to have a lower affinity then one would expect with its highly favourable dispersion forces (owing to its large interface). The role of the hinge function is to compensate for the entropy reduction resulting from the disorder to order transitions occurring in a loop and at the C-terminal region of 2OZA. The second most significant descriptor is the vibrational entropy term S_VIB. At low values, its contribution is approximately zero, but becomes linear for higher values. This is consistent with the interpretation that, because this descriptor is approximate (Carrington and Mancera, 2004), the learner is presumably choosing to use it when its contribution to the binding energy is sufficient to outweigh the noise it introduces. This for example cannot be

performed with linear regression and such a descriptor might be completely down weighted and not form any significant contribution to the final model.

*RBF base learner*: The RBF model does not learn weights on descriptors, rather it down weights and up weights examples according to how much they help the prediction of other examples. The prediction function of the RBF model takes the form of

$$F(x) = \mu + \sum_{i=1}^{N} a_i \phi(\| \mathrm{x} - \mathrm{x}_i \|) \qquad 3.5$$

Where *u* is the mean of the training affinities. The affinity prediction of a complex *x*, is lower than the mean affinity if $a_i$ is negative, and higher if $a_i$ is positive. The magnitude of this deviation from the mean depends on how many complexes are furthest from it in feature space. Conversely, the closer it is to complexes in the training set, the closer is its predicted affinity to the mean. The weights $a_i$ learnt by the RBF function trained on the validated set are presented in Figure 3.6. High affinity complexes tend to have a negative weight $a_i$. This means that if a complex affinity is to be predicted, the further in feature space it is from high affinity complexes, the lower from the mean its affinity would be. Effectively this translates to a low affinity prediction. In this way, the model is a plausible and the coefficients learnt are in line with what one would expect to see. One may also appreciate that certain complexes have an $a_i$ very close to zero and have no significant contribution in the final RBF model.

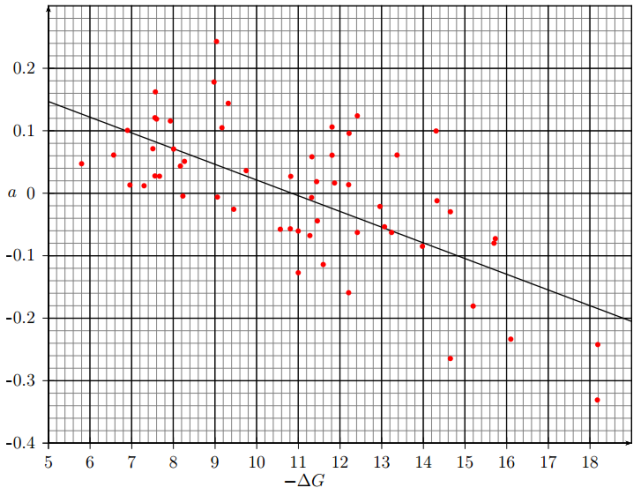



**Figure 3.6: The distribution of regression coefficients learnt by the RBF model versus binding affinity.**
Negative coefficients weaken the predicted affinity of complexes far away in feature space. Positive coefficients strengthen the predicted affinity of complexes far away in feature space.

## 3.5 Discussion

This work on binding affinity prediction is a first in many ways. Previous to the described consensus model, the datasets used to train and benchmark models did not include a diverse enough set of complex families with a wide range of conformational changes. There were no such limitations to this study. Namely; both the unbound and bound structures were considered; energetics were also calculated on a set of sampled conformational states; and complexes from a variety of complex families, undergoing a broad range of conformational changes included. The use of non-linear machine learning methods for the prediction of affinities was also put forward as an alternative to linear regression. The consensus model, which is the mean prediction of four machine learning algorithms, achieves a correlation with experimental binding affinities of R=0.77 on a validated set of experimental affinities. The consensus model outperforms other previously published methods both on this dataset and when tested on different subsets of the dataset.

*Two major influences on the prediction of affinities*. The results from this work uncovered two major concerns. Firstly, there is still a large discrepancy in our

ability to accurately predict affinities for flexible cases (those undergoing significant conformational change) as opposed to rigid cases. To address this, several descriptors related to entropy were included, along with calculations of descriptors on the unbound complexes. Such additions specifically increased the accuracy of flexible complexes, yet correlations are still in the region 0.6 with experimental affinities. On the other hand, for rigid cases ($R_{VAL\text{-}RIG}$=0.9), the RMSE of 1.67 kcal $mol^{-1}$ is within the variation expected due to experimental errors and unaccounted environment factors, around 1.4 kcal $mol^{-1}$ (Kastritis et al., 2011). A second finding, and that which had a striking effect on the prediction of affinities, is the training and benchmarking of models on affinities corroborated by different experimental methods or studies; one should be critical of generalizations concerning the importance of descriptors when used in models trained on non-validated affinities as the descriptors in such models might be only acting as noise-compensators rather than their intended purpose. Linear regression models are even more susceptible to using descriptors in this unintended way.

*Consistent determinants of affinity.* In this work, four machine-learning algorithms, with different conceptual attributes, were used. Their prediction outputs were then combined in a consensus model. Analysis of the most important descriptors for each model showed that certain descriptors were common to all models. This inspite of the ML algorithms having significantly different methods in the way they build their models. Such consistent descriptors include; London dispersion forces (ROS_FA_ATR), several hydrogen bonding terms, and the change in vibrational entropy (S_VIB). The two most common statistical potentials were the DDFIRE and OPUS PSP. For the most part however, descriptors which were important to each model for its prediction were not shared across models. Again, this should be taken into consideration when making outright claims on the importance of a descriptor from the feature importance list of only one learning model.

*Binding affinity prediction methods published after this work*. Subsequent to the work described in this chapter, three notable BAP methods were published (Vreven et al., 2012, Yan et al., 2013, Ma et al., 2014), all of which use the same benchmark set of complexes used in this work (Kastritis et al., 2011). In the work of Vreven et al. (2012), a linear combination of nine terms (ZAPP) achieves a correlation coefficient of $R_{ALL}$=0.63 with the whole benchmark of experimental affinities. This is in comparison to $R_{ALL}$=0.55 that was achieved in this work, on the same full benchmark. Given the reduction in accuracy reported in this thesis work when complexes with non-validated experimental affinities are used, the subset of predictions from ZAPP which form part of the validated set were extracted. The correlation for the whole validated set, rigid and flexible subsets respectively, were ($R_{VAL}$=0.72, $R_{VAL\text{-}RIG}$=0.78, $R_{VAL\text{-}FLEX}$=0.65). This in comparison to the consensus model of this thesis with ($R_{VAL}$=0.77, $R_{VAL\text{-}RIG}$=0.9, $R_{VAL\text{-}FLEX}$=0.59). The higher accuracy ($R_{ALL}$=0.55) achieved on the whole benchmark by ZAPP, is most likely due to the fact that most of the terms in their final model are residue-based, which as they claim, introduce less-noise than atomic-based terms. With this in mind, once the validated set is considered, ZAPP performs worse than the consensus model reported in this work, most notably for the rigid cases. One should also note that in ZAPP, no unbound structures are used for the calculation of descriptors. Yet still, the correlation on the flexible cases is consistent and better than the consensus model. Interestingly, the only feature which may account for the flexible cases in ZAPP is 'MisRes'. This feature counts the number of residues in the interface that are present in the bound, but not in the unbound form (Vreven et al., 2012). One point that is not clear in the ZAPP method is how the final set of nine features are chosen in the final model. The authors' state that they were chosen from a larger set of features, yet details on any separate dataset for this feature selection is not given. Therefore, one could not rule out biased results. A second work which attempts BAP on the same benchmark set of structures, is the scoring function SPA-PP (Yan et al., 2013). Here, a statistical potential which incorporates both the specificity and affinity of an interaction into the optimisation is developed. Comparison is only provided on the whole benchmark where SPA-PP achieves ($R_{ALL}$=0.39, $R_{ALL\text{-}RIG}$=0.63, $R_{ALL\text{-}FLEX}$=0.24), compared to ($R_{ALL}$=0.55, $R_{ALL\text{-}RIG}$=0.70, $R_{ALL\text{-}FLEX}$=0.36) for the

consensus model developed in this chapter. The third BAP method developed after the publication of the consensus model was that of Ma et al. (2014). A RF learner is trained on a set of 154 features and the method benchmarked on a test set of 31 samples, which is a subset of the affinity benchmark. The authors' correlation on this 31 sample test set is $R_{SAMPLE}$=0.91, compared to $R_{SAMPLE}$=0.89 for the consensus model and $R_{SAMPLE}$=0.88 for the ZAPP method (Vreven et al., 2012). On this test set, there are no significant differences between the three methods. An interesting aspect of the work of Ma et al. (2014), is the attempt to introduce categorical variables which indicate what type of complex family the interaction is part of. In theory, this could lead the algorithm to apply different models according to the complex family. However, the authors failed to note that RF would only make a split and apply separate models if the immediate split decreases the MSE. To achieve the intended goal of the authors, one would need look-ahead-regression models, or explicitly separate models for each complex family type.

*Future directions.* The prediction of protein-protein binding affinities would benefit from the derivation of features, which are able to accurately characterise the affinity of complexes that undergo significant conformational changes and account for entropic contributions. As starting points, it has been shown in this work how the vibrational entropy term and the inclusion of unbound structures into the modelling process, improve the prediction of these flexible cases. Similarly the 'MisRes' feature developed in ZAPP (Vreven et al., 2012) has an equally contributing effect. It might be the case that the descriptors for rigid complexes and flexible complexes are incompatible within the same model. To rule out this effect, models specifically trained on flexible cases only should be investigated. This principle could also be applied to complexes derived from different functional families. As stated by Wallqvist et al. (1995), though the free energy change of binding has many known contributions, it is not always possible to invoke any general statements about the relative importance of each of these terms, as diverse arrangements occur that can contradict any attempted generalizations. For example, though on average the protein interface is more hydrophobic (Cherfils et al., 1991), the distribution of the composition of the

hydrophobic residues at an interface reveals a larger variability than in the interior of proteins (Tsai et al., 1997). In addition, there are many examples of complexes whose interfaces are largely hydrophilic in nature (Xu et al., 1997). This provides two possible routes for future BAP methods; The first one, having family-specific models for complexes of different biological function that are further partitioned according to the extent of predicted conformational change upon complex formation. The second and alternative route is that of having one learning model and with carefully designed descriptors which together are able to account for such diversity. Therefore further investigations on ML models and categorical descriptors of this sort I feel is a fruitful pursuit in this regard.

# Chapter 4

# 4 Models for Predicting Changes in Binding Affinity upon Mutation

## 4.1 Introduction

Having an accurate binding affinity predictor is straightforwardly generalizable to the prediction of changes in binding affinity upon mutation. Rather than predicting the affinity of a number of unrelated complexes, the problem shifts to that of predicting the affinity of a single complex and a number of mutations applied to it. Effectively, models for the prediction of *wild-type* binding affinities, as presented in Chapter 3, are a precursor to affinity optimisation in computational drug design methods (Fleishman et al., 2011).

This chapter specifically deals with the design of a scoring function which is able to rank order tentative mutations on a computationally designed interface. Such a scoring function must be able to detect those few rare mutations that are able to enhance the affinity of the designed interaction, even further. For practical purposes, the detection of these beneficiary mutations must also be accompanied by a low number of false-positives. Both the design and benchmarking of the scoring function developed in this work, is benchmarked in a special round of CAPRI (Round 26), on two targets T55 and T56, described below in section 4.2.

Traditionally, CAPRI is a blind-trial community-wide benchmark for docking algorithms. In this round however, the structures of the target complexes were given, and the task for the community was to predict the effect on binding for each mutation in a large set of around 1800 single-point mutations on two hemagglutinin influenza protein binding constructs. CAPRI round 26 was divided into two phases, and the prediction models designed for each are detailed in sections 4.3.1 and 4.4.1 respectively. In sections 4.3.2 and 4.4.2, the results for our laboratory's model predictions and all the participant groups are presented for round 1 and round 2 respectively. Throughout both rounds, most particularly in the detection of the rare beneficiary mutations, the model derived in this work was one of the top performing predictors (Moretti et al., 2013). The novelties, merits and shortcomings of this approach are discussed in sections 4.3.2, 4.4.2 and 4.4.3.

## 4.2 CAPRI Round 26 Targets T55 and T56: Blind Trial Prediction of Mutations on *de novo* Protein Drugs to Bind the Flu Virus Hemagglutinin.

In (Fleishman et al., 2011), two computationally designed hemagglutinin influenza protein construct binders HB36.4 and HB80.3 (See Figure 4.1) were used as a starting point for a large set of single-point mutants to further enhance the affinity of their interaction. For each of the 53 and 45 positions of HB36.4 (T55) and HB80.3 (T56), single-point mutations were created to all 20 amino acids. An experimental enrichment value, used as a proxy for ΔΔG, was measured for each of the mutations as described in (Moretti et al., 2013, Whitehead et al., 2012).

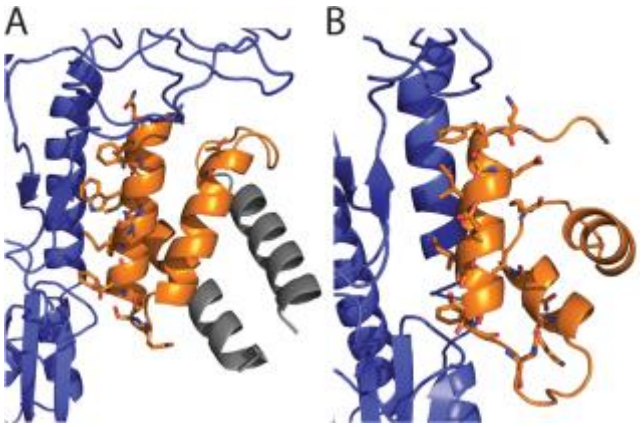



**Figure 4.1: The structures of (A) HB36 (B) HB80 in complex with the flu virus hemagglutinin.**
Residues for which experimental enrichment values were available and given to the community are in orange; the remainder are in grey. Interface residues are shown as sticks. As observed, mutations on these residues may also affect binding by affecting monomer stability. Figure taken from (Moretti et al., 2013).

The participants were asked to rank the mutations according to how beneficiary they are to the stability of the complex on an arbitrary scale of 0-1. In addition, it was also necessary to assign a class to each mutation (beneficial / neutral / deleterious). As starting structures, HB36.3 (PDB code, 3R2X) and HB80.4 (PDB code, 4EEF) were provided to the community. The difference between HB36.3 and HB36.4, and their respective *wild-type* structures, on which experimental mutations were made and measured, is a K64N mutation for HB80.4, and the mutations G12K, L17I, L21I, A35K and S42K for HB80.3. Two phases were set for the community. In the first, participants had to make predictions on 1007 and 856 mutations for T55 and T56 respectively. In the second, the enrichment ratios of half of the mutations for each residue site mutated were given to the community to train on. In this way it could be evaluated to which extent the prediction is enhanced upon having some mutational data on the complexes in question.

## 4.3 CAPRI Round 26 Targets T55 and T56: Round 1

### 4.3.1 Dataset / Molecular Descriptors / Learning Model and Training Results

The problem of predicting the enrichment ratio for the single-point mutants defined in section 4.2 was treated as a ΔΔG prediction problem i.e. predicting the change in binding affinity upon mutation. The methodology employed was similar to the one used in Chapter 3. The major difference is that whereas for the *wild-type* ΔG prediction in Chapter 3, calculations on molecular descriptors took the form of

$$\Delta G = \text{Complex} - (\text{Receptor} + \text{Ligand}) \quad (4.1)$$

Here, for the change in binding affinity, the following equation was used

$$\Delta\Delta G = [\text{Complex} - (\text{Receptor} + \text{Ligand})]_{MUT} - [\text{Complex} - (\text{Receptor} + \text{Ligand})]_{WT} \quad (4.2)$$

*Training dataset:* A training set of mutations was first compiled from the literature. This amounted to a dataset of 645 single-point/multi-point mutations (with experimentally measured ΔΔGs) on 40 protein-protein complexes determined by X-ray crystallography.

*Molecular feature set*. Taking note of the successful features (a combination of physics-based, statistical potentials and miscellaneous descriptors) for the consensus *wild-type* binding affinity models described in Chapter 3, a series of features were similarly calculated. These include a number of energy terms from FoldX (Carra et al., 2012, Schymkowitz et al., 2005), CHARMM (Brooks et al., 2009), PyRosetta (Chaudhury et al., 2010) and STC (Lavigne et al., 2000). These include all of the standard thermodynamic equations, including solvation, electrostatics and entropy terms. In addition a number of statistical potentials were also calculated DFIRE/DCOMPLEX (Zhang et al., 2004), OPUS-PSP (Lu et al., 2008), GOAP (Zhou and Skolnick, 2011) , GEOMTRIC and DECK (Liu and Vakser, 2011). Finally, miscellaneous descriptors such as the change in solvent accessibility and change in normalised solvent accessibility, residue conservation, Interface Packing (NIP) and Surface Complementarity (NSC) terms

were also calculated (Mitra and Pal, 2010). Since monomer stability may also effect binding indirectly, some of the statistical potentials included are originally folding potentials. In addition I-Mutant 2.0 (Capriotti et al., 2005), a predictor of changes in protein stability upon mutation, was also added to the feature set. Unbound structures of the complex were not considered in the calculations; however, 100 structural ensembles were generated using CONCOORD (de Groot et al., 1997), and feature calculations averaged over these structures.

*Model training results:* The RF regression algorithm was employed for learning and prediction. The number of forest trees was set to 1000 and the *mtry* parameter to 15. To assess the generalization ability of the RF model, in advance of predictions for round 1, a 10-fold cross-validation was performed. A correlation coefficient of R=0.86 between experimental and predicted ΔΔGs was achieved, the scatter plot of which is shown in the left hand panel of Figure 4.2.

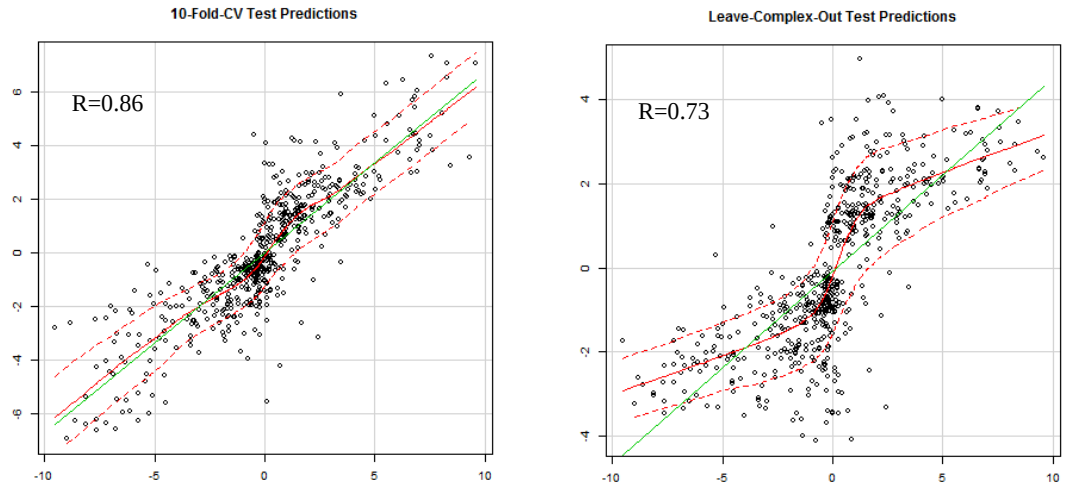



**Figure 4.2: Cross-validated test predictions for the RF model on 645 experimental single-point and multi-point ΔΔGs for 40 protein-protein complexes**.
Experimental ΔΔG values are shown on the x-axis and predicted ΔΔG values on the y-axis. 10-fold cross-validation is shown in left panel. To gain a more representative generalization ability of the blind prediction on T55 and T56, a leave-complex-out cross-validation was also performed (right panel). Here all mutations of a particular complex are left out as a test set in each fold. This has the effect of underestimating the magnitude of affinity increasing/decreasing mutations.

Leave-complex out cross validation, where a whole complex along with its respective mutations is taken out for each fold, was also performed. The correlation coefficient in this scenario dropped to R=0.73 (Figure 4.2 right panel), but gives a better indication of the model's generalization ability when making predictions on unseen complexes such as those for T55 and T56.

### 4.3.2 Affinity Prediction Results on T55 and T56

The Capri organisation committee required participants to submit both numerical and categorical predictions for the 1007 and 856 single-point mutations on T55 and T56 respectively. For the numerical predictions, the RF predictions were scaled to [0, 1] with the ΔΔG=0 neutral point at 0.8246. Using the unscaled prediction from the RF, the thresholds for destabilizing, neutral and stabilizing mutations were set at $\Delta\Delta G > 1$ kcal $mol^{-1}$, 0 kcal $mol^{-1} < \Delta\Delta G < 1$ kcal $mol^{-1}$ and $\Delta\Delta G < 0$ kcal $mol^{-1}$ respectively. After submission of all participant predictions, results for the continuous predictions were evaluated using the Kendall tau-b correlation to the log2 (enrichment ratio) values. In this metric, all possible pairs of predictions are evaluated as concordant (e.g. enrichment 1 > enrichment 2 and prediction 1 > prediction 2) or discordant (e.g. enrichment 1 > enrichment 2 and prediction 1 < prediction 2). Then

Tau-b = (concordant pairs – discordant pairs) / $N_x N_p$,

Where $N_x$ and $N_p$ are the number of total pairs not tied on experimental and predicted values, respectively. For categorical prediction, the F1-score was used (see methods section 2.4.

Figure 4.3 and Figure 4.4 show the results for the 22 participant groups. The reference BLOSUM62 prediction is highlighted in blue, and our group (G05s) is highlighted in green. For numerical predictions, our model performs worse than BLOSUM 62. Numerical predictive performance is largely affected by how well the destabilizing mutations are ranked against each other. For computational design purposes however, the interest lies more in the ability to detect those few mutations that are beneficial (Moretti et al., 2013), irrespective of how well destabilizing mutations are ranked against each other. For HB36 (T55),

beneficial mutations amount to only 3.4% of the substitutions, and only 2.4% for HB80 (T56) (Moretti et al., 2013). In contrast to the performance on numerical predictions, our RF model excels at the categorical detection of beneficial mutations. For T55, our group ranked 1st from the 22 groups (F1-Score=0.34), and 6$^{th}$ on T56 (F1-Score=0.14). From 22 groups, only our model and two other were able to achieve precisions better than 10% for both proteins (Moretti et al., 2013).

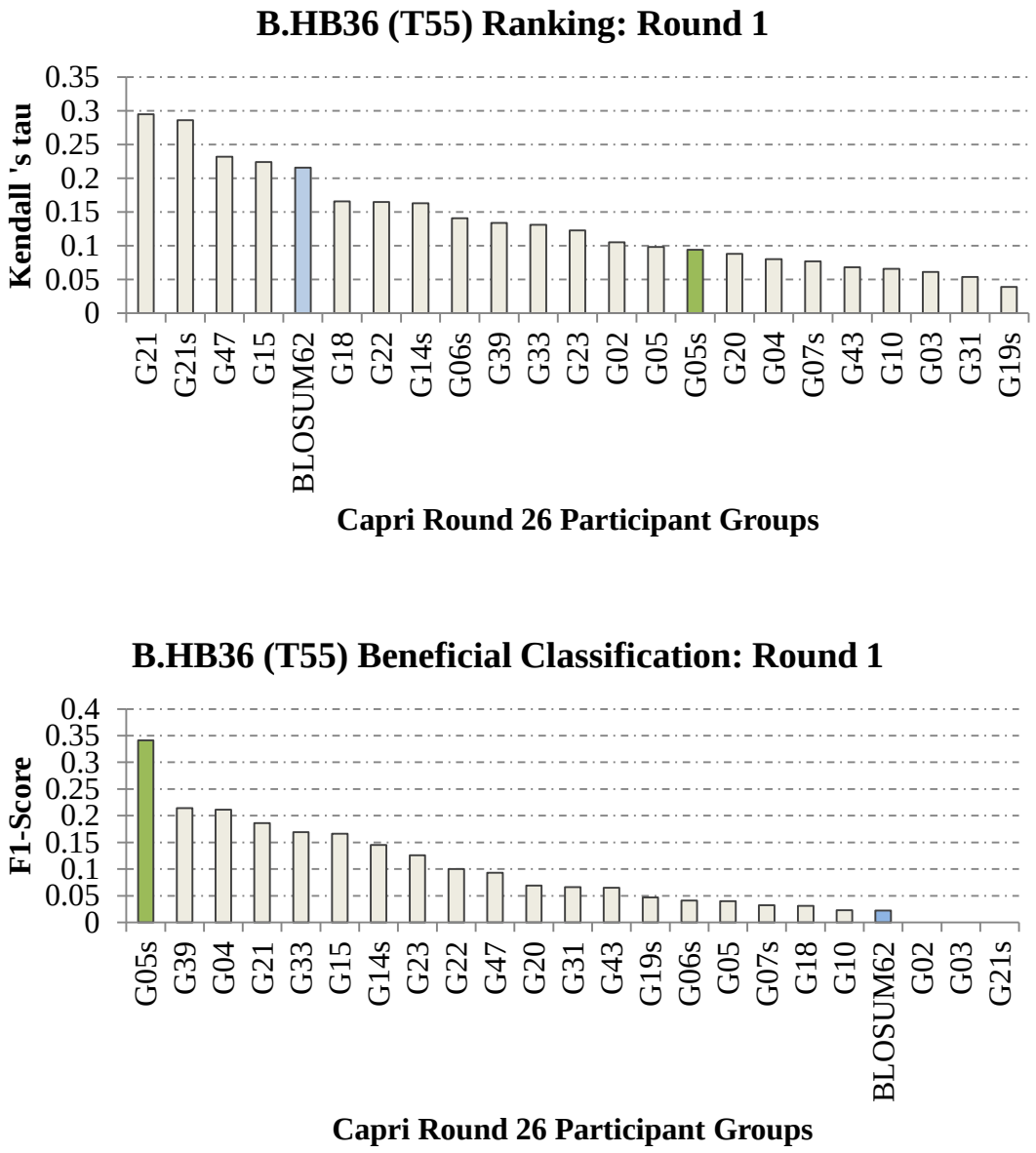



**Figure 4.3: CAPRI 26, target T55 round 1, prediction performance of all participant groups.**

A number of residue sites on T55 are mutated to each of the 20 amino-acid residues and an experimental enrichment ratio (proxy to the ΔΔG) is measured for each. This amounts to 1007 mutations on T55, for which the participants were asked to submit their blind predictions. Predictions were submitted in numerical form, scaled in the range of [0,1] according to how beneficial the mutation is thought to be. The groups' numerical predictions are ranked according to the Kendall's tau (top panel). Predictions were also submitted in categories (Deleterious/Neutral/Beneficial) and the F1-Score used for ranking the performance of the beneficial mutation detection (bottom panel). In blue are the reference BLOSUM62 predictions and in green (group G05s) the prediction of our group.

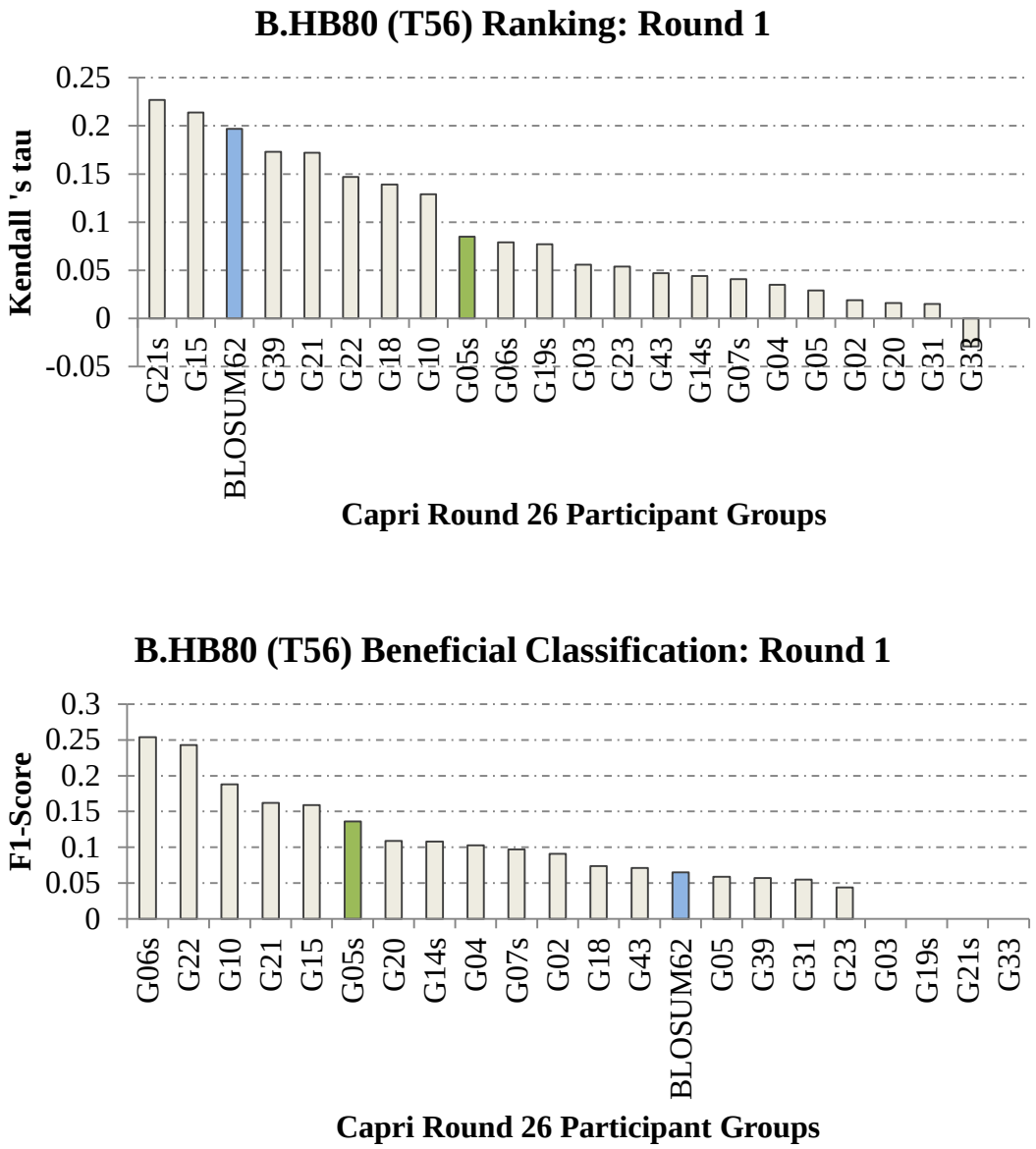



**Figure 4.4**: **CAPRI 26, target T56 round 1, prediction performance of all participant groups.**

A number of residue sites on T56 are mutated to each of the 20 amino-acid residues and an experimental enrichment ratio (proxy to the ΔΔG) is measured for each. This amounts to 856 mutations on T56, for which the participants were asked to submit their blind predictions. Predictions were submitted in numerical form, scaled in the range of [0,1] according to how beneficial the mutation is thought to be. The groups' numerical predictions are ranked according to the Kendall's tau (top panel). Predictions were also submitted in categories (Deleterious/Neutral/Beneficial) and the F1-Score used for ranking the performance of the beneficial mutation detection (bottom panel). In blue are the reference BLOSUM62 predictions and in green (group G05s) the prediction of our group.

After careful analysis of the results, two things in our methodology might have negatively affected our predictions. Firstly, our descriptor calculations were not performed on the exact structures used to derive the experimental enrichment ratios from the mutations. T55 differed by one mutation to the structure our calculations were made on, however T56 differed by 5 mutations. Given that each residue site was mutated to every other possible residue, then 1.9% (19 of 1007 mutations) and 11% (95 of 855 mutations) of mutation predictions, were affected for T55 and T56 respectively. In addition to this, the results for neighbouring residues which were also mutated are also affected. Given that >95% of all residues of T55 and T56 were mutated, it is reasonable to assume that for each mutated residue site, there are at least 3 other residues sites in contact with it which are also mutated (and hence required predictions for). This potentially increases the amount of mutations affected for T55 and T56 to ~6% and ~30% respectively. Hence, whereas for T55 the predictions are largely unaffected, the results for T56 must be treated with caution. This also should explain why the performance on T56 was markedly lower than that of T55. A second aspect of our methodology which could have negatively affected the results, was the use of the coarse-grain conformational ensembles generated using the program CONCOORD. The subtle effect of the single-point mutations for which predictions were to be made, was more than likely subdued by the conformations generated by CONCOORD. In fact redoing the methodology without the CONCOORD structures revealed that this was a significant source of error. With this in mind the use of such ensembles was avoided in round 2, however structures used remained the same and thus predictions for T56 remained severely constrained even in round 2.

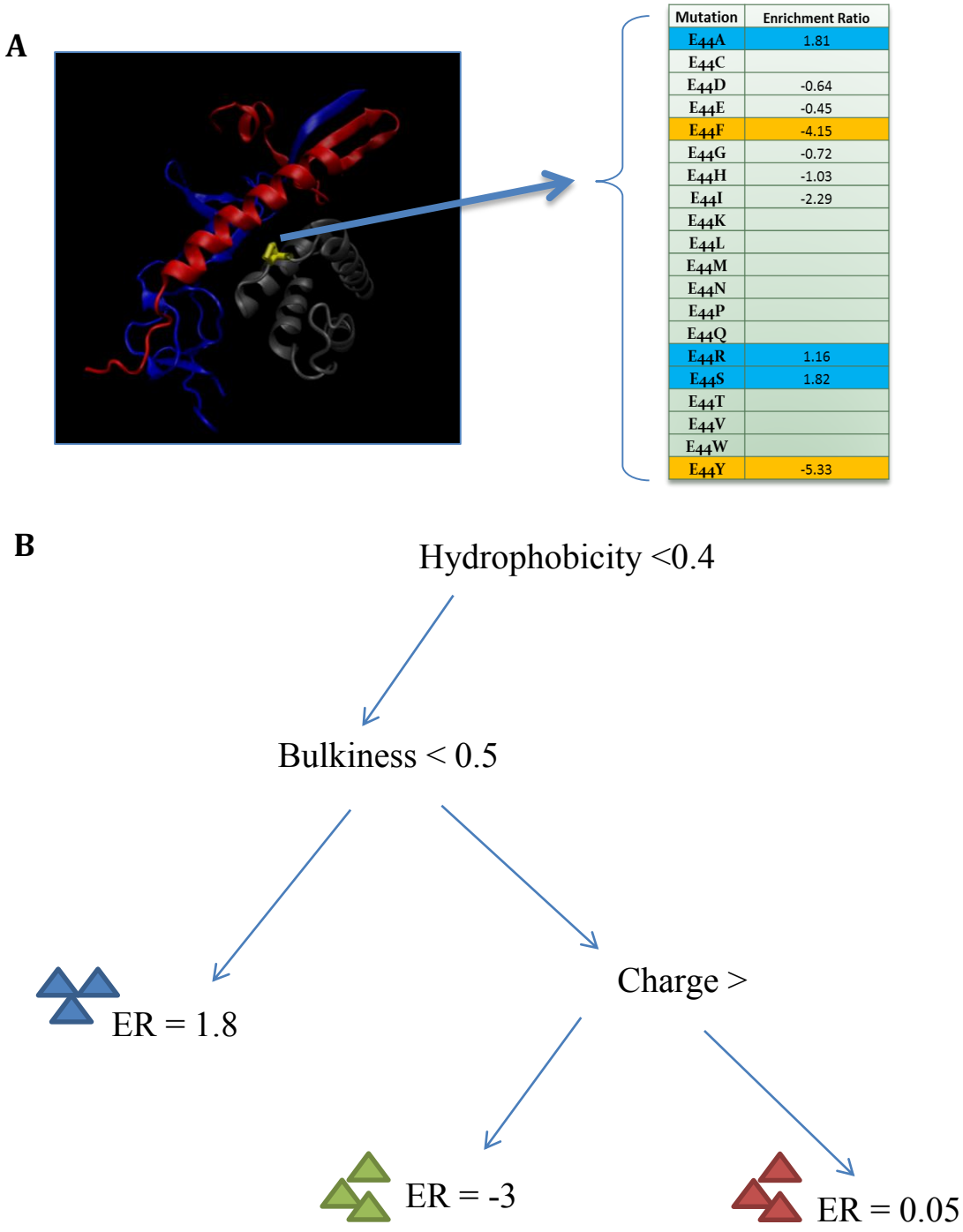



**Figure 4.5: Depiction of the Capri 26, round 2, strategy.**
(A) For each mutated site, the participants were given the enrichment values for 9 randomly selected mutations to train upon. Our approach was to try and uncover correlations between similar residues at a given site. A Position-Specific Model (PSM) was built for this purpose at each mutated site. This PSM model learns why at a given mutation-site, certain mutations have similar Enrichment Ratios (ER). For example, the residue groups (A,R,S) in blue and (F,Y) in yellow depicted above. (B) Shows a toy-example of a tree model that might be learnt at a given mutation-site using the amino-acid properties from

Table **4.1** as features.

## 4.4 CAPRI Round 26 Targets T55 and T56: Round 2

To test whether having some mutational data on T55 and T56 available to the participants, would help prediction, an extended round of predictions was performed. For each residue position, the experimental enrichment values for half of the mutations were given to the participants. The 9 mutations were selected randomly at each position. The task for the community was to make predictions on the remaining half of mutations at each position as depicted in Figure 4.5.

### 4.4.1 Design of a Position Specific Model for ΔΔG Prediction

The most straightforward use of the additional mutation data provided in round 2, is to use it to extend the ΔΔG training dataset built in round 1, and retrain the RF model. However, this method does not exploit the wealth of information we have at each position and a novel position-specific model (PSM) was developed for this extended round. This position specific model is based on the hypothesis that

*'At a given position, 'similar' residue substitutions should act 'similarly''*

The emphasis here is that similar residues should act similarly only at a given residue site i.e. where the context surrounding is controlled for, and constant. By 'act similarly' the hope is that similar residues would have comparable enrichment ratios. The central issue to this method is that defining residue similarity is non-trivial, and depends on what amino-acid property one considers. For example at one residue site, the enrichment ratios for leucine are similar to other hydrophobic residue mutations, whereas at another residue site, they are more similar to mutations to large residues. Therefore, rather than limiting ourselves to one specific amino-acid property for calculating residue similarity, the similarity was instead learnt using a learning model. More precisely, a site-specific model learns which properties are indicative of

correlations between enrichment values for a set of amino acids at a given position. For each residue site (53 positions on T55 and 46 positions on T56), the 9 training mutations, for which the enrichment ratios were available, were used as training data for a RF regression model. The feature set of which, consisted of a set of amino-acid properties shown in

Table **4.1**. The prediction of enrichment ratios for each residue site therefore has, their own unique RF model, training data, and features as depicted in Figure 4.5.

**Table 4.1: A selection of amino-acid properties that form the feature set available to each PSM model.**

| Feature Type | Amino-Acid features |
|---|---|
| **Numerical** | Hydrophobicity |
| **Numerical** | Isoelectric point |
| **Numerical** | Molecular weight |
| **Categorical** | Acyclic |
| **Categorical** | Aliphatic |
| **Categorical** | Hydrophobic |
| **Categorical** | Negative |
| **Categorical** | Positive |
| **Numerical** | Hydropathy |
| **Numerical** | Solvation Potential |
| **Numerical** | Linker Propensity |
| **Numerical** | Surface Exposure |
| **Numerical** | Polarity |
| **Numerical** | Hydrophobicity |
| **Numerical** | Flexibility |
| **Numerical** | Coil Propensity |
| **Numerical** | Bulkiness |

To estimate the generalization ability of the PSM model in advance of the round 2 predictions, 9-fold CV was performed at each position. The PSM-Score achieved an Area Under the Curve (AUC) value of 0.88 for T55 and 0.83 for T56 assuming enrichment ratio of >0 as stabilizing. For the predictions on the blind test set of 1040 mutations as required for round 2 of Capri 26, the thresholds for beneficial mutations were taken at > 0.5 and > -1 for T55 and T56 respectively.

### 4.4.2 Results for T55 and T56 using PSMs

Similar metrics to the ones described in section 4.3.2 in round 1 were used to assess the predictions of all participants. Numerical prediction performance ranked according to the tau-b metric and detection of beneficial mutations ranked according to F1-Score. The results for all participant groups are shown in Figure 4.6 for T55 and Figure 4.7 for T56. The BLOSUM62 is highlighted in blue as a reference and our laboratory's group (G05s) predictions in green. Our model for T55 ranked 1st in both the numerical estimation (Kendal-tau=0.51) and categorical prediction of beneficial mutations (F1-Score=0.49). For T56 results were also consistent for the numerical estimation of enrichment ratios, ranking 2$^{nd}$ with Kendal-tau of 0.41. The detection of beneficial mutations however was not satisfactory and ranked very low with and F1-Score of 0.12. The discrepancy in performance for T56 on the detection of beneficial mutations is more than likely due to using a structure which differed by 5 mutations to that on which the experimental mutations were made on. The improvement in results from round 1 to round 2, which is shared across most of the groups, clearly shows that indeed having some mutational information on the complex in question increases prediction performance. With this in mind, the only other group to show similar success to our group (G-21 Fernandez Recio) was the only other group to exploited positional information in their model. In their case, rather than having regression models unique to each position, amino-acid properties were added to the feature-set. Though not a purely position specific model, the learner in this case is potentially able to exploit correlations across different mutations using their amino-acid properties. Monomer stability may still affect binding indirectly and accounting for monomer stability was also found to be a key component in the top performing groups, including our own (Moretti et al., 2013).

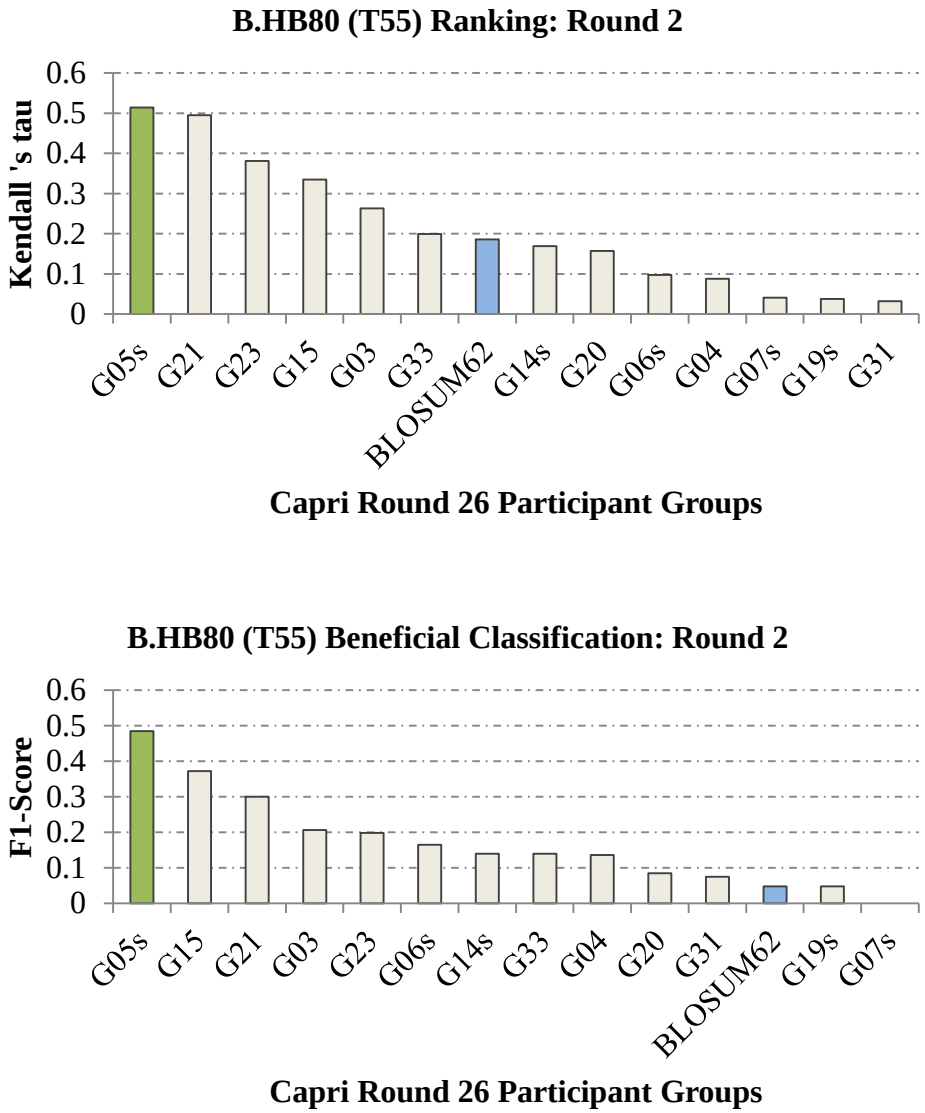
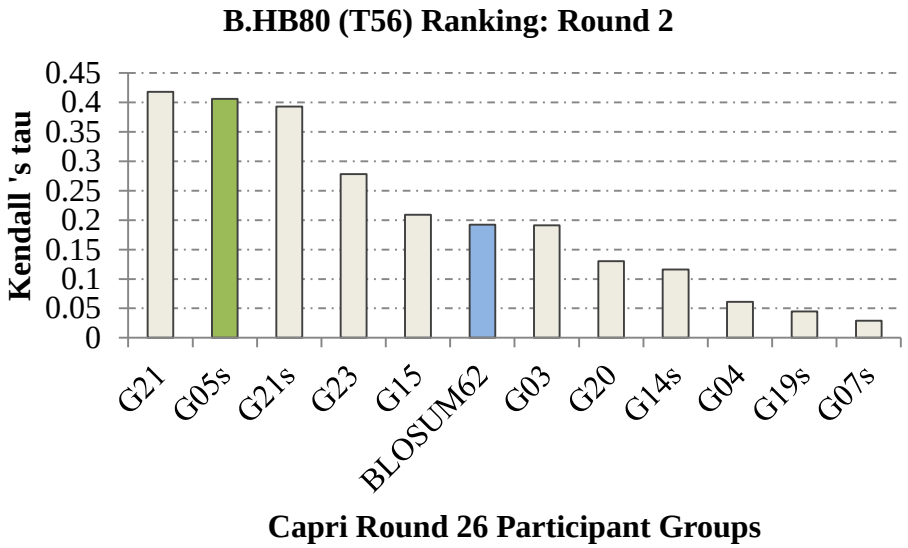



**Figure 4.6: CAPRI 26 T55, round 2 prediction performance of all participant groups.**

A number of residue sites on T55 are mutated to each of the 20 amino-acid residues and an experimental enrichment ratio (proxy to the ΔΔG) is measured for each. Round 2 differs from round 1 in that at each residue site mutated, the enrichment ratios of 10 randomly selected mutations are given to the participants to train upon. The participants are asked to make predictions on the remaining mutations. Predictions were submitted in numerical form, scaled in the range of [0,1] according to how beneficial the mutation is thought to be. The groups' numerical predictions are ranked according to the Kendall's tau (top panel). Predictions were also submitted in categories (Deleterious/Neutral/Beneficial) and the F1-Score used for ranking the performance of the beneficial mutation detection (bottom panel). In blue are the reference BLOSUM62 predictions and in green (G05s) are the prediction result assessment scores from our group.



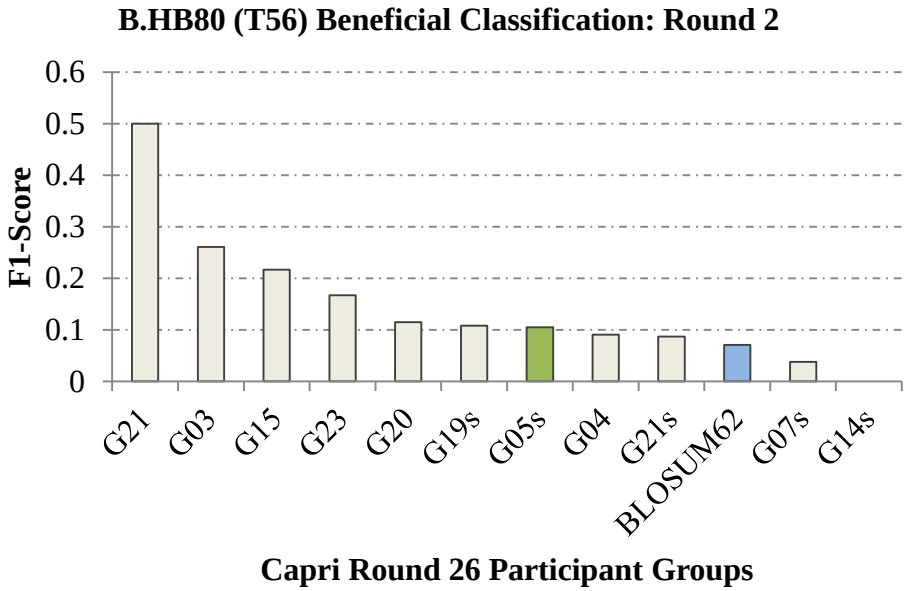




**Figure 4.7: CAPRI 26 T56, round 2 prediction performance of all participant groups. Figure legend details as for Figure 4.8.**

### 4.4.3 Contribution of the PSM Model to Prediction Accuracy

Though the groups exploiting positional information were identified as the top performing groups in Capri 26 round 2, to assess whether the PSM-Score was in fact the driving force behind the success of our results, further analysis was performed. Three sets of test predictions on T55 were generated. Training was performed on half of the T55 mutations given in round 1 and testing performed on the second half. T56 was excluded from this analysis as part of the structure was incorrect (see section 4.3.2). The first set 'All Molecular', refer to predictions from a RF model trained on the feature set compromising of all molecular features described in 4.3.1 (i.e. without the PSM-Score as a feature). A second set 'PSM-Score', are predictions from the PSM model and a third set, predictions from a RF model trained on all features including the PSM-Score combined. Several classification measures (described in methods section 2.4) were used to assess the performance of each model and presented in Table 4.2. The PSM-Score on its own (F1-Score=0.571) performs markedly better than the 'All Molecular' model (F1-Score=0.273). The addition of the PSM-Score with the 'All Molecular' feature set improves the performance over that of the 'All Molecular' set (F1-Score=0.444). To further confirm the prominence of the PSM-Score, the RF feature importance measures were invoked for the model trained on the 'All molecular' features and that when the PSM-Score was added to the feature set. The top 10 features are presented in Table 4.3. These include, statistical potentials, OPUS_PSP (Lu et al., 2008), and GOAP (Zhou and Skolnick, 2011), together with physics-based terms such as the Lenard-Jones repulsive terms (fa_rep) from PyRosetta (Chaudhury et al., 2010) and FoldX (Schymkowitz et al., 2005); and the position of the applied mutation as captured by its solvent accessibility using Naccess. Upon including the PSM-Score to this feature set, the RF model ranks it as the most important feature, superseding the importance (5-Fold increase over 2nd ranked descriptor) of any other molecular descriptor available. From these results, we can conclude that the success of our group's predictions for T55 in Round 2 was mostly attributed to the PSM-Score.

**Table 4.2: Classification Performance for 3 RF Classifiers on T55 Test Mutations.**

All RF Models are trained on half of the T55 mutations (enrichment ratios) given to the participants of Capri 26 round 1. Performance shown in table that of the test predictions for the remaining half of the T55 mutations (enrichment ratios). 'All Molecular', refer to predictions from a RF model trained on the feature set compromising of all molecular features described in 4.3.1 (i.e. without the PSM-Score as a feature). A second set 'PSM-Score', are predictions from the PSM model and a third set, predictions from a RF model trained on all features including the PSM-Score combined.

| Feature Set | TPR | FPR | MCC | F1 | Acc | Spec | Prec | Rec |
|---|---|---|---|---|---|---|---|---|
| All Molecular | 0.158 | 0.000 | 0.391 | 0.273 | 0.969 | 1.000 | 1.000 | 0.158 |
| PSM-Score | 0.421 | 0.002 | 0.603 | 0.571 | 0.977 | 0.998 | 0.889 | 0.421 |
| All Molecular + PSM-Score | 0.316 | 0.004 | 0.475 | 0.444 | 0.971 | 0.996 | 0.750 | 0.316 |

**Table 4.3: Comparison of Top 10 Features for 'All Molecular' Model and 'All Molecular + PSM-Score' Model**.

Importance is extracted from the in-built RF feature importance measure.

| All Molecular | | All Molecular + PSM-Score | |
|---|---|---|---|
| **Feature Name** | **RF Importance** | **Feature Name** | **RF Importance** |
| fxGOAP2 | 209.6 | PSM-Score | 853.7 |
| SolvAccessNorm | 181.7 | fxGOAP2 | 151.8 |
| fxOPUS1 | 171.4 | SolvAccessNorm | 117.4 |
| pyfa_rep | 164.9 | fxOPUS1 | 99.6 |
| fxGOAP1 | 127.4 | pyfa_rep | 96.7 |
| fxfa_rep | 126.1 | fxGOAP1 | 94.0 |
| SolvAccess | 101.6 | fxfa_rep | 78.4 |
| dFOLDX | 96.6 | SolvAccess | 74.6 |
| fxvdwaals | 92.5 | fxOPUS3 | 72.1 |
| fxelec | 81.6 | Fxvdwaals | 66.6 |

## 4.5 Discussion and Conclusions

The blind test trial in the first round, were no mutational information was given for the targets in question, does highlight the current inability to characterise mutations on unseen complexes for most models. The mean and maximum values from all participating groups show an increase in predictive performance upon the availability of mutational information on the target complexes (See Table 4.4).

**Table 4.4: Performance comparison of group predictions for T55 and T56 from round 1 to round 2.**

Shown are the mean Kendall-tau and F1-Score values for all groups; the Kendall-tau and F1-Score values for the best performing group; the Kendall-tau and F1-Score values for our Laboratory's group (G05s).

| **Performance Measure** | **Round 1** | | **Round 2** (available mutational information) | |
|---|---|---|---|---|
| | T55 | T56 | T55 | T56 |
| **Kendall-tau (mean / max / G05s)** | 0.14 / 0.30 / 0.09 | 0.09/ 0.23 / 0.09 | 0.21 / 0.51 / 0.51 | 0.21 / 0.42 / 0.41 |
| **F1-Score (mean / max / G05s)** | 0.11 / 0.34 / 0.34 | 0.12 / 0.26 / 0.14 | 0.19 / 0.49 / 0.49 | 0.16 / 0.5 / 0.11 |

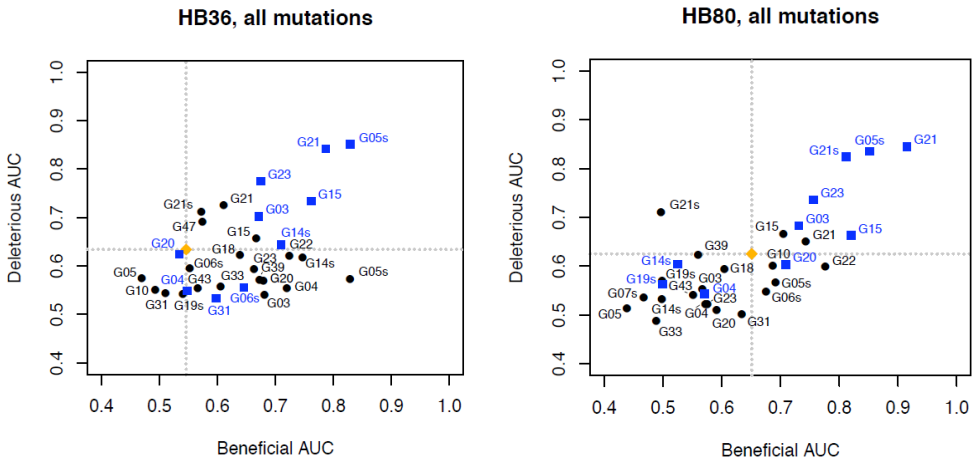



**Figure 4.8: Plot of the AUC for the detection of beneficial mutations vs. that of deleterious mutations.**
Left panel HB36 (T55) and right panel HB80 (T56). Black circles show the prediction for the first round and blue squares show the predictions for the second round. The AUC values for the BLOSUM prediction are indicated by an orange diamond and grey dotted line. Figure taken from supplementary information of (Moretti et al., 2013).

A similar observation is made on assessment of the AUCs when having no mutational information (Figure 4.8 black dots) to having mutational information (Figure 4.8 blue dots). Though clearly present, targeting the source of this reduction in ability to predict changes in binding affinity upon mutations on complexes not in the training set is not trivial. This may be due to insufficiently diverse datasets of protein-protein structures on which training is performed or an underrepresentation of certain residue types. For example most ΔΔG training sets are dominated by alanine-scanning data. The training set derived from the literature for the training of our round 1 model contained > 50% single-point alanine mutations. In contrast, only 5% of the mutations on T55 and T56 were alanine mutations. This also stresses the importance of performing leave-complex-out and more stringent forms of cross validation which account for related complexes when evaluating models for ΔΔG prediction (Moal et al., 2011, Lise et al., 2011).

Whereas the participant' results from round 2 show that the availability of mutational information for the complex on which other mutations must be predicted, helps this endeavour, the PSM model takes this one step further. This,

by showing that the availability of mutational information at the given site on which other mutations must be predicted, helps further the prediction. Interestingly enough, an approach similar to the one taken here, with the difference that it is applied for the ΔΔG of protein stability and not binding, also confirmed that using known ΔΔG values of mutations at the query position improves the accuracy of ΔΔG predictions for other mutations in that position (Wainreb et al., 2011). Unfortunately, the necessity of this site specific mutational informational for training, limits the PSMs application, as rarely is such experimental data on binding available.

# Chapter 5

# 5 Prediction of Hotspot Residues on Protein-Protein Interfaces

## 5.1 Introduction

In Chapters 6-8, a number of descriptors derived from hotspot counts, energies and distribution are designed. These are termed as hotspot descriptors, and are used for characterizing the change in off-rate of protein interactions upon mutation. To generate the hotspot descriptors six hotspot prediction algorithms are used. Two are designed in this work; *RFSpot* and *RFSpot_KFC2*. In addition, hotspot descriptors were also generated using publicly available hotspot predictors *KFC2* (*KFC2a*, *KFC2b*) (Zhu and Mitchell, 2011) and *Hotpoint* (*RFHotpoint1*, *RFHotpoint2)* (Tuncbag et al., 2010).

## 5.2 Methods

### 5.2.1 RFSpot and RFSpot_KFC2

For training and benchmarking the hotspot predictors *RFSpot* and *RFSpot_KFC2*, the SKEMPI alanine dataset described in 2.1.4 is used. For each training example in the dataset and hence *wild-type* complex PDB structure, a number of molecular descriptors, describing various aspects of the interaction, were calculated. These descriptors have already proven successful in our previous work related to the prediction of *wild-type* binding free energies (Moal et al., 2011) and *wild-type* kinetic rate constants (Moal and Bates, 2012). A full list and explanation of the molecular descriptors can be found in the Table 2.1 under the column HS. After calculation of the molecular descriptors on the *wild-type* complex PDB structure, each respective structural mutation was made using FoldX (Schymkowitz et al., 2005) and the same set of molecular descriptors recalculated. Each descriptor, fed into the learning model, is determined as the difference between the mutant and *wild-type* descriptor value:

$$\Delta E_{Desc} = \Delta E_{Desc}^{MUT} - \Delta E_{Desc}^{WT} \qquad 5.1$$

As a learning algorithm the Random Forest (RF) classifier model is employed (Breiman, 2001b), using 1000 trees and an mtry (*i.e.* number of random variables sampled as candidates for a split) of 15. The RF learner is well suited for high dimension datasets, such as the one described here with 110 features. Throughout the thesis, this RF hotspot classifier algorithm is referred to as *RFSpot*. *RFSpot_KFC2* is a similar classifier model to *RFSpot* with the difference that it adds to the 110 molecular features set, 13 features from the original *KFC2a* and *KFC2b* models. These include: *res_hp*, *pos_per*, *delta_tot*, *core_rim*, *rot5*, *plast4*, *plast5*, *fp10* from *KFC2a* and *res_size*, *ratio5*, *rot4*, *hp5*, *fp9* from *KFC2b*. Details on the calculation of each specific descriptors are described in (Zhu and Mitchell, 2011), most notably they include features which position the mutation using solvent accessibility. This enables the model to exploit the fact that

hotspots tend to occur in regions of low solvent accessibility (Bogan and Thorn, 1998) and is the key difference between *RFSpot* and *RFSpot_KFC2*.

### 5.2.2 RFHotpoint1, RFHotpoint2, KFC2a and KFC2b

Similar to *RFSpot* and *RFSpot_KFC2*, *RFHotpoint1* and *RFHotpoint2*, are Random Forest hotspot classifiers trained on the SKEMPI alanine dataset which use only features from the original *Hotpoint* server as features. These include: *relativeComplexASA*, *relativeMonomerASA*, *pairPotential*, *complexASA* as described in (Tuncbag et al., 2010). *RFHotpoint2* differs from *RFHotpoint1* in that for the former, the threshold is lowered to allow for more hotspot detections at the cost of a higher FPR. The reason behind developing the *RFHotpoint* models is due to the fact that the original *Hotpoint* server does not associate an energetic or confidence value to its hotspot prediction, hence hotspot descriptors which make use of hotspot energies cannot be calculated. *RFHotpoint* models therefore enable us to use *Hotpoint* features, trained on a larger dataset of SKEMPI instead of ASEdB (as in the original Hotpoint algorithm) and most importantly, associated confidence values to our hotspot predictions using the Random Forest model. To validate *RFHotpoint1* and *RFHotpoint2* as a representative alternative to *Hotpoint*, predictions from *Hotpoint* server were generated for the SKEMPI alanine dataset. Any prediction for mutations also in *ASEdB* where removed since Hotpoint uses *ASEdB* as training data. The predictions are compared to the 20-fold test predictions of *RFHotpoint1* and *RFHotpoint2* for the same mutations and classification results are shown below in Table 5.1. Both *RFHotpoint1* and RFHotpoint2 achieve higher MCCs than *Hotpoint* and are therefore fare representations of this hotspot predictor.

**Table 5.1: Performance comparison of *RFHotpoint1* and *RFHotpoint2* with server prediction of Hotpoint on SKEMPI.**

| Hotpot Predictor | TPR | FPR | MCC | F1 | Acc | Spec | Prec |
|---|---|---|---|---|---|---|---|
| Hotpoint | 0.500 | 0.379 | 0.113 | 0.424 | 0.584 | 0.621 | 0.368 |
| RFHotpoint1 | 0.360 | 0.128 | 0.268 | 0.437 | 0.715 | 0.872 | 0.554 |
| RFHotpoint2 | 0.570 | 0.303 | 0.253 | 0.505 | 0.658 | 0.697 | 0.454 |

For *KFC2a* and *KFC2b*, no models needed to be re-trained again, as the original predictions from *KFC2* server have associated with them an energetic value that can be directly used for the calculation of the hotspot descriptors.

### 5.2.3 Generation of Hotspot Energies

In the Random Forest classifier model used in *RFSpot*, *RFSpot_KFC*, *RFHotpoint1* and *RFHotpoint2*, each tree in the 1000-tree forest makes is own class prediction (Hotspot / Non-Hotspot) of the mutation in question. The class which accumulates the majority of tree-votes, is the predicted class, and the difference in the number of votes for the hotspot class relative to the non-hotspot class ($Votes_{Hotspot}$- $Votes_{Non\text{-}Hotspot}$ ) indicates the model's confidence in the predicted class. In this work, these confidence values are used as an estimation of hotspot *ΔΔG*s. The rationale is that the higher the confidence value, the more trees have predicted this to be a hotspot, implying that larger numbers of different feature subsets consider this to be a hotspot also. Given that several different aspects of the protein interaction have vouched for the example to be a hotspot, then it is expected that the hotspot *ΔΔG* is larger in magnitude. To confirm this, RF regression models are trained on the same training data as *RFSpot, RFSpot_KFC2, RFHotpoint1 & RFHotpoint2* RF classifiers, in order to generate true *ΔΔG* predictions and compared to the confidence values generated by each of them. Note that *RFHotpoint1* and *RFHotpoint2* use same confidence values and only differ by their threshold on those confidence values; hence one correlation for the confidence values is presented for both. The confidence values of the classifier models, show correlations of R=0.88, R=0.86, R=0.86 with the regression models' *ΔΔG* predictions for *RFSpot, RFSpot_KFC2* and *RFHotpoint1 & 2* respectively. Therefore, apart from the differences in their absolute values, the confidence values do provide relative values, which have a direct linear relationship to *ΔΔG*. On assessment of the MCCs of the regression RF models at a threshold of >=2 kcal/mol, the regression models achieve lower MCCs to that of the classifier models, all of which are the result of higher FPR. Given that 75% of the *ΔΔG* data is of the negative non-hotspot class, minimal increases in the FPR add a significant number of false-positives, which would subdue the gain of additional hotspots, correctly detected. Therefore the use of a classifier in

*RFSpot, RFSpot_KFC2, RFHotpoint1 and RFHotpoint2,* enables us to achieve a lower false positive rate, to that of a regression model, but still be able to have confidence values that relate directly to *ΔΔG*. For the sake of simplicity, we refer to the *ΔΔG* confidence values extracted by the method described here as *ΔΔG*s.

## 5.3 Results

### 5.3.1 Performance of Hotspot Predictors on the SKEMPI Alanine Dataset

The predictive accuracy of the hotspot predictors from which the hotspot descriptors are generated from (i.e. *RFSpot*, *RFSpot_KFC2*, *RFHotpoint1*, *RFHotpoint2*, *KFC2a* and *KFC2b*), is assessed on the SKEMPI alanine dataset using a number of classification performance measures. All hotspot prediction algorithms and performance measures in this section are related to how well the hotspot predictors are able to detect ΔΔGs of single-point alanine mutations which satisfy ΔΔG > 2kcal/mol i.e. the prediction of hotspots. For *RFSpot*, *RFSpot_KF2*, *RFHotspoint1* & *RFHotpoint2*, the prediction results from a 20–Fold cross-validation are used, whereas for *KFC2a* and *KFC2b* the predictions from KFC2 (Zhu and Mitchell, 2011) server are used. Note that for *KFC2a* and *KFC2b*, the predictions for the data, which is in SKEMPI and not in ASEdB is presented, as KFC2 server algorithm uses ASEdB mutations for model design and training. The predictions are compared to a number of hotspot prediction algorithms (*KFC2* (Zhu and Mitchell, 2011) *HotPoint* (Tuncbag et al., 2010), *Robetta* (Kortemme and Baker, 2002), *RFMirror* (Wang et al., 2012), and *TSVM* (Lise et al., 2009)). Details on each hotspot predictors and the sources of their predictions are presented in Table 5.2.

**Table 5.2. Summary of hotspot predictors benchmarked in this work and the datasets used.**

| Hotpot Predictor | Description and Source of Predictions |
|---|---|
| **RFSpot** | RFSpot*: Random Forest Model trained on SKEMPI alanine data, with molecular features.<br>RFSpot_KFC2*: Random Forest Model trained on SKEMPI alanine data, with molecular features and features from KFC2a and KFC2b for which KFC2a and KFC2b Features from KFC2 server where available. Random Forest threshold adjusted to achieve same FPR of RFSpot for comparison. |
| **KFC2 (Zhu and Mitchell, 2011)** | KFC2a_Orig: uses predictions on ASEdB.<br>KFC2b_Orig: uses predictions on ASEdB<br>KFC2a *: uses server predictions on SKEMPI alanine data which is not in ASEdB from KFC2 Server<br>KFC2b*: uses server predictions on SKEMPI alanine data which is not in ASEdB from KFC2 Server |
| **Robetta (Kortemme and Baker, 2002)** | Robetta: uses predictions on ASEdB. |
| **Hotpoint (Tuncbag et al., 2009)** | Hotpoint_Orig: uses predictions on ASEdB.<br>Hotpoint uses server prediction on SKEMPI alanine data which is not in ASEdB from Hotpoint Prediction Server<br>RFHotpoint1*: Random Forest Model trained on SKEMPI alanine data, with original Hotpoint Features for which Hotpoint Features from Hotpoint server where available<br>RFHotpoint2*: Random Forest Model trained on SKEMPI alanine data, with Hotpoint Features for which Hotpoint Features from server where available. Random Forest Threshold lowered to allow for more TPs |
| **RFMirror (Wang et al., 2012)** | RFMirror: uses predictions on ASEdB. |
| **SVM score (Lise et al., 2009)** | SVM score: uses predictions on ASEdB. |
| **TSVM score (Lise et al., 2009)** | TSVM score: uses predictions on ASEdB. |

** Indicate Hotspot Predictors used for the generation of hotspot descriptors*

The performance of each hotspot predictor is shown in the Table 5.3. The test sets on which the performance measures are calculated, are different for each predictor. Namely, each test set is the intersection between SKEMPI and the original test set used in each respective work. The highest MCC is achieved by TSVM score (Lise et al., 2011). Note however that, though ranked according to MCC, Table 5.3 shows their performance on different mutations, therefore cannot be relatively compared. A relative comparison between two predictors can only be performed on the intersections of mutations for which both algorithms have unbiased predictions for. Since the all-vs-all comparison of the hotspot predictor algorithms is beyond the scope of this work, the comparison is only made for the two-hotspot predictors developed in this work namely *RFSpot*

and *RFSpot_KFC2*. Given that *RFSpot* and *RFSpot_KFC2* both use SKEMPI, the dataset intersection with other predictors' datasets is the same. First, for each predictor, the test set intersection with SKEMPI is extracted, and each predictor's performance on this intersection is shown in Table 5.4. For comparison, the performance of *RFSpot* and *RFSpot_KFC2* on these same dataset intersections is presented in Table 5.5 and Table 5.6 respectively. Therefore the performance values in corresponding cells of in Table 5.4, Table 5.5 and Table 5.6 are comparable. For better visual inspection, in blue are highlighted the instances in which *RFSpot* or *RFSpot_KFC2* perform better than the respective hotspot predictor. *RFSpot_KFC2* outperforms all hotspot predictors with the exception of TSVM (Lise et al., 2011). TSVM achieves a higher MCC as a result of a higher TPR, though *RFSpot_KFC2* achieves higher accuracy specificity and precision than that of TSVM. With this, it can be concluded that *RFSpot_KFC2* is as good as the best hotspot prediction algorithm available. This gives confidence that the hotspot predictions from *RFSpot_KFC2* that will be used in further chapters for the generation of hotspot descriptors and off-rate prediction are high quality predictions.

**Table 5.3. Performance of Hotspot Descriptors - part 1.**
Performance of Hotspot Predictors on intersection of original data used in original hotspot predictors and SKEMPI. Predictors are ranked according to MCC.

| Predictor | Intersection with SKEMPI | TPR | FPR | MCC | F1 | Acc | Spec | Prec |
|---|---|---|---|---|---|---|---|---|
| **TSVM score** | **TSVM score** | 0.673 | 0.136 | 0.508 | 0.619 | 0.823 | 0.864 | 0.574 |
| **RFSpot_KFC2*** | **RFSpot_KFC2*** | 0.490 | 0.083 | 0.452 | 0.560 | 0.814 | 0.917 | 0.652 |
| **SVM score** | **SVM score** | 0.615 | 0.152 | 0.438 | 0.566 | 0.798 | 0.848 | 0.525 |
| **RFMirror** | **RFMirror** | 0.500 | 0.094 | 0.434 | 0.545 | 0.816 | 0.906 | 0.600 |
| **KFC2a*** | **KFC2a*** | 0.734 | 0.279 | 0.402 | 0.568 | 0.724 | 0.721 | 0.463 |
| **KFC2b** | **KFC2b** | 0.452 | 0.103 | 0.383 | 0.509 | 0.789 | 0.897 | 0.583 |
| **Hotpoint_Orig** | **Hotpoint_Orig** | 0.552 | 0.196 | 0.368 | 0.593 | 0.707 | 0.804 | 0.640 |
| **KFC2b*** | **KFC2b*** | 0.436 | 0.129 | 0.328 | 0.477 | 0.764 | 0.871 | 0.526 |
| **Robetta** | **Robetta** | 0.458 | 0.155 | 0.295 | 0.440 | 0.769 | 0.845 | 0.423 |
| **RFSpot*** | **RFSpot*** | 0.268 | 0.083 | 0.237 | 0.350 | 0.761 | 0.917 | 0.506 |
| **RFHotpoint1*** | **RFHotpoint1*** | 0.319 | 0.125 | 0.229 | 0.395 | 0.711 | 0.875 | 0.517 |
| **RFHotpoint2*** | **RFHotpoint2*** | 0.504 | 0.277 | 0.218 | 0.466 | 0.658 | 0.723 | 0.433 |
| **KFC2a_Orig** | **KFC2a_Orig** | 0.258 | 0.124 | 0.159 | 0.314 | 0.727 | 0.876 | 0.400 |
| **Hotpoint** | **Hotpoint** | 0.500 | 0.379 | 0.113 | 0.424 | 0.584 | 0.621 | 0.368 |

** Indicate Hotspot Predictors used for the generation of hotspot descriptors.*

**Table 5.4: Performance of Hotspot Predictors on SKEMPI part 2.**
Performance of Hotspot Predictors on intersection of original data used in original hotspot predictors and SKEMPI. Positioned for comparison with Table 5.5 and Table 5.6.

| Predictor | Intersection with SKEMPI | TPR | FPR | MCC | F1 | Acc | Spec | Prec |
|---|---|---|---|---|---|---|---|---|
| **RFSpot_KFC2*** | **RFSpot_KFC2*** | 0.490 | 0.083 | 0.452 | 0.560 | 0.814 | 0.917 | 0.652 |
| **RFSpot*** | **RFSpot*** | 0.268 | 0.083 | 0.237 | 0.350 | 0.761 | 0.917 | 0.506 |
| **KFC2a*** | **KFC2a*** | 0.734 | 0.279 | 0.402 | 0.568 | 0.724 | 0.721 | 0.463 |
| **KFC2b*** | **KFC2b*** | 0.436 | 0.129 | 0.328 | 0.477 | 0.764 | 0.871 | 0.526 |
| **Hotpoint** | **Hotpoint** | 0.500 | 0.379 | 0.113 | 0.424 | 0.584 | 0.621 | 0.368 |
| **RFHotpoint1*** | **RFHotpoint1*** | 0.319 | 0.125 | 0.229 | 0.395 | 0.711 | 0.875 | 0.517 |
| **RFHotpoint2*** | **RFHotpoint2*** | 0.504 | 0.277 | 0.218 | 0.466 | 0.658 | 0.723 | 0.433 |
| **KFC2a_Orig** | **KFC2a_Orig** | 0.258 | 0.124 | 0.159 | 0.314 | 0.727 | 0.876 | 0.400 |
| **KFC2b_Orig** | **KFC2b_Orig** | 0.452 | 0.103 | 0.383 | 0.509 | 0.789 | 0.897 | 0.583 |
| **Robetta** | **Robetta** | 0.458 | 0.155 | 0.295 | 0.440 | 0.769 | 0.845 | 0.423 |
| **Hotpoint_Orig** | **Hotpoint_Orig** | 0.552 | 0.196 | 0.368 | 0.593 | 0.707 | 0.804 | 0.640 |
| **RFMirror** | **RFMirror** | 0.500 | 0.094 | 0.434 | 0.545 | 0.816 | 0.906 | 0.600 |
| **SVM score** | **SVM score** | 0.615 | 0.152 | 0.438 | 0.566 | 0.798 | 0.848 | 0.525 |
| **TSVM score** | **TSVM score** | 0.673 | 0.136 | 0.508 | 0.619 | 0.823 | 0.864 | 0.574 |

** Indicate Hotspot Predictors used for the generation of hotspot descriptors*

**Table 5.5: Performance of RFSpot and Hotspot Predictors .**
Performance of *RFSpot* on intersection of original data used in original hotspot predictors and SKEMPI. Highlighted in Blue are instances where *RFSpot* performs better than respective hotspot predictor, as compared with values in the corresponding cells of Table **5.4**.

| Predictor | Intersection with SKEMPI | TPR | FPR | MCC | F1 | Acc | Spec | Prec |
|---|---|---|---|---|---|---|---|---|
| **RFSpot** | **RFSpot_KFC2*** | 0.268 | 0.083 | 0.237 | 0.350 | 0.761 | 0.917 | 0.506 |
| **RFSpot** | **RFSpot*** | 0.268 | 0.083 | 0.237 | 0.350 | 0.761 | 0.917 | 0.506 |
| **RFSpot** | **KFC2a*** | 0.255 | 0.080 | 0.230 | 0.340 | 0.756 | 0.920 | 0.511 |
| **RFSpot** | **KFC2b*** | 0.255 | 0.080 | 0.230 | 0.340 | 0.756 | 0.920 | 0.511 |
| **RFSpot** | **Hotpoint** | 0.279 | 0.118 | 0.199 | 0.361 | 0.698 | 0.882 | 0.511 |
| **RFSpot** | **RFHotpoint1*** | 0.291 | 0.110 | 0.223 | 0.374 | 0.713 | 0.890 | 0.526 |
| **RFSpot** | **RFHotpoint2*** | 0.291 | 0.110 | 0.223 | 0.374 | 0.713 | 0.890 | 0.526 |
| **RFSpot** | **KFC2a_Orig** | 0.323 | 0.072 | 0.316 | 0.417 | 0.781 | 0.928 | 0.588 |
| **RFSpot** | **KFC2b_Orig** | 0.323 | 0.072 | 0.316 | 0.417 | 0.781 | 0.928 | 0.588 |
| **RFSpot** | **Robetta** | 0.417 | 0.062 | 0.418 | 0.500 | 0.835 | 0.938 | 0.625 |
| **RFSpot** | **Hotpoint_Orig** | 0.310 | 0.087 | 0.287 | 0.429 | 0.680 | 0.913 | 0.692 |
| **RFSpot** | **RFMirror** | 0.296 | 0.079 | 0.272 | 0.376 | 0.784 | 0.921 | 0.516 |
| **RFSpot** | **SVM score** | 0.269 | 0.073 | 0.252 | 0.350 | 0.786 | 0.927 | 0.500 |
| **RFSpot** | **TSVM score** | 0.269 | 0.073 | 0.252 | 0.350 | 0.786 | 0.927 | 0.500 |

** Indicate Hotspot Predictors used for the generation of hotspot descriptors.*

**Table 5.6: RFSpot_KFC2 and Hotspot Predictors.**
Performance of *RFSpot_KFC2* on intersection of original data used in original hotspot predictors and SKEMPI. Highlighted in Blue are instances where *RFSpot_KFC2* performs better than respective hotspot predictor as compared with values in the corresponding cells of Table **5.4**.

| Predictor | Intersection with SKEMPI | TPR | FPR | MCC | F1 | Acc | Spec | Prec |
|---|---|---|---|---|---|---|---|---|
| **RFSpot_KFC2** | **RFSpot_KFC2*** | 0.490 | 0.083 | 0.452 | 0.560 | 0.814 | 0.917 | 0.652 |
| **RFSpot_KFC2** | **RFSpot*** | 0.490 | 0.083 | 0.452 | 0.560 | 0.814 | 0.917 | 0.652 |
| **RFSpot_KFC2** | **KFC2a*** | 0.500 | 0.098 | 0.436 | 0.556 | 0.803 | 0.902 | 0.627 |
| **RFSpot_KFC2** | **KFC2b*** | 0.500 | 0.098 | 0.436 | 0.556 | 0.803 | 0.902 | 0.627 |
| **RFSpot_KFC2** | **Hotpoint** | 0.535 | 0.133 | 0.424 | 0.582 | 0.765 | 0.867 | 0.639 |
| **RFSpot_KFC2** | **RFHotpoint1*** | 0.511 | 0.113 | 0.431 | 0.574 | 0.776 | 0.887 | 0.655 |
| **RFSpot_KFC2** | **RFHotpoint2*** | 0.511 | 0.113 | 0.431 | 0.574 | 0.776 | 0.887 | 0.655 |
| **RFSpot_KFC2** | **KFC2a_Orig** | 0.452 | 0.082 | 0.419 | 0.528 | 0.805 | 0.918 | 0.636 |
| **RFSpot_KFC2** | **KFC2b_Orig** | 0.452 | 0.082 | 0.419 | 0.528 | 0.805 | 0.918 | 0.636 |
| **RFSpot_KFC2** | **Robetta** | 0.583 | 0.052 | 0.583 | 0.651 | 0.876 | 0.948 | 0.737 |
| **RFSpot_KFC2** | **Hotpoint_Orig** | 0.483 | 0.087 | 0.451 | 0.596 | 0.747 | 0.913 | 0.778 |
| **RFSpot_KFC2** | **RFMirror** | 0.500 | 0.063 | 0.495 | 0.581 | 0.841 | 0.937 | 0.692 |
| **RFSpot_KFC2** | **SVM score** | 0.519 | 0.068 | 0.499 | 0.587 | 0.844 | 0.932 | 0.675 |
| **RFSpot_KFC2** | **TSVM score** | 0.519 | 0.068 | 0.499 | 0.587 | 0.844 | 0.932 | 0.675 |

** Indicate Hotspot Predictors used for the generation of hotspot descriptors.*

## 5.4 Discussion

In this chapter, the computational prediction of hotspot residues was investigated. Using a large set of mutations from SKEMPI, two hotspot predictors were built and benchmarked against a number of hotspot predictors. Both *RFSpot* and *RFSpot_KFC2* use a set of statistical and physical descriptors to characterize the $\Delta\Delta G$ of the mutated residue in question. In addition to these features, *RFSpot_KFC2* also uses features which estimate the local flexibility of the neighbouring residues, and the solvent accessibility of the residue in question. To maintain an unbiased prediction scheme, based purely on molecular and physical descriptors, the inclusion of such descriptors in *RFSpot* is intentionally avoided, and this is the probable reason for its low TPR. It is understood that the addition of descriptors which relate to solvent accessibility may increase the TPR of *RFSpot* as this would enable the RF learner to distinguish between mutations performed at the core as opposed to those at the rim, where less hotspots occur (Bogan and Thorn, 1998). Indeed, it has been shown that a predictor with just 3 solvent accessibility features can result in a

sensitivity of 0.87 (*i.e.* TP/(TP+FN)) for the BID test set (Zhu and Mitchell, 2011). This is also confirmed using *RFSpot_KFC2* which introduces features related to solvent accessibility and upon setting the threshold to achieve the same FPR to that of *RFSpot*, the TPR is increased from 0.27 (in *RFSpot*) to 0.49 (in *RFSpot_KFC2*). With this in mind, low solvent accessibility is not a sufficient indicator of hotspots as most residues at the core are still non-hotspots (Bogan and Thorn, 1998, Zhu and Mitchell, 2011). Such models are biased towards predicting hotspots at the core regions and may lay the risk of not being able to detect mutations in other regions, as accurately. The risk is higher if the training set is small and other regions outside the core are underrepresented. Though *RFSpot_KFC2* uses such solvent accessibility descriptors, the model uses other molecular features and is trained on a more diverse alanine dataset of SKEMPI as opposed to the ASEdB. The use of a data-partitioning model such as the random forest is particularly useful in this scenario; as it is able to characterize mutations in regions of low solvent accessibility with different features to those in regions of high solvent accessibility.

# Chapter 6

# 6 Characterizing Change in Off-Rate upon Mutation using Hotspot Energy and Architecture

*'Non agunt nisi fixat' - A substance will not work unless it is bound* [Paul Ehrlich]

## 6.1 Introduction

In this chapter, the stability of protein-protein complexes is further probed using features and models specifically designed for the prediction of off-rates. Several mutational studies show that the off-rate can be independently modulated with no change to on-rate of an interaction (Moal and Fernandez-Recio, 2012), which suggests that at least in particular situations, some energetic factors are specific to the off-rate. As described in section 2.1.2, in contrast to on-rates, the off-rates are generally insensitive to ionic strength (Moore et al., 1999, Alexander-Brett and Fremont, 2007) indicating short-range forces should be more prominent. In the work of Moal and Bates (2012), a large set of molecular descriptors are

assessed for their correlation with $k_{on}$ and $k_{off}$. In contrast to $k_{on}$, no significant correlations where found for $k_{off}$. This suggests that currently available descriptors are not able to represent the correlative effects of the dissociation rate, or more possibly there is no single dominant contributor to the stability to $k_{off}$. Counter intuitive to what one would expect, the most prominent descriptors were found to be *coarse-grained* statistical potentials, rather than *fine-grained* atomic potentials (Moal and Bates, 2012). Nevertheless, correlations for $k_{off}$ were still lower than 0.35 and insignificant with ($p>0.01$). Here, it is proposed that the change in off-rate brought about by an interface mutation, can be explained by changes in the hotspot energy landscape as a result of the same mutation. For a mutation in question, computational alanine scans are preformed pre- and post-mutation to determine the hotspot energy landscape in each case. Hotspot descriptors are then designed to capture the changes in the pre- and post-mutation landscape. Using these hotspot descriptors, several off-rate prediction models are then developed and their performance compared to models using standard molecular descriptors.

Work on hotspots can be divided into three categories; the design of hotspot predictor algorithms (Tong et al., 2004, Wang et al., 2012, Xia et al., 2010, Darnell et al., 2007, Morrow and Zhang, 2012, Tuncbag et al., 2010, Zhu and Mitchell, 2011, Lise et al., 2009, Lise et al., 2011, Kortemme and Baker, 2002); the study of distribution related properties of hotspots (Bogan and Thorn, 1998, Keskin et al., 2005); and investigations into the use of hotspot regions as drug target sites (Grosdidier and Fernandez-Recio, 2012, Ma and Nussinov, 2007, Thangudu et al., 2012). However, no work has used hotspot architecture to infer the dynamics of complex dissociation. In this chapter, it is shown how using hotspot descriptors and hence, the energies of single-point mutations to alanine, can be used to describe off-rate changes by mutations other than alanine, and also, multi-point mutations.

### 6.1.1 Anchor Points of Interaction – Hypothesis for Linking Hotspots Energy and Distribution to the Off-Rate

The hotspot descriptors designed in this work aim to capture both the energetics and distribution (referred to as the 'hotspot landscape' for sake of clarity) of hotspot residues across the interface – more precisely, they capture the changes to the hotspot landscape brought about by the mutation in question. To link this to change in off-rate, the following hypothesis is proposed:

*'Thinking of hotspot residues as the 'anchor points' of an interaction, if a mutation increases the number of 'anchor points' (or more precisely, the hotspot landscape has more favourable energy and distribution), this will result in higher complex stability, which manifests itself as a decrease in the off-rate'*

### 6.1.2 Why Hotspots?

There are two main properties of hotspots that could qualify them to be the ideal candidates to use as features over conventional molecular features:

1. Synergy: What makes a hotspot? Essentially, the occurrence of a hotspot is not limited to any particular physical phenomena. Instead, hotspots result from the synergistic effect of a number of factors. These may include physicochemical and structural properties (Ofran and Rost, 2007). Therefore, in terms of computation, hotspots prediction algorithms may encompass the combined contribution from a number of features; for example in this work, the hotspot prediction algorithm is built using a broad range of molecular features listed in Table 2.1 of Methods section 2.2

2. Distribution: Hotspots share defined patterns of distributions that are not always observable in more traditional protein-protein interface features. It is known that hotspots tend to cluster into hotregions, within which, hotspots are suggested to be energetically cooperative (Keskin et al.,

2005, Reichmann et al., 2005). It has also been shown that hotspots tend to occur more at the core regions as opposed to the rims; however, low solvent accessibility is not a sufficient property for a residue to be a hotspot (Bogan and Thorn, 1998). Even though these distributional patterns have been found at protein-protein interfaces, their effect on protein complex stability has not yet been investigated, nor exploited. Trying to uncover the advantage, if any, of these properties in complex stability is the aim of using some of the hotspot descriptors designed in this work.

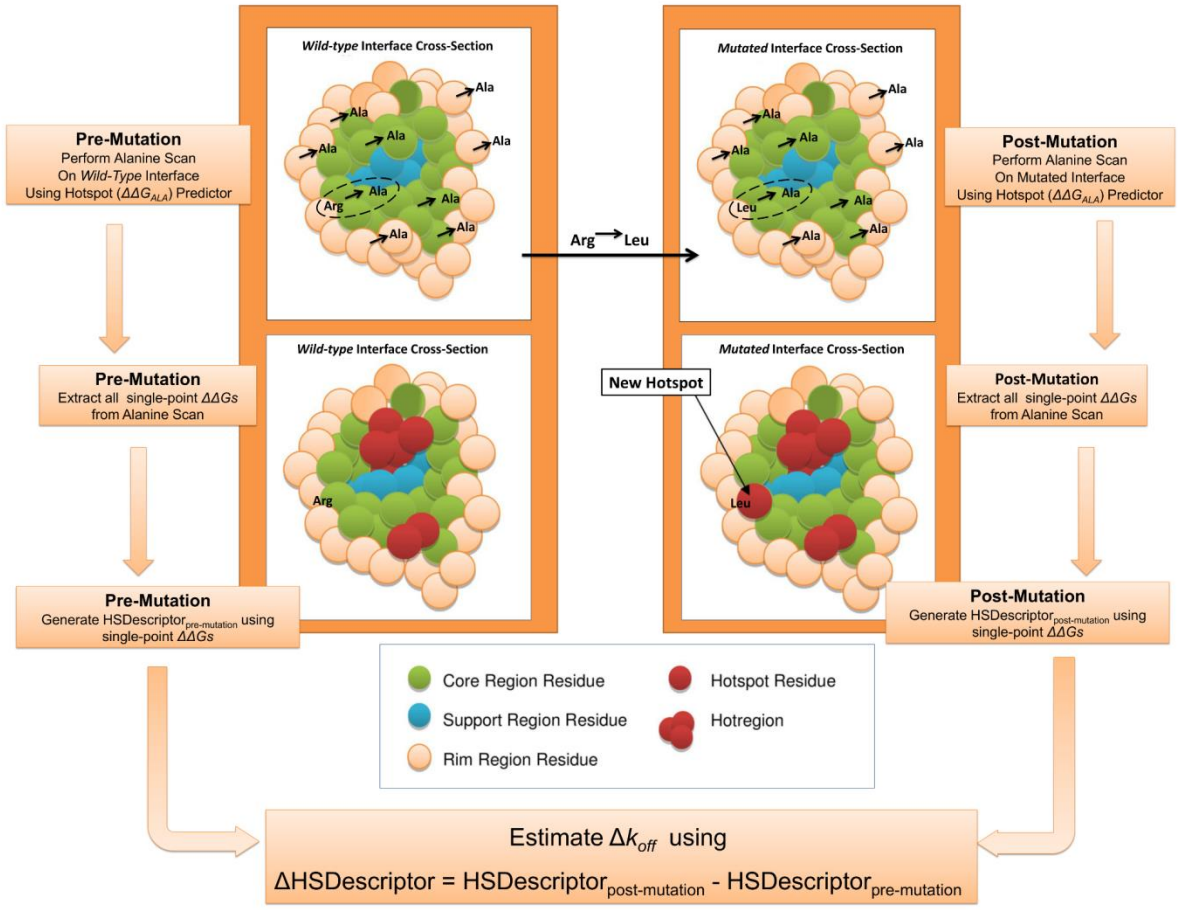
3. Simplicity: The description of an interface though hotspots is conceptually and visually straightforward. From a computational stand-point, the advantage is that one is able to represent an interface with a much smaller set of features without compromising accuracy, as the effects of several phenomena is still encompassed within the hotspots themselves. This reduction in feature set size is also particularly attractive in the context of machine learning algorithms.

## 6.2 Methods

### 6.2.1 Hotspot Descriptor Generation

The use of hotspot predictions and subsequently hotspot descriptors for characterizing off-rates is depicted in Figure 6.1. First a pre-mutation alanine scan is performed; essentially this translates to using a hotspot predictor of choice on each residue at the interface. This generates a collection of single-point alanine $\Delta\Delta G$s that are then employed differently depending on the hotspot descriptor in question (See Table 6.3 for a list of hotspot descriptors). For example if one uses *Int_HS_Energy*, then this hotspot descriptor will sum all the energies of only the hotspot residues. After all the hotspot descriptors for the *wild-type* complex are calculated, the mutation in question is applied using FoldX (Schymkowitz et al., 2005), such as the Arg to Leu mutation in Figure 6.1. Then, using a hotspot predictor as in the *wild-type* scan, another computational alanine scan is

performed on the mutated interface. Again, all single-point alanine *ΔΔG*s are then fed into the hotspot descriptors. Continuing with the example of *Int_HS_Energy* as a hotspot descriptor, here the *ΔΔG*s of only the hotspot residues on the mutated interface are summed, and the final descriptor value will be the change in the sum of the single-point *ΔΔG*s to alanine of all hotspot residues pre- and post-mutation. This value is then correlated to $\Delta k_{off}$ Leu->Arg.




$$Int\_HS\_Energy = (\sum_{nHS=1}^{N} \Delta\Delta G_{nHS \rightarrow Ala})^{\text{post-mutation}} - (\sum_{nHS=1}^{N} \Delta\Delta G_{nHS \rightarrow Ala})^{\text{pre-mutation}} \propto \Delta k_{off}$$

**Figure 6.1: Off-rate estimation using hotspot energies and organization.**
In this work a set of hotspot descriptors for characterizing off-rate changes upon mutation is generated. The hotspot descriptors use single-point alanine ΔΔGs from computational alanine-scans performed by hotspot prediction algorithms. The single-point alanine ΔΔGs are encapsulated in hotspot descriptors, which are then applied to the prediction of changes in off-rate upon single-point and multi-point mutations to all residue types. To do so, for a given wild-type complex structure, the interface is scanned for hotspots using a hotspot prediction algorithm. The single-point alanine ΔΔGs from the scan are extracted and stored.

Next, the structural mutation in question is applied and the mutated interface re-scanned for hotspots. This generates a new set of single-point alanine ΔΔGs for the mutated interface. Note that the mutation in question may also affect the hotspot energies of other neighbouring residues that are not mutated. The two sets of ΔΔGs are then used to generate a set of hotspot descriptors, where the final hotspot descriptor value is the change in the descriptor's value from mutant to wild-type. For example in the case of Int_HS_Energy, the final value is the change in the sum of the ΔΔGs, of all hotspot residues, pre- and post-mutation.

## 6.3 Results

### 6.3.1 Hypothesis Validation Part 1 - Explaining Off-Rate Changes Using ΔΔG Energies from Single-Point Alanine Mutations

The hotspot descriptors map the effects of single-point and multipoint mutations to all residue types into energies of only single-point alanine mutations. Therefore, this enables off-the-shelf hotspot predictors to be used for a new application, the prediction of off-rate change upon mutation. To assess the proposition, a representative hotspot descriptor *Int_HS_Energy* is used as an example:

$$Int\_HS\_Energy = \left( \sum_{n_{HS}=1}^{N} \Delta\Delta G_{n_{HS} \to Ala} \right)^{MUT} - \left( \sum_{n_{HS}=1}^{N} \Delta\Delta G_{n_{HS} \to Ala} \right)^{WT} \quad 6.1$$

where the difference of hotspot energies pre- and post-mutation is used as an estimate for the $k_{off}$ term, $\Delta\log_{10}(k_{off})$. *Int_HS_Energy* shows a PCC of R = -0.51 with the 713 experimental $\Delta\log_{10}(k_{off})$ mutation values found in the SKEMPI database. This correlation (R=-0.51) is the average of six PCC values, as generated per each *Int_HS_Energy* descriptor of the six-hotspot predictors assessed in this work. One should note that even through the hotspot predictors in this work employ different features and learning algorithms to build their hotspot models (See 5.2), all hotspot predictors except one (the worst performing hotspot predictor) show a consistent correlation of |R|>0.5 with *Int_HS_Energy* (See Table 6.1). The strength of observed correlations of *Int_HS_Energy* with $\Delta\log_{10}(k_{off})$ is in line with the hypothesis proposed and shows that when considering changes in off-rate, single-point and multi-point mutations to all residue types can be explained by energies of only single-point alanine mutations. In addition, the negative sign of the correlation confirms that

an increase in 'anchor points' – more precisely an increase in hotspot energies (from *wild-type* to mutant), results in a more stable interface and hence a lower off-rate.

**Table 6.1: Pearson's Correlation Coefficient (PCC) of hotspot descriptors with experimental $\Delta\log_{10}(k_{off})$.**

Correlations are shown for the 713 off-rate mutations in the SKEMPI database and for hotspot descriptors generated by each hotspot predictor.

| Hotspot Descriptor | RFHotpoint1 | RFHotpoint2 | KFC2a | KFC2b | RFSpot | RFSpot_KFC2 | Mean PCC | Variance in PCC |
|---|---|---|---|---|---|---|---|---|
| **Int_Energy_1** | -0.312 | -0.312 | -0.472 | -0.432 | -0.182 | -0.289 | -0.333 | 0.105 |
| **No_HS** | -0.433 | -0.266 | -0.429 | -0.496 | -0.493 | -0.496 | -0.436 | 0.089 |
| **Int_HS_Energy** | -0.568 | -0.312 | -0.546 | -0.527 | -0.532 | -0.559 | -0.508 | 0.097 |
| **No_Clusters** | 0.101 | -0.069 | -0.075 | -0.272 | -0.284 | -0.285 | -0.147 | 0.159 |
| **MaxClusterSize** | -0.225 | 0.022 | 0.094 | 0.052 | -0.163 | -0.292 | -0.085 | 0.162 |
| **AVG_HS_PathLength** | -0.152 | -0.139 | -0.031 | -0.197 | -0.110 | -0.016 | -0.108 | 0.071 |
| **CoreHSEnergy** | -0.608 | -0.365 | -0.369 | -0.427 | -0.541 | -0.560 | -0.479 | 0.105 |
| **RimHSEnergy** | -0.415 | 0.020 | -0.100 | 0.000 | -0.367 | -0.329 | -0.198 | 0.194 |
| **SuppHSEnergy** | -0.153 | -0.162 | -0.617 | -0.489 | -0.385 | -0.465 | -0.379 | 0.187 |
| **CoreHS** | -0.413 | -0.281 | -0.232 | -0.476 | -0.342 | -0.440 | -0.364 | 0.095 |
| **RimHS** | -0.319 | -0.071 | -0.181 | 0.000 | -0.128 | -0.176 | -0.146 | 0.109 |
| **SuppHS** | -0.156 | -0.153 | -0.430 | -0.344 | -0.480 | -0.441 | -0.334 | 0.146 |
| **HSEner_NegCoop** | -0.487 | -0.282 | -0.475 | -0.260 | -0.414 | -0.514 | -0.405 | 0.109 |
| **HS_NegCoop** | -0.330 | 0.013 | -0.049 | -0.356 | -0.415 | -0.460 | -0.266 | 0.198 |
| **HSEner_PosCoop** | -0.278 | -0.192 | -0.218 | -0.437 | -0.573 | -0.444 | -0.357 | 0.150 |
| **HS_PosCoop** | -0.013 | -0.256 | -0.138 | -0.154 | -0.494 | -0.457 | -0.252 | 0.190 |

### 6.3.2 Hypothesis Validation Part 2 - Explaining Off-Rate Changes Using ΔΔG Energies from Single-Point Alanine Mutations

In 4.3.1 it was shown how the change in the sum of ΔΔG energies of single-point mutations to alanine pre- and post-mutation shows a significant correlation to the change in off-rate. Though using experimental off-rates, this was found to be true; there are two confounding factors that need to be addressed. Firstly, ΔΔGs are used to estimate off-rates. Secondly, single-point ΔΔGs of alanine mutations are used to estimate the off-rate of mutations to non-alanine mutations, which also include multi-point mutations (along with single-point alanine mutations).

Starting first by addressing the second point; Let us assume we have a single-point mutation at a specific residue position, from Leu to Arg, then the $\Delta k_{off}$ may be calculated as

$$\Delta k_{off}^{LEU->ARG} = k_{off}^{\mathrm{LEU}} - k_{off}^{\mathrm{ARG}} \quad 6.2$$

If we assume that most of the change in binding free energy of the mutation $\Delta\Delta G$, is a result of a change in $k_{off}$ with minimal change in $k_{on}$, then

$$\Delta k_{off}^{LEU->ARG} \alpha \Delta\Delta G^{LEU->ARG} \quad 6.3$$

Now,

$$\Delta\Delta G^{LEU->ARG} = \Delta G^{\mathrm{ARG}} - \Delta G^{\mathrm{LEU}} \quad 6.4$$

$$= [G^{\mathrm{ARG}}{}_{\mathrm{Complex}} - (G^{\mathrm{ARG}}{}_{\mathrm{Receptor}} + G^{\mathrm{ARG}}{}_{\mathrm{Ligand}})] \quad 6.5$$
$$- [G^{\mathrm{LEU}}{}_{\mathrm{Complex}} - (G^{\mathrm{LEU}}{}_{\mathrm{Receptor}} + G^{\mathrm{LEU}}{}_{\mathrm{Ligand}})]$$

$$= G^{\mathrm{ARG}}{}_{\mathrm{Complex}} - G^{\mathrm{ARG}}{}_{\mathrm{Receptor}} - G^{\mathrm{ARG}}{}_{\mathrm{Ligand}} \quad 6.6$$
$$- G^{\mathrm{LEU}}{}_{\mathrm{Complex}} + G^{\mathrm{LEU}}{}_{\mathrm{Receptor}} + G^{\mathrm{LEU}}{}_{\mathrm{Ligand}}$$

Approaching this using only ALA mutations we have,

$$\Delta\Delta G^{\mathrm{ARG}->ALA} - \Delta\Delta G^{\mathrm{LEU}->ALA} = \quad 6.7$$
$$[G^{\mathrm{ALA}}{}_{\mathrm{Complex}} - G^{\mathrm{ALA}}{}_{\mathrm{Receptor}} - G^{\mathrm{ALA}}{}_{\mathrm{Ligand}} - G^{\mathrm{ARG}}{}_{\mathrm{Complex}} + G^{\mathrm{ARG}}{}_{\mathrm{Receptor}} + G^{\mathrm{ARG}}{}_{\mathrm{Ligand}}]$$
$$- [G^{\mathrm{ALA}}{}_{\mathrm{Complex}} - G^{\mathrm{ALA}}{}_{\mathrm{Receptor}} - G^{\mathrm{ALA}}{}_{\mathrm{Ligand}} - G^{\mathrm{LEU}}{}_{\mathrm{Complex}} + G^{\mathrm{LEU}}{}_{\mathrm{Receptor}} + G^{\mathrm{LEU}}{}_{\mathrm{Ligand}}]$$

$$= - G^{\mathrm{ARG}}{}_{\mathrm{Complex}} + G^{\mathrm{ARG}}{}_{\mathrm{Receptor}} + G^{\mathrm{ARG}}{}_{\mathrm{Ligand}} + G^{\mathrm{LEU}}{}_{\mathrm{Complex}} - G^{\mathrm{LEU}}{}_{\mathrm{Receptor}} - G^{\mathrm{LEU}}{}_{\mathrm{Ligand}} \quad 6.8$$

$$= - \Delta\Delta G^{LEU->ARG} \quad 6.9$$

Therefore,

$$- ( \Delta\Delta G^{\text{ARG}->ALA} - \Delta\Delta G^{\text{LEU}->ALA} ) = \Delta\Delta G^{LEU->ARG} \alpha \Delta k_{off}^{LEU->ARG} \quad 6.10$$

Hence for the mutation Leu->Arg,

where $\Delta\Delta G^{\text{ARG}->ALA} > \Delta\Delta G^{\text{LEU}->ALA}$ , the change in off-rate is negative (i.e. the off-rate of mutant is lower than the off-rate of the wild-type), and hence the mutation is stabilising. Conversely, if $\Delta\Delta G^{\text{ARG}->ALA} < \Delta\Delta G^{\text{LEU}->ALA}$ then this results in a positive $\Delta k_{off}$ and hence the mutation is destabilizing.

The key assumption here is that the change in binding free energy is mostly reflected through a change in the off-rate rather than the on-rate (See 6.3). To validate this, for the 713 off-rate mutations used in this work, the corresponding experimental values for ΔΔG are correlated to the $\Delta\log_{10}(k_{on})$ and $\Delta\log_{10}(k_{off})$ values for the same mutations. The respective PCCs between them are shown in Table 3A.

**Table 6.2: Relationship between experimental ΔΔG, $\Delta\log_{10}(k_{off})$, $\Delta\log_{10}(k_{on})$ and change in interface hotspot energy (Int_HS_Energy) for 713 mutations in the SKEMPI database.**

(A) PCC between experimental ΔΔG with the respective $\Delta\log_{10}(k_{off})$ and $\Delta\log_{10}(k_{on})$ for single-point alanine, single-point non-alanine, multi-point and all 713 mutations. (B) PCC between Int_HS_Energy with the respective, ΔΔG with$\Delta\log_{10}(k_{off})$ and $\Delta\log_{10}(k_{on})$ for single-point, single-point non-alanine, multipoint and all 713 mutations. Experimental values for the 713 mutations used here are extracted from SKEMPI and detailed in Methods section 2.1.2.

| A | ***ΔΔG*** | **Single-point alanine** | **Single-point alanine** | **Multi-point alanine** | **All Types** |
|---|---|---|---|---|---|
| | **$\Delta\log_{10}(k_{off})$** | 0.57 | 0.92 | 0.96 | 0.83 |
| | **$\Delta\log_{10}(k_{on})$** | -0.56 | -0.65 | -0.65 | -0.60 |
| B | **Int_HS_Energy** | **Single-point alanine** | **Single-point alanine** | **Multi-point alanine** | **All Types** |
| | **$\Delta\log_{10}(k_{off})$** | -0.33 | -0.34 | -0.62 | -0.51 |
| | **$\Delta\log_{10}(k_{on})$** | 0.12 | 0.08 | 0.22 | 0.17 |
| | ***ΔΔG*** | -0.48 | -0.29 | -0.57 | -0.53 |

The correlations are calculated for single-point alanine mutations, single-point non-alanine, multi-point, and on all mutations. Namely, $\Delta\Delta G$, shows a correlation of R=0.83 with $\Delta\log_{10}(k_{off})$ and R=-0.6 with $\Delta\log_{10}(k_{on})$. More notable is that the $\Delta\Delta G$ of multi-point and a non-alanine mutation is strongly reflected through a change in $\Delta\log_{10}(k_{off})$ (R=0.96, R=0.92 respectively). Other lines of evidence also show that the change in binding free energy is largely explained through a change in off-rate; For example, mutagenesis studies (Castro and Anderson, 1996, Jin and Wells, 1994) have shown that increases in dissociation rate constants are the dominant cause for a decrease in binding affinity, and work on the related phenomenon of protein-DNA binding shows that 78% of the variance of $\log_2(k_{off})$ is explained by the variance of information of the binding site sequence as opposed to 49% of the variance of $\log_2(k_{on})$ (Shultzaberger et al., 2007). Similarly, work on the enhancement of protein-protein association rate shows that mutations that affect binding free energy, as a result of affecting the on-rate with no change in the off-rate, are found at surface-exposed sites and located at the vicinity of, but outside, the binding site - as those within the binding site are generally off-rate modulating (Kiel et al., 2004). Thus, for the 713 off-rate mutation dataset, only 25% of the mutants are located at the edges (Rim) or outside the binding site (Surface), hence it may also be expect that the larger portion of mutants in the data used in this analysis, to predominantly affect the off-rate, as is also confirmed by the correlations in Table 3A.

### 6.3.3 The Hotspot Descriptors and Hotspot Predictors

Using *Int_HS_Energy* derived from our hypothesis which links hotspot energies to off-rates, an additional 15 hotspot descriptors were designed. The motivations and calculation for each of the 16-hotspot descriptors is detailed in 2.3.7. In summary (See Table 6.3); *Int_HS_Energy*, is the difference in the sum of hotspot residue energies pre- and post-mutation. *HSEner_PosCoop* and *HSEner_NegCoop* are identical to *Int_HS_Energy* except that, in order to account for positive and negative cooperativity effects between hotspots within a hotregion, the hotspot

energies are down-weighted and up-weighted accordingly to the size of hotregion they are in. *CoreHSEnergy*, *RimHSEnergy* and *SuppHSEnergy*, are

**Table 6.3: Summary of Hotspot Descriptors.**
The functional form of each hotspot descriptor and the motivations behind its design, are detailed in the methods sections of 2.3.7.

| Hotspot Descriptor | Description |
|---|---|
| **Int_Energy_1** | Change in Total Interface $\Delta\Delta G_{ALA}$ Energy |
| **Int_HS_Energy** | Change in Total Interface $\Delta\Delta G_{ALA}$ Energy of Hotspots |
| **No_HS** | Change in Number of Hotspots |
| **No_Clusters** | Change in Number of Unique Hotregions |
| **MaxClusterSize** | Change in Number of Hotspots in Largest Hotregion |
| **AVG_HS_PathLength** | Change in Hotspot Coverage |
| **CoreHSEnergy** | Change in Total $\Delta\Delta G_{ALA}$ Energy of Hotspots in Core Region |
| **CoreHS** | Change in Number of Hotspots in Core Region |
| **RimHSEnergy** | Change in Total $\Delta\Delta G_{ALA}$ Energy of Hotspots in Rim Region |
| **RimHS** | Change in Number of Hotspots in Rim Region |
| **SuppHSEnergy** | Change in Total $\Delta\Delta G_{ALA}$ Energy of Hotspots in Support Region |
| **SuppHS** | Change in Number of Hotspots in Support Region |
| **HSEner_PosCoop** | Change in Total Hotspot $\Delta\Delta G_{ALA}$ Energy Accounting for Positive Cooperativity in Hotregions |
| **HS_PosCoop** | Change in Hotspot Counts Accounting for Positive Cooperativity in Hotregions |
| **HSEner_NegCoop** | Change in Total Hotspot $\Delta\Delta G_{ALA}$ Energy Accounting for Negative Cooperativity in Hotregions |
| **HS_NegCoop** | Change in Hotspot Counts Accounting for Negative Cooperativity in Hotregions |

similar to *Int_HS_Energy*, except that changes in hotspot energies are limited to the given region on the interface. Each of the six descriptors also have their coarse-grain counterparts (*No_HS*, *HS_PosCoop*, *HSNegCoop*, *CoreHS*, *RimHS* and *SuppHS*), where only hotspot counts instead of energies are used in the calculations. Other hotspot descriptors include the change in the size of the largest hotregion (*MaxClusterSize*), the number of hotregions (*No_Clusters*), the spread of the hotspots at the interface (*AVG_HS_PathLength*) and *Int_Energy_1* that sums the changes of all single-point alanine mutations at the interface.

A number of hotspot predictors are investigated for the generation of hotspot descriptors, and in total, six sets of hotspot descriptors are generated. These

include hotspot descriptors generated from available hotspot prediction servers, *KFC2a*, *KFC2b* (Zhu and Mitchell, 2011), *RFHotpoint1* and *RFHotpoint2* (Tuncbag et al., 2010), along with the hotspot descriptors generated from hotspot prediction algorithms developed in this work (*RFSpot*, *RFSpot_KFC2*). Performance comparisons of the hotspot prediction algorithms can be found in 5.2. The use of multiple hotspot predictors enables us to probe consistencies and anomalies in the predictive abilities of the hotspot descriptors.

### 6.3.4 Comparison of Hotspot Descriptors with Molecular Descriptors

The PCCs achieved by *Int_HS_Energy* and the rest of the hotspot descriptors with $\Delta\log_{10}(k_{off})$, is compared to those achieved by the benchmark set of molecular descriptors (See Table 2.1 in Methods section 2.2 for the full list). The molecular descriptor set consists of a complex and comprehensive set of 110 structure-related descriptors characterizing various aspects of protein-protein interactions and their energetics. For the hotspot descriptors, 16 hotspot descriptors as generated by each of the six-hotspot predictors are assessed. With 713 mutations in the dataset, all absolute correlations of $|R|>0.1$ are significant with $p<0.001$. Figure 6.2(A) shows all descriptors (Hotspot descriptors in green superimposed on the molecular descriptors in black) ranked according to their absolute PCCs. Consistently higher PCC values are observed for the hotspot descriptors. It should be noted that here one is comparing the raw predictive power of each descriptor in estimating $\Delta\log_{10}(k_{off})$; this is independent of any learning models trained on $\Delta\log_{10}(k_{off})$ data.

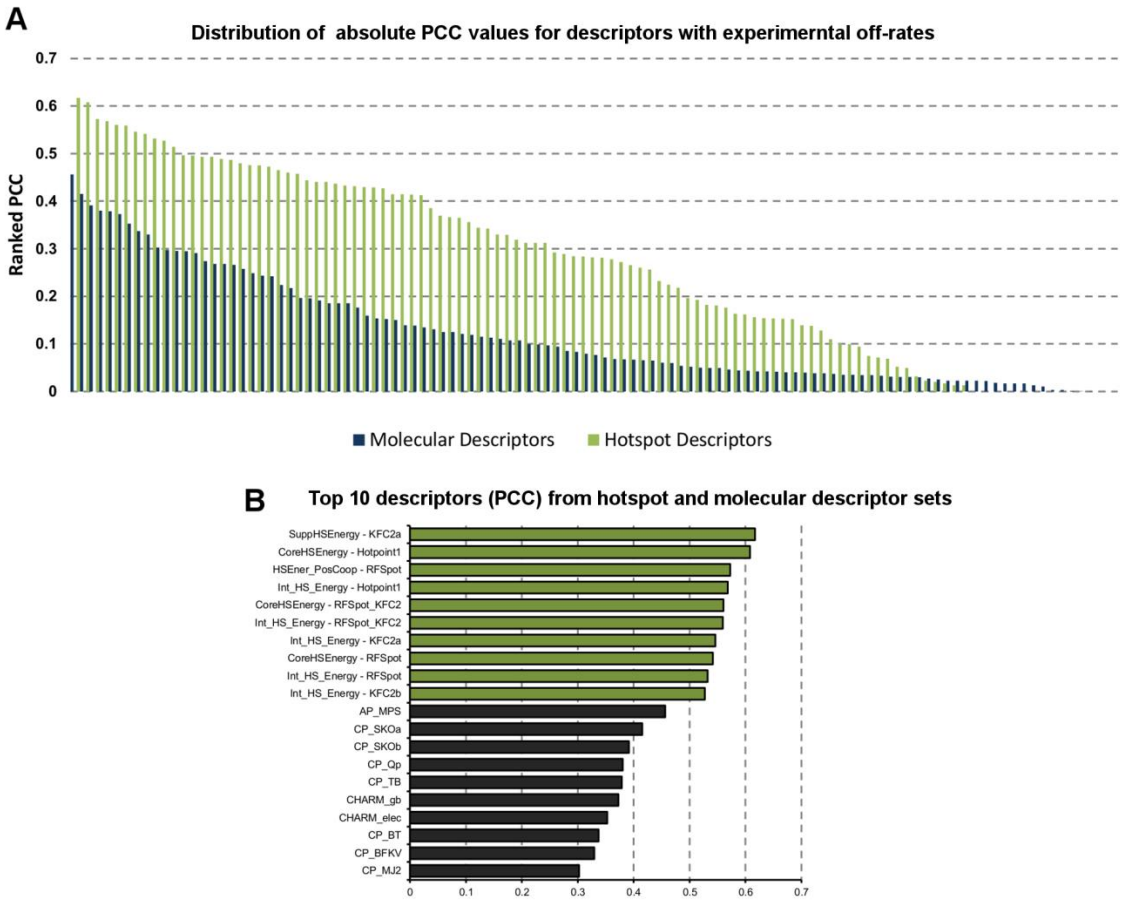



**Figure 6.2: Hotspot and molecular descriptors for estimating change in off-rate using PCC.**

The hotspot descriptors designed in this work are benchmarked against a set of 110 molecular descriptors; both in their ability to estimate $\Delta\log_{10}(k_{off})$ and in their ability to detect stabilizing mutations of $\Delta\log_{10}(k_{off}) < -1$. The performance measures shown here enable us to assess the raw predictive power of the descriptors independent of any learning models. Green and black bars highlight descriptors from the hotspot and molecular descriptor sets respectively. (A) Comparison of the distribution of the absolute PCC values for the hotspot descriptors designed in this work against that for the molecular descriptors. (B) The top ten hotspot descriptors (green) followed by the top ten molecular descriptors (black), ranked according to the PCC. The highest ranked descriptors all relate to energetic changes in hotspots suggesting that changes in hotspot counts is not sufficient to characterise changes in off-rate. The most common of which are the *Int_HS_Energy* and *CoreHSEnergy*, where the latter only considers changes in hotspot energies in the core region of protein-protein interfaces. Apart from the DARS atomic potential, *AP_MPS* (Chuang et al., 2008), designed for protein-protein docking with |R|=0.46, the top 10 molecular descriptors are dominated by coarse-grain statistical potentials. The bias toward coarse-grain potentials is similar to that observed previously (Moal and Bates, 2012) where

models for *wild-type* off-rates were built. The correlation power of the molecular descriptors decreases markedly down to |R|=0.3 at just the 10th ranked molecular descriptor, in contrast to the hotspot descriptors where |R|=0.5.

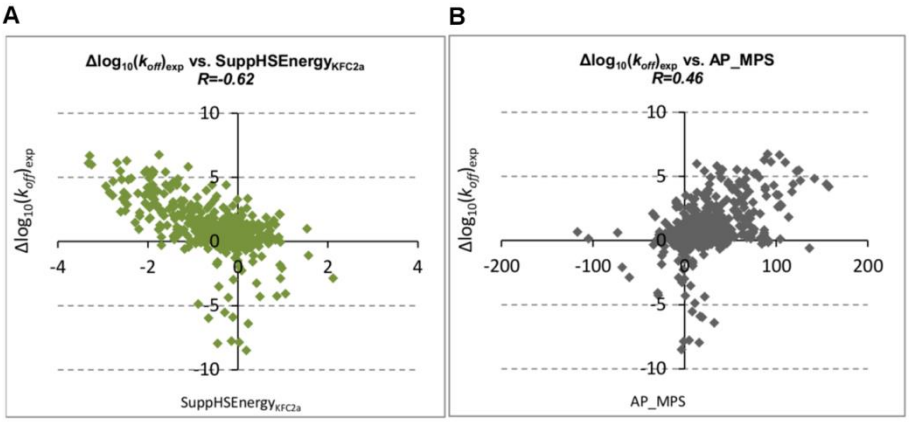



**Figure 6.3: Scatter plots of best performing hotspot and molecular descriptors according to PCC.**
The relationship between experimental values for $\Delta\log_{10}(k_{off})$ and (A) hotspot descriptors showing highest correlation with $\Delta\log_{10}(k_{off})$ ($SuppHSEnergy_{KFC2a}$ - changes in hotspot energies in the support region as predicted by KFC2a (Zhu and Mitchell, 2011)) and (B) molecular descriptor showing highest correlation with $\Delta\log_{10}(k_{off})$ (AP_MPS - the DARS atomic potential (Chuang et al., 2008)).

Scatter plots of the top performing (according to PCC) hotspot descriptor $SuppHSEnergy_{KFC2a}$ (R=-0.62) and the top performing molecular descriptor *AP_MPS (R=0.46.)* are shown in Figure 6.3. These descriptors are the best from their set at globally estimating the changes in off-rate, which as observed in Figure 6.3. Their accuracy mostly stems from their ability to model the destabilizing (off-rate increasing) portion of the dataset. The underestimation of stabilising (off-rate decreasing) mutations and descriptors, which are better able to detect such mutations, is investigated in subsequent sections.

### 6.3.5 Detection of Complex Stabilizing Mutations

To assess the discriminatory power of the hotspot and molecular descriptors, the dataset is partitioned into ($\Delta\log_{10}(k_{off}) < -1$), representing the stabilizing portion of the dataset, and ($\Delta\log_{10}(k_{off}) > 0$), representing the neutral to destabilizing portion of the dataset (referred to as CDS1 – Classification Dataset 1). Another dataset, which removes the neutral mutations as detailed in 2.1.3, is also used (referred to as CDS2). For an unbiased assessment of descriptor discrimination ability, two discrimination performance measures are calculated; the Mann Whitney U-Test (Figure 6.4 C,E), the MCC (Figure 6.4 D,F) – both of which are described in methods section 2.4. Similar to the correlations with $\Delta\log_{10}(k_{off})$ (i.e Figure 6.2 A,B), the changes in hotspot descriptors show better discrimination abilities than changes in molecular descriptors (Figure 6.4 C-F). This confirms that the hotspot descriptors, as well as possessing fine-grain predictive ability, also possess coarse-grain predictive ability. For example, the most discriminatory hotspot descriptor under the U-Test, achieves a 4-fold (CDS1) and 5-fold (CDS2) increase in discriminatory power over the most discriminatory molecular descriptor. Once again, those hotspot descriptors that use the energies of hotspots, as opposed to counts, dominate.

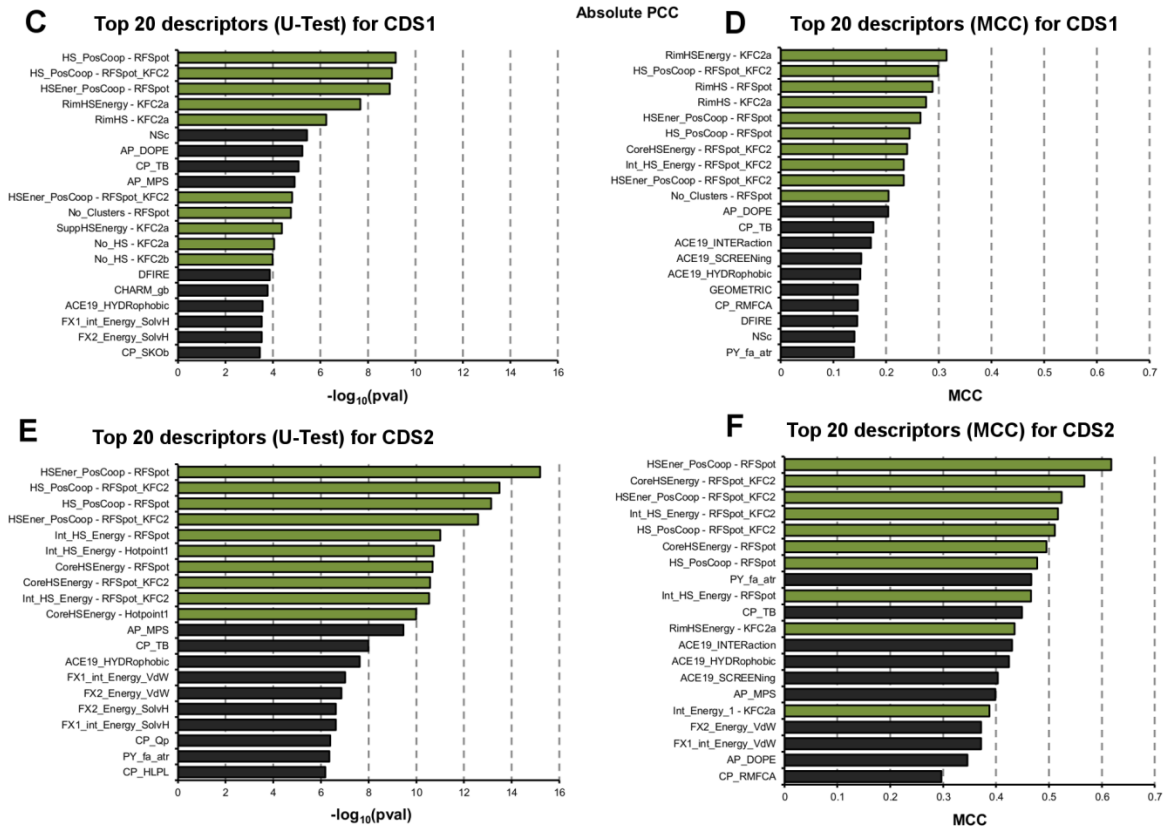



**Figure 6.4: Hotspot and molecular descriptors for estimating change in off-rate using the MCC/U-Test.**

The hotspot descriptors designed in this work are benchmarked against a set of 110 molecular descriptors; both in their ability to estimate $\Delta\log_{10}(k_{off})$ and in their ability to detect stabilizing mutations of $\Delta\log_{10}(k_{off}) < -1$. The performance measures shown here enable assessment of the raw predictive power of the descriptors independent of any learning models. Green and black bars highlight descriptors from the hotspot and molecular descriptor sets respectively. (C) Mann Whitney U-Test rankings for all descriptors where values are ranked according to $-\log_{10}$ (pval) and represent the discrimination ability of the descriptors for the detection of stabilizing mutants ($\Delta\log_{10}(k_{off}) < -1$) from neutral to destabilizing mutants ($\Delta\log_{10}(k_{off}) > 0$) (Referred to as CDS1). This dataset contains 31 stabilizing mutants and 503 neutral to destabilizing mutants. (D) Matthew's Correlation Coefficient (MCC) rankings for all descriptors on same dataset. (E) and (F) are identical to (C) and (D) except that results are for off-rates that satisfy $|\Delta\log_{10}(k_{off})| > 1$. This dataset contains 31 stabilizing mutants and 213 destabilizing mutants (referred to as CDS2).

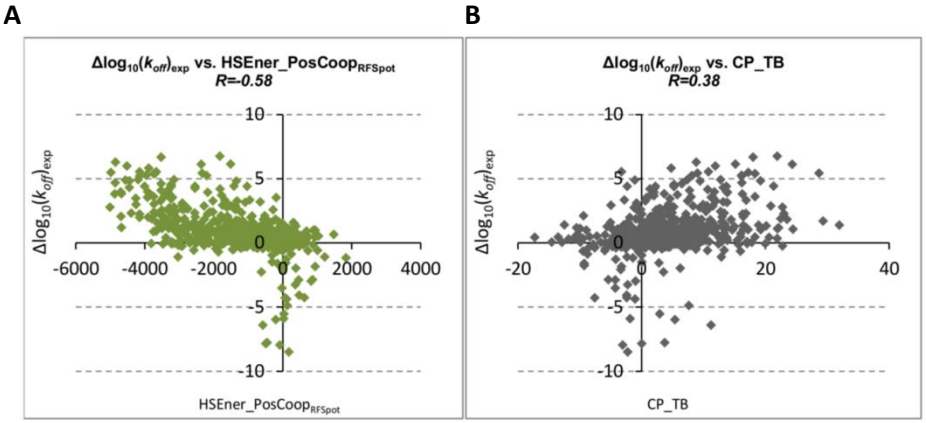



**Figure 6.5: Scatter plots of best performing hotspot and molecular descriptors according to MCC.**
The relationship between experimental values for $\Delta\log_{10}(k_{off})$ and (A) top performing hotspot descriptor for the detection of stabilizing mutants ($HSEner\_PosCoop_{RFSpot}$ – changes in hotspot energies on accounting for positive cooperativity in hotregions) and (B) top performing molecular descriptor for the detection of stabilizing mutants (CP_TB – coarse grained protein-protein docking potential).

Scatter plots of a representative hotspot (*HSEnerPosCoop*$_{RFspot}$) and molecular descriptor (*CP_T*B) (Tobi, 2010), are shown in Figure 6.5 C,D. These descriptors do well on both CDS1 and CDS2, though they still show a tendency to underestimate stabilizing mutations. For both CDS1 and CDS2, the positive cooperativity descriptors *HSEner_PosCoop*/*HS_PosCoop* dominate the ranked lists (Figure 6.4 C-F) and RimHSEnergy/RimHS for CDS1 (Figure 6.4 D). For example, *HSEner_PosCoop*$_{RFSpot}$ achieves a TPR/FPR/MCC of 0.58/0.05/0.62 for the detection of stabilizing mutants on CDS2. Given that *HSEner_PosCoop*$_{RFSpot}$ supersedes *Int_HS_Energy* (additivity within hotregions assumption) and *HSEner_NegCoop* (negative cooperativity within hotregions assumption), applying the general assumption of positive cooperativity between hotspots within a hotregion, and accounting for it, provides higher detection rates of stabilizing mutations (*i.e.* $\Delta\log_{10}(k_{off}) < -1$). It should be noted however, that out of the three-hotspot predictors, which generate the most discriminatory hotspot descriptors (i.e. *RFSpot*, *RFSpot_KFC2* and *KFC2a)*, the positive cooperativity descriptors which show high discrimination abilities, are limited to those generated by *RFSpot* and *RFSpot_KFC2*. The relationship of $\Delta\log_{10}(k_{off})$ and

cooperative effects within hotregions is discussed further in the subsequent section 8.3 of Chapter 8. An interesting observation is that whereas for a good global correlation, the coarse-grain statistical potential are preferred to their atom-based counterparts, this is not the case for the detection of stabilizing mutations. Here, the physics-based energetic terms play a more important role.

## 6.4 Discussion

In this chapter, it is proposed that the energetics and architecture of hotspots on protein-protein interfaces can be used to estimate changes in off-rate. More specifically, the change in the sum of hotspot energies at an interface pre- and post-mutation correlates to a change in off-rate brought about by that mutation. Using a dataset of 713 mutations with experimental off-rates, correlations as a high as R=0.62 are observed for hotspot descriptors based on this proposition. More importantly, such hotspot descriptors perform consistently better than the more traditional molecular descriptors.

This begs the question, why do hotspot descriptors dominate over molecular descriptors? To gain insight into why this is so, the key differences between hotspot descriptors and molecular descriptors are discussed. Essentially, a hotspot is the realisation of an optimal 'state' of several energetic factors. A state here represents the residue in question and its context i.e. surrounding residues, solvent accessibility etc. For an optimal state, a number of energetic contributors favourably combine. Naturally, one would expect that changes to these states (as represented by changes caused to hotspot residues) would have a more correlative effect to the off-rate than molecular descriptors, which consider a very specific energetic contribution. With this in mind, several of the molecular descriptors are statistical potentials, which themselves implicitly encompass a number of energetic contributions. Hence synergy alone is not sufficient to explain the better performance of hotspot descriptors over molecular descriptors.

Another aspect of hotspot descriptors, which distinguished them from molecular descriptors, is the way they are calculated. Take for example the electrostatic interaction between all pairs of contacts;

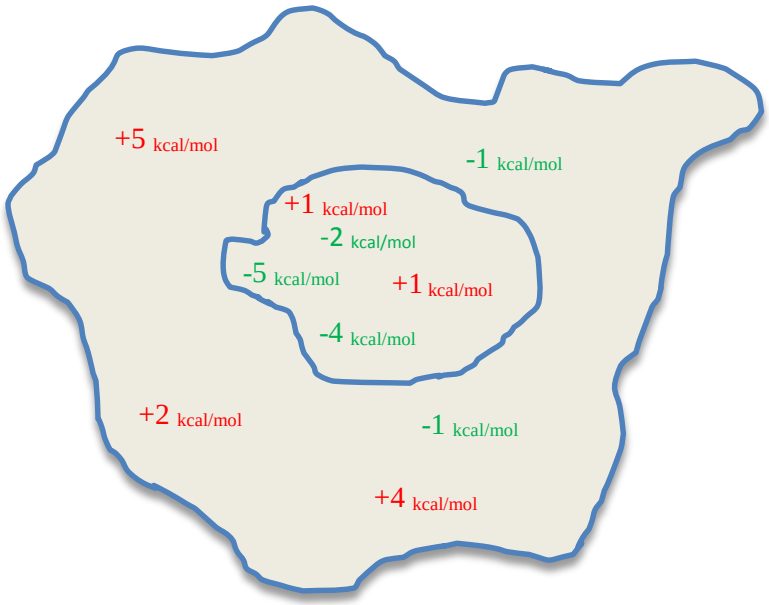



**Figure 6.6: Distribution of energy across a protein-protein interface. Favourable interactions across a complex interface are not distributed homogenously.**

Molecular descriptors traditionally perform a summation over all favourable and unfavourable contributions that in turn result in an averaging out effect. Effectively, any positional of favourable interactions is lost. Hotspot descriptors, on the other hand, consider the contribution of only a subset of residues or those in specific regions of the interface. Hence the averaging out effect is less prominent. If distribution and not only the sum of favourable interactions contribute to complex stability, then hotspot descriptors may indeed capture effects critical for the stability of a complex that traditional molecular descriptors will miss.

Once the summation is performed over all pairs, the resulting effect is essentially an averaging out effect of favourable and unfavourable interactions at the interface. Therefore, any positional information is lost. The cost of this loss of information, if any, is explained using a toy scenario, Figure 6.6. Let's assume *EnergyX* is an enthalpic contributor, important for the stability of the complex. In this example, *EnergyX's* total energetic contribution after summation is neutral at

0 kcal/mol. Before summation, one could appreciate a central core of favourable energy (-9 kcal/mol) surrounded by less favourable interactions distributed sparsely in the outer shell. Does this central core of favourable energy override the global neutrality of the total energetic sum? Whether or not this is so, the summation of energetic contribution across all contacts of an interface does not allow for this effect to be accounted for. In contrast, the hotspot descriptors are not a sum over all residues of the interface. *Int_HS_Energy* (mean PCC of *R=-0.51 with* $\Delta log_{10}(k_{off})$) for example, only sums the energies of hotspots – hence ignoring the energetic contribution of the majority of the residues. Interestingly, when one considers the energetic contribution of all interface residues, and not only hotspots, the signal degrades, where *Int_Energy* has a mean PCC of *R=-0.33*. Therefore, any signal observed if one only considers the energies of hotspot residues, is lost after including the energies of the remaining non-hotspot interface residues. Hotspot descriptors, such as *Core_HS_Energy*, are even more selective concerning which residues they score. In this case, only the core region hotspots are considered yet a mean PCC of *R=-0.48* and a maximum of *R=-0.61* is achieved. Even though, on average some signal is lost when not considering hotspots outside the core region, one must keep in mind that its maximum *R=-0.61,* is still significantly higher to that achieved by the best preforming molecular descriptor *AP_MPS (R=0.46)* with p*<<0.001.*

# Chapter 7

# 7 Prediction of Off-Rate Changes upon Mutation Using Machine Learning Models and Hotspot Descriptors

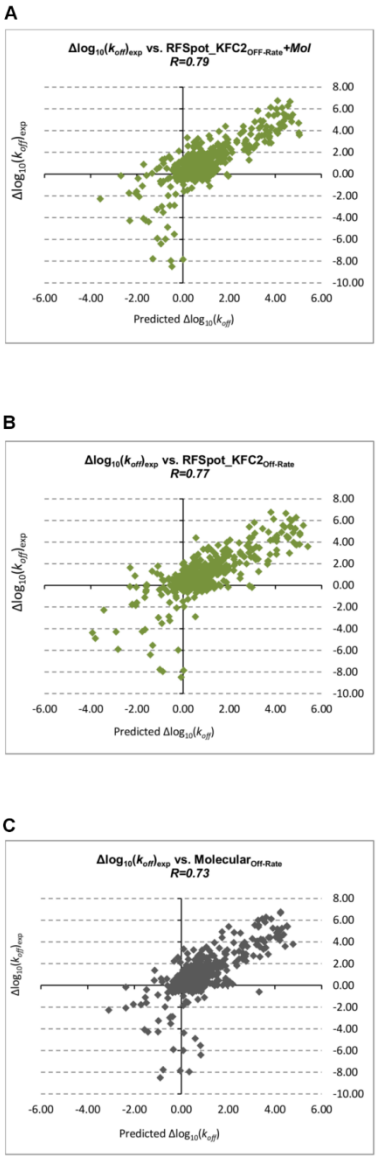
In Chapter 6 it was described how changes in the hotspot energy landscape can be transformed into hotspot descriptors, which are able to estimate changes in the off-rate. Most importantly, the individual power of hotspots surpasses that of standard physics-based energetic terms statistical potentials. In this chapter, the aforementioned descriptors are combined using machine learning models to achieve even higher predictive performance in the prediction of off-rate change upon mutation. In section 7.1 several regression models using both hotspot and molecular descriptors are built and the detection of rare residence-time increasing (off-rate decreasing) mutations is investigated using a number of classification models in 7.2. In section 7.3, the 713 mutations

in the off-rate dataset are categorized into what may be termed, data regions. Such data regions represent mutations that have a common physical property, or come from a similar type of complex or region at the interface. The effects of mutations within a particular data region might be more or less difficult to predict than mutations within another. Therefore, data region analysis enables us to identify current strengths in the prediction of off-rates and conversely, mutations, which are consistently harder to characterise. The effects of conformational changes and cross-validation routines are presented in the remaining sections.

## 7.1 Off-Rate Prediction Using Machine Learning Models with Hotspot and Molecular Descriptors

Confirming that energy estimates of single point-alanine mutations can be used to describe the effects of off-rate changes of single- and multi-point mutations not limited to alanine, it is next assessed whether the whole set of 16 hotspot descriptors, from each hotspot prediction algorithm, can be combined synergistically in a model for off-rate prediction to achieve even higher correlations. A separate Random Forest (RF) regression model is trained on the 713 off-rate mutant dataset using the hotspot descriptors generated by each hotspot predictor ($RFSpot_{Off\text{-}Rate}$, $RFSpot\_KFC2_{Off\text{-}Rate}$, $RFHotpoint1_{Off\text{-}Rate}$, $RFHotpoint2_{Off\text{-}Rate}$, $KFC2a_{Off\text{-}Rate}$ and $KFC2b_{Off\text{-}Rate}$). In addition, models that add the set of 110 molecular descriptors to the hotspot descriptors ($RFSpot+Mol_{Off\text{-}Rate}$, $RFSpot\_KFC2+Mol_{Off\text{-}Rate}$, $RFHotpoint1+Mol_{Off\text{-}Rate}$, $RFHotpoint2+Mol_{Off\text{-}Rate}$, $KFC2a+Mol_{Off\text{-}Rate}$ and $KFC2b+Mol_{Off\text{-}Rate}$) are also built for comparison. Note that the *Off-Rate* subscript is used to distinguish the off-rate predictor trained on hotspots, from the actual hotspot predictor generating the hotspot descriptors in question. The 20-fold cross-validation (CV) results are concatenated to form of a set of 713 test predictions and their PCC with $\Delta\log_{10}(k_{off})$ calculated and presented in Table 7.1.

**Table 7.1 PCC values of off-rate regression models.**
PCC values are generated using the experimental values for Δlog10(koff) from 713 off-rate mutations in the SKEMPI database and the predicted values for Δlog10(koff) from the regression models using 20-Fold CV. CP – Coarse-Grain Potentials, AP – Atomic-Based Potentials, CP-AP, All statistical Potentials, PB – Physics based Energy Terms.




| Hotspot Descriptor + Molecular Descriptor Models | |
|---|---|
| **Model** | **PCC** |
| *RFSpot*$_{Off\text{-}Rate}$, + MOL | 0.78 |
| *RFSpot_KFC2*$_{Off\text{-}Rate}$ + MOL | 0.79 |
| *RFHotpoint1*$_{Off\text{-}Rate}$ + MOL | 0.77 |
| *RFHotpoint2*$_{Off\text{-}Rate}$ + MOL | 0.75 |
| *KFC2a*$_{Off\text{-}Rate}$ + MOL | 0.74 |
| *KFC2b*$_{Off\text{-}Rate}$ + MOL | 0.75 |
| **Hotspot Descriptor Models** | |
| **Model** | **PCC** |
| *RFSpot*$_{Off\text{-}Rate}$, | 0.74 |
| *RFSpot_KFC2*$_{Off\text{-}Rate}$ | 0.77 |
| *RFHotpoint1*$_{Off\text{-}Rate}$ | 0.73 |
| *RFHotpoint2*$_{Off\text{-}Rate}$ | 0.7 |
| *KFC2a*$_{Off\text{-}Rate}$ | 0.7 |
| *KFC2b*$_{Off\text{-}Rate}$ | 0.71 |
| **Molecular Descriptor Models** | |
| **Model** | **PCC** |
| *CP*$_{Off\text{-}Rate}$ | 0.68 |
| *AP*$_{Off\text{-}Rate}$ | 0.61 |
| *CP_AP*$_{Off\text{-}Rate}$ | 0.69 |
| *PB*$_{Off\text{-}Rate}$ | 0.72 |
| *Molecular*$_{Off\text{-}Rate}$ | 0.73 |

**Figure 7.1: Scatter plots of best performing off-rate regression models.**
The relationship between experimental values for $\Delta\log_{10}(k_{off})$ and predicted values for $\Delta\log_{10}(k_{off})$ with (A) RFSpot_KFC2$_{Off\text{-}Rate}$+Mol, the best performing off-rate prediction model combining hotspot and molecular descriptors. Hotspot descriptors for this model are generated using the RFSpot_KFC2 hotspot prediction algorithm. (B) RFSpot_KFC2$_{Off\text{-}Rate}$+Mol, the best performing off-rate

prediction model using only hotspot descriptors. Hotspot descriptors for this model are again generated using the RFSpot_KFC2 hotspot prediction algorithm. (C) Molecular$_{\text{Off-Rate}}$, off-rate prediction model using molecular descriptors. The addition of hotspot descriptors as observed in (A) compared to the molecular descriptor model only as shown in (B) notably improves the prediction of stabilizing mutants, which are all found in the lower left quadrant for *RFSpotKFC2$_{Off\text{-}Rate}$+Mol*.

Figure 7.1, shows the best performing off-rate models from each class. The class here refers to the type of features used for off-rate prediction – molecular descriptors only (Figure 7.1c R=0.73), hotspot descriptors only (Figure 7.1b R=0.77 and $p<0.05$ to R=0.73), hotspot and molecular descriptors (Figure 7.1a R=0.79 and $p<0.005$ to R=0.73). Besides exhibiting higher correlations, off-rate models using hotspot descriptors (Figure 7.1a and Figure 7.1b), show fewer mutations in the lower right quadrant, and hence better at identifying off-rate decreasing mutations than a model using molecular descriptors (Figure 7.1c). This is investigated more specifically in 7.2 using classification models and performance measures. The performance of models created from different categories of molecular descriptors is also investigated (See Table 7.1 ). These include Atomic Potentials (AP), Coarse-grain Potentials (CP) and Physics-Based energy terms (PB). The physics-based descriptors model (PB$_{Off\text{-}Rate}$, R=0.72) which include CHARMM (Brooks et al., 2009), FoldX (Schymkowitz et al., 2005) and PyRosetta (Chaudhury et al., 2010) energy terms performs better than the coarse-grain (CP$_{Off\text{-}Rate}$, R=0.68) and atomic (AP$_{Off\text{-}Rate}$, R=0.61) statistical potentials alone or combined (CP_AP$_{Off\text{-}Rate}$, R=0.69). *RFSpot_KFC2$_{Off\text{-}Rate}$* (R=0.77) built on hotspot descriptors only, achieves higher PCC than a model with all molecular descriptors combined (CP_AP_PB$_{Off\text{-}Rate}$, R=0.72), whereas the highest correlation is still achieved when combining both molecular and hotspot descriptors (*RFSpot_KFC2+Mol$_{Off\text{-}Rate}$*, R=0.79), as previously highlighted.

## 7.2 Prediction of Stabilizing Mutations

Similar to the regression Random Forest models, several Random Forest classification models are also built for the detection of stabilizing (*i.e.* $\Delta\log_{10}(k_{off}) < -1$) mutants and models trained and tested on both Classifier Dataset 1 (CDS1) and Classifier Dataset 2 (CDS2) using 20-Fold CV. Comparison of the MCCs achieved on CDS1 and CDS2 shows that the ability to detect stabilizing mutants is diminished when neutral mutations are present. The highest MCC achieved for CDS1 is achieved by $RFSpot\_KFC2_{Off\text{-}RateC}$ (MCC=0.60, TPR=0.45, FPR=0.01) and $RFSpot{+}Mol_{Off\text{-}RateC}$ for CDS2 (MCC=0.82, TPR=0.84, FPR=0.02).

**Table 7.2: MCC values of off-rate classification models.**

MCC values for off-rate classifier model predictions for classification data sets CDS1 and CDS2. CDS1 includes neutral mutations whereas CDS2 excludes neutral mutations; hence the detection of stabilizing mutants is enhanced in the latter, though results for CDS1 are more relevant for interface design scenarios. All results are based on 20-fold CV. CP – Coarse-Grain Potentials, AP – Atomic-Based Potentials, CP-AP, All statistical Potentials, PB – Physics based Energy Terms.

| **Hotspot + Molecular Descriptor Models** | | |
|---|---|---|
| **Model** | **MCC CDS1** | **MCC CDS2** |
| $RFSpot_{Off\text{-}RateC}$, + MOL | 0.56 | 0.82 |
| $RFSpot\_KFC2_{Off\text{-}RateC}$ + MOL | 0.58 | 0.72 |
| $RFHotpoint1_{Off\text{-}RateC}$ + MOL | 0.53 | 0.55 |
| $RFHotpoint2_{Off\text{-}RateC}$ + MOL | 0.56 | 0.63 |
| $KFC2a_{Off\text{-}RateC}$ + MOL | 0.44 | 0.6 |
| $KFC2b_{Off\text{-}RateC}$ + MOL | 0.5 | 0.63 |
| **Hotspot Descriptor Models** | | |
| **Model** | **MCC CDS1** | **MCC CDS2** |
| $RFSpot_{Off\text{-}RateC}$ | 0.53 | 0.73 |
| $RFSpot\_KFC2_{Off\text{-}RateC}$ | 0.6 | 0.79 |
| $RFHotpoint1_{Off\text{-}RateC}$ | 0.3 | 0.5 |
| $RFHotpoint2_{Off\text{-}RateC}$ | 0.53 | 0.62 |
| $KFC2a_{Off\text{-}RateC}$ | 0.4 | 0.66 |
| $KFC2b_{Off\text{-}RateC}$ | 0.43 | 0.55 |
| **Molecular Descriptor Models** | | |
| **Model** | **MCC CDS1** | **MCC CDS2** |
| $CP_{Off\text{-}RateC}$ | 0.43 | 0.54 |
| $AP_{Off\text{-}RateC}$ | 0.35 | 0.43 |
| $CP\_AP_{Off\text{-}RateC}$ | 0.5 | 0.51 |
| $PB_{Off\text{-}RateC}$ | 0.4 | 0.53 |
| $Molecular_{Off\text{-}RateC}$ | 0.53 | 0.68 |

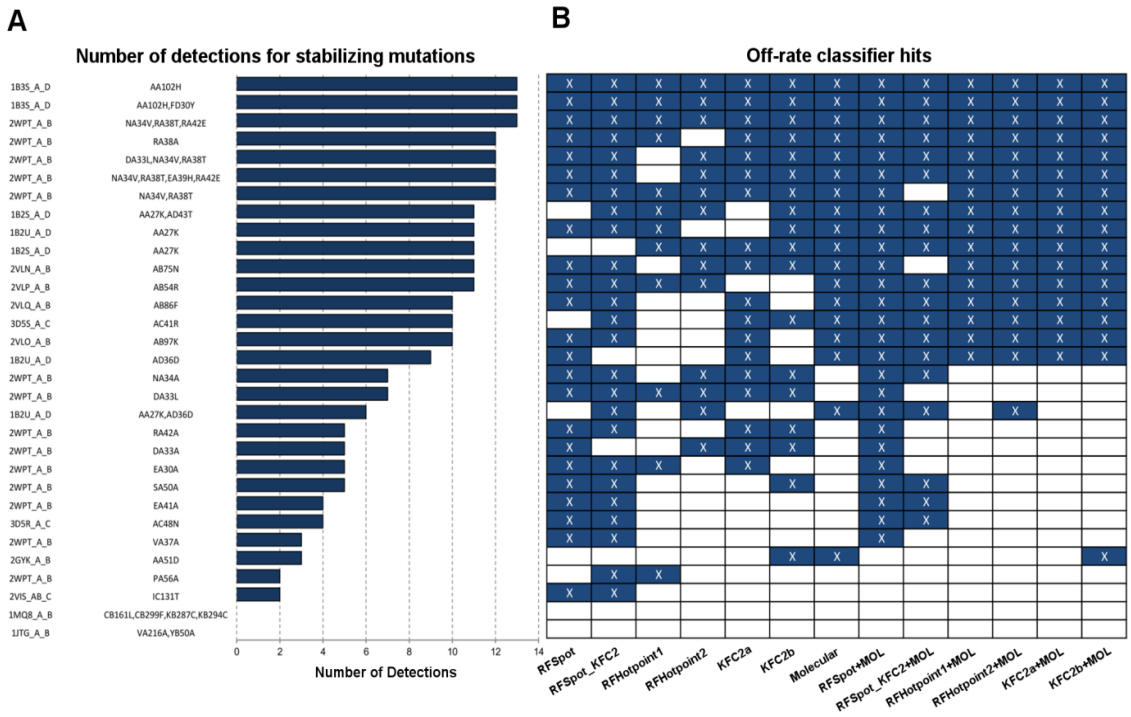



**Figure 7.2: Detection of rare complex stabilizing mutations using off-rate classification models.**

(A) Ranked list of 31 stabilizing mutations ($\Delta\log_{10}(k_{off})$ <-1) in the SKEMPI off-rate dataset. The list is ranked according to the number of off-rate prediction classification models that detect the mutation in question as stabilizing. Detections per model (B) are highlighted with white cross in a blue box, and non-detections highlighted with a white box. The lower portion of (A) is dominated by single-point mutations to alanine residues, which suggests that the stabilizing effects of these mutations, as opposed to their more common neutralizing/destabilizing effects, are much harder to characterise.

Figure 7.2 shows the list of 31 stabilizing mutants ($\Delta\log_{10}(k_{off})$ < -1) sorted according to the number of classifiers that detect the given mutation as stabilizing. Of particular interest are those stabilizing mutations that go undetected, and therefore only data from CDS2 is used as all mutations undetected in CDS2 were also undetected in CDS1 (though not the contrary). Two stabilizing mutants go undetected by of the all the predictors, namely the double alanine mutant VA216A-YB50A for protein complex 1JTG (RSCB protein data bank code) and the 4-point mutant CB161L-CB299F-KB287C-KB294C for protein complex 1MQ8. The mutations, which tend to be undetected by most of

the classifiers, generally involve a mutation to an alanine residue. Alanine mutations are generally neutral in their ability to destabilize a complex. For the rare situation in which an alanine mutation actually stabilizes a complex interface the likely interpretation is that several side-chains may sometimes hinder binding (Clackson et al., 1998, Cunningham and Wells, 1989). For example several alanine-shaving experiments show an increase in the binding affinity and an octa-alanine mutant of hGH, binding hGHbp 50-fold times tighter than the wild-type (Clackson et al., 1998, Cunningham and Wells, 1989).

## 7.3 Prediction Patterns and Data Region Analysis

So far it is observed that an off-rate prediction model with hotspot descriptors does better than one with molecular descriptors and the highest performance is achieved when combining both sets of descriptors. As a consequence, several questions come to mind; for which mutation types are the hotspot descriptor off-rate models achieving an enhanced performance compared with molecular descriptor off rate models? What information is gained by adding the molecular descriptors to the hotspot descriptor off-rate models? Is there any orthogonal information that the hotspot and molecular descriptors are capturing?

To address these questions, the performance of the off-rate models is assessed at subsets of the dataset, termed 'Data Regions'. The data regions include mutations at the core/rim/support (CORE/RIM/SUPP) regions; mutations on complexes with large/small interface areas (LIA/SIA); mutations to alanine (ALA) and non-alanine (N_ALA); mutations to polar/hydrophobic/charged (POL/HYD/CHARG) residues and finally single-point (SP) mutations as well as multi-point (MP) mutations.

### 7.3.1 Orthogonal Information Content in Hotspot and Molecular Descriptor Models

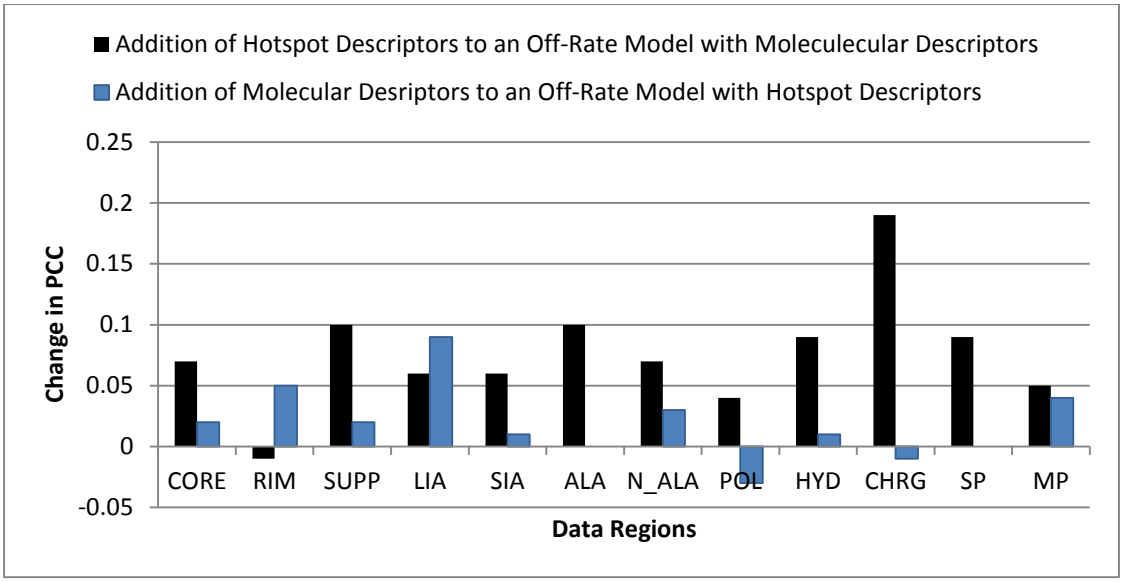



**Figure 7.3: Orthogonal information content in hotspot and molecular descriptor models.**

The 713 off-rate data set is divided into data regions where the performance of the off-rate regression models can be assessed on the subset of mutations that are in the given region. The data regions include mutations at the core/rim/support (CORE/RIM/SUPP) regions; mutations on complexes with large/small interface areas (LIA/SIA); mutations to alanine (ALA) and non-alanine (N_ALA); mutations to polar/hydrophobic/charged (POL/HYD/CHARG) residues and finally single-point (SP) mutations as well as multi-point (MP) mutations. Black bars indicate the change in PCC for the given data region when adding hotspot descriptors to a molecular descriptor off-rate model. Conversely, blue bars represent the change in PCC for the given data region when adding molecular descriptors to a hotspot descriptor an off-rate model.

Keeping in mind that Molecular$_{Off\text{-}Rate}$ is a model trained only using molecular descriptors, and RFSpot_KFC2$_{Off\text{-}Rate}$ is one trained using only hotspot descriptors, the model which combines both hotspot descriptors and molecular descriptors (RFSpot_KFC2$_{Off\text{-}Rate}$+Mol) can be assessed in two ways – the improvement in correlation achieved by adding hotspot descriptors to Molecular$_{Off\text{-}Rate}$ (Figure 7.3 black bars) or vice-verse, the improvement in correlation achieved by adding molecular descriptors to RFSpot_KFC2$_{Off\text{-}Rate}$ (Figure 7.3 blue bars). The magnitude of each positive change indicates the extent that the addition of the descriptor adds new (or rather orthogonal)

information to the model, which is then exploited by the learning model. This is what is observed for ten of the twelve data regions when hotspot descriptors are added to Molecular$_{Off\text{-}Rate}$ (as shown by the positive black bars in Figure 7.3). On the other hand, no and minimal change (characteristic of the blue bars), suggest that the addition of the molecular descriptors does not provide any new information that was not available in the existing hotspot descriptor model, or in some situations, even compromise the accuracy of the hotspot descriptor model by contributing to noise in model output as a consequence of being a weak descriptor.

### 7.3.2 Accurate and Weak Regions of Accuracy in the Prediction of Off-Rates Using Data Regions

The performance of regression models trained on the whole 713 off-rate dataset is assessed by calculating the PCC on the mutations of each data-region. Figure 7.4 shows, for each data region, a heatmap with these PCCs. From this it can be observed:

1. Rim regions are poorly characterised: All off-rate predictors obtain good correlation for core mutations, less so for support region mutations, and the weakest correlations are found for rim region mutations. The addition of molecular descriptors to the models, as presented in the lower half of the heatmap, increases the accuracy of the predictors both at the core and support regions, though rim regions are still inadequately characterised. One should note that there are 355 mutations, which affect the core region in the dataset to only 148 and 182 affecting the rim and support regions. This imbalance in the dataset may attribute to a weaker performance in the rim region mutations. The lowest correlations are for the rim regions, and found for the models derived from statistical potentials ($CP_{Off\text{-}Rate}$, AP$_{Off\text{-}Rate}$, CP_AP$_{Off\text{-}Rate}$). The lack of predictive power here may lie in their inability to model solvation effects, which are more prominent at the rim. Off-rate models using the physics-based descriptors show a higher correlation of *R=0.4* at the rim regions. Models with only hotspot descriptors, on the other hand, generally achieve correlations of R>0.5 with $R_{MAX}$=0.5.

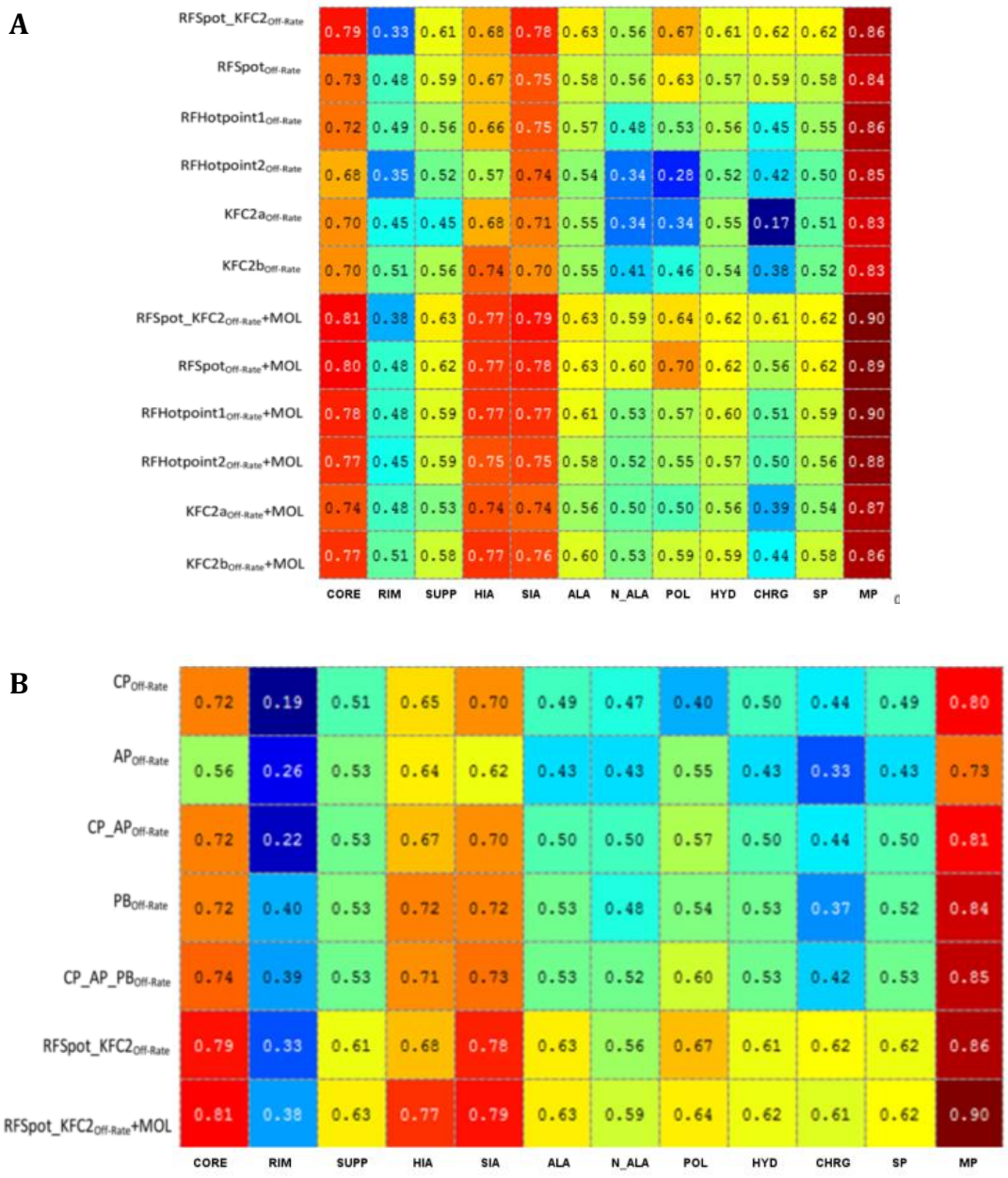



**Figure 7.4: Correlation heatmap of off-rate regression algorithms on data regions.**

(A) Off-Rate regression models that use hotspot descriptors, or a combination of hotspot and molecular descriptors. The different methods indicate the hotspot prediction method by which the hotspot descriptors where generated from. The respective data regions are shown on the x-axis and values in matrix show the PCC achieved by the given model for the given data region. (B) is similar to (A) except that off-rate prediction models using subsets of molecular descriptors are investigated. CP – Coarse-Grain Potentials; AP – Atomic-Based Potentials; CP-AP – All Statistical Potentials; PB – Physics Based Energy Terms. As a benchmark comparison, results for *RFSpot_KFC2Off-Rate* (best performing off-rate predictor

using hotspot descriptors) and *RF_Spot_KFC2$_{Off\text{-}Rate}$+MOL* (best performing off-rate predictor using hotspot and molecular descriptors) are also included.

2. Predictions on Large-Interface-Area Complexes, for Non-Alanine Mutations, improved with molecular descriptors: The hotspot descriptor predictors are better at capturing effects of mutants on Small-Interface-Area (SIA) than Large-Interface-Area (LIA) complexes. This discrepancy is alleviated with the addition of molecular descriptors to the models. Single-point mutations to alanine are generally better characterised than single-point mutations to non-alanine. This discrepancy is most accentuated for the less accurate hotspot predictor models and less so for the molecular descriptor models.

3. The hotspot descriptor model outperforms the molecular descriptor model for mutations to polar, hydrophobic and charged residues: The all-molecular descriptor off-rate model achieves PCCs of R=0.6, R=0.53, R=0.42 on polar, hydrophobic and charged residues respectively. Even though certain molecular descriptors are designed specifically for addressing electrostatics, degradation in performance is observed for charged residues. Interestingly, an accurate hotspot descriptor off-rate model, such as *RFSpot$_{Off\text{-}Rate}$*, achieves PCCs of R=0.67 (p=0.12), R=0.61 (p=0.28) and R=0.62 (p<0.001) for the same data regions, and shows significant increases in correlation to its molecular model counterpart on the prediction of charged residues; here p-values show the significance of the difference in PCC when compared to the molecular descriptor models.

4. Multi-point mutations are notably better characterised than single-point mutations: Correlations for multi-point (MP) mutations have an average PCC of $R_{MEAN}$=0.85 and are as high as $R_{MAX}$=0.9 for certain models. This is in contrast to the PCCs achieved for single-point (SP) mutations ($R_{MEAN}$=0.55) and indicates that the subtleties of SP mutations are harder to characterise than the collective effect of multi-point mutations. Note that, though theoretically, MP mutations have the potential to cause off-rate changes of larger magnitudes, this is not so in the present dataset, where the mean and standard deviation of $|\Delta\log_{10}(k_{off})|$ for MP mutations is 0.96 and 1.4 compared to 1.17 and 1.48 for SP mutations. Therefore, one cannot conclude that the reason for better prediction of multi-

point mutations is related to being able to predict extreme changes in $\Delta\log_{10}(k_{off})$ better than subtle changes in $\Delta\log_{10}(k_{off})$.

### 7.3.3 Specialized Feature Selection Models for Off-Rate Prediction

The above analysis was performed using models trained on all the 713 off-rate mutations in the dataset, of which the predictions were then subset into data regions for separate analysis. Here, off-rate models, which are only trained on subsets of mutations, as defined by the data regions, are investigated. Separate models are built for the different data regions of the dataset using a Genetic Algorithm for Feature Selection (GA-FS). All 110 molecular descriptors and 16 hotspot descriptors generated from the *RFSpot_KFC2* hotspot predictor were made available for feature selection. The feature set size was set to five features to avoid over-fitting and both non-linear (using Support Vector Machines, SVM) and linear (using Linear Regression, LR) models were investigated. For every data region, 50 separate GA-FS runs were performed; an inner-cross validation loop was used for FS (and SVM parameter optimization), whereas an outer-cross validation loop was used for testing the final model. The results of specialized models for the data regions are shown in Figure 7.5 (GS-FS LR in Red and GA-FS SVM in blue). The performance of the specialized models on the data regions is compared to that of the best performing global off-rate prediction model (*i.e.* $RFSpot\_KFC2+Mol_{Off\text{-}Rate}$). For most of the data regions, there is no advantage in having such specialized models, as having a global one-fits-all model suffices. However, for mutations in the rim region, and mutations to charged residues, having a specialized model significantly increases the correlation of off-rates in these data regions ($p=0.04$ and $p<<0.001$ respectively).

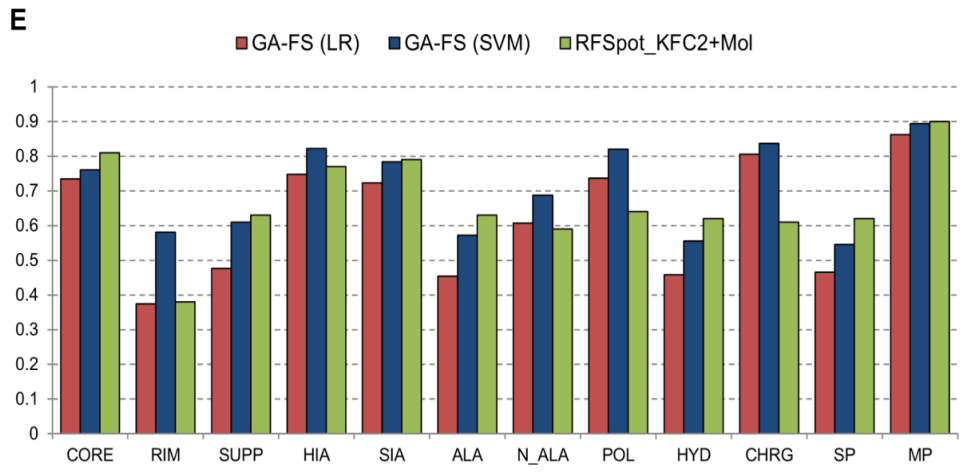



**Figure 7.5: Performance comparison of specialized models against one-fits all model.**

GA-FS Feature Selection Models using Genetic Algorithm are run for different data regions of the off-rate dataset for which both linear (using Linear Regression - GA-FS (LR) ) and non-linear (using SVM Regression GA-FS (SVM)) models are investigated. The figure shows the mean PCC of the optimal models found by the GA-FS runs for each data region. For comparison, PCC results on the data regions results are also shown for RFSpot_KFC2$_{\text{Off-Rate}}$+Mol. Note that the latter model is trained on all 713 off-rate mutations, and the predictions are separated post prediction into data regions and analysed for their PCC. This effectively compares the predictions of specialized models vs. one-fits-all model. Though there is no overall evidence that specialized models perform better than a one-fits-all model, certain subsets of mutations, such as those at the rim regions, show notable improvements when a specialized model is employed.

#### 7.3.3.1 Broadly Predictive and Highly Specific Descriptors for Off-Rate Data Regions

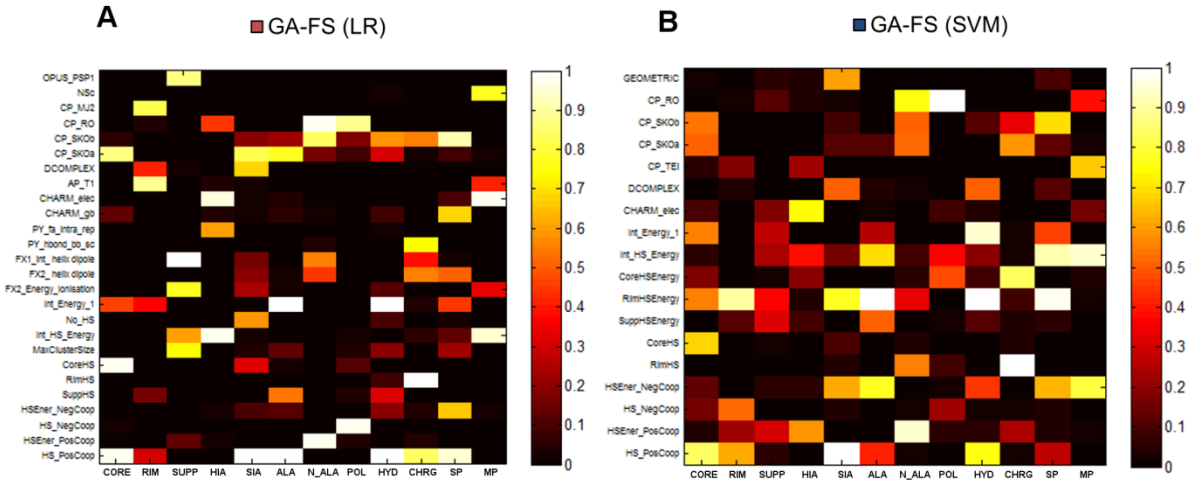



**Figure 7.6: Dataset heterogeneity in the 713 off-rate dataset.**
(A) and (B) shows the importance of the most selected features for each data region. The features shown are those that are part of the final model, for any data region for more than 50% of the GA-FS runs; colour bar displays this percentage. The features on the y-axis are ranked starting from (Coarse-grain Potentials, Atomic-based Potentials, Physics-Based Energy Terms and Hotspot Descriptors). The heat maps show how different descriptors are needed to accurately predict the mutations on different data regions.

Here, the relationship between the descriptors and data regions is investigated using the final features selected in the GA-FS runs. Initially available for the GA-FS algorithms are a set of 16 hotspot descriptors (generated from the hotspot predictor *RFSpot_KFC2*) and 110 molecular descriptors. For each region, the descriptors which are part of the final model, in at least half of the total number of runs, are singled out for analysis and presented in heat maps which indicate their importance to the given data region (Figure 7.6a: GS-FS (LR) and Figure 7.6b GS-FS (SVM)). On the y-axis, the singled out descriptors are listed and sorted according to descriptor type (CP, AP, PB, and hotspot descriptors from top to bottom), with each data region shown on the x-axis. Globally, it is observed that whereas for LR models, top features are distributed throughout the four main feature categories, for the non-linear SVM models, 61% of the features are hotspot descriptors, suggesting that non-linear relationships between hotspot

descriptors can be better exploited for the predictions of off-rates. Note that, if hotspot descriptors were equally important to molecular descriptors, only 12% of the final features would be expected to be hotspot descriptors.

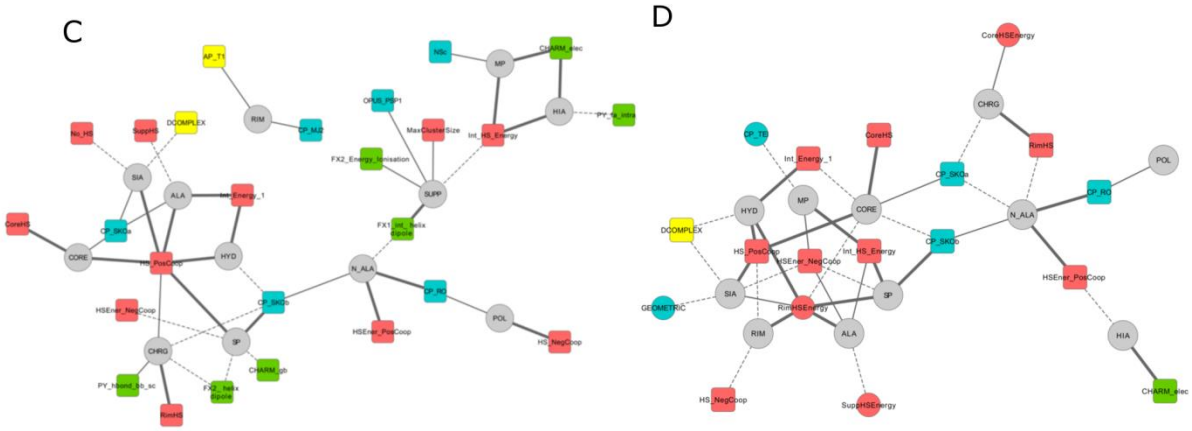



**Figure 7.7: Descriptor – data region networks.**
(C) and (D) are descriptor-data region networks for Figure 7.6a and Figure 7.6b respectively. Circled nodes represent data regions and square nodes represent features; therefore, only edges between circle and square nodes are present. An edge is present if the feature is in the final model for the given data region in more than 50% of the GA-FS runs (dotted edge), between 70-90% of the GA-FS runs (normal edge), more than 90% of the GA-FS runs (bold edge). Coarse-grain Potentials (blue), Atomic-based Potentials (yellow), Physics-Based Energy Terms (green), Hotspot descriptors (pink) and data regions (gray). From the descriptor-data region networks, descriptors highly specific to certain classes of off-rate mutations can be observed. Conversely, as in the case of the GA-FS (SVM) data region network, a cluster of broadly predictive hotspot descriptors is also shown.

To visualize the interconnections between descriptors and data regions, descriptor-data region networks are generated for both the LR (Figure 7.7 C) and SVM (Figure 7.7 D) GA-FS runs. An edge between a descriptor and a data region is shown if the given descriptor is part of the final GA-FS model in at least 50% of the GA-FS runs for the given data region (with increasing edge weight for > 50%). Several descriptors are highly specific to certain data regions. For instance in the LR model (Figure 7.7 D), two statistical potentials, (*AP_T1* (Tobi, 2010) and *CP_MJ2* (Miyazawa and Jernigan, 1996)), are specific to rim region mutations. Whereas others, such as *HS_PosCoop,* as highlighted by their high degree, are

broader in their predictive value and can explain off-rate changes in a number of data regions collectively. Interestingly, for the support regions, *MaxClusterSize* is invoked which suggests that larger hotregions in the support regions may be important for complex stabilization. Certain descriptor-data region relationships hold for both LR and SVM models, such as the electrostatic contributions (*CHARMM_elec* (Brooks et al., 2009)) from mutations on complexes of large-interface-area (LIA). The ability to model nonlinearities between features, invokes some different descriptors. Most notably, a key observation specific to the SVM descriptor-data region network is a central cluster of highly interconnected hotspot descriptors and data regions, which involve *HS_PosCoop*, *HSEner_PosCoop*, *Int_HS_Energy* and *RimHSEnergy*.

## 7.4 Off-rate Prediction and Conformational Changes

Predictions of all off-rate regression models are analysed separately for mutations on complexes that show significant backbone conformational changes for, either or both, binding partners upon complex formation. The subset of complexes for which the unbound crystal structures of the *wild-type* complex are available, were singled-out and their I_RMSD values for backbone conformational rearrangements were extracted from the work of Kastritis et al. (2011). This subset of complexes for which unbound crystal structures are available, amounts to 17 complexes and 332 mutations. A total of 67 mutations on four complexes show significant conformational changes with (I_RMSD >1.5 Å), and if the threshold is lowered to (I_RMSD >1 Å), this results in 119 mutations on six complexes.

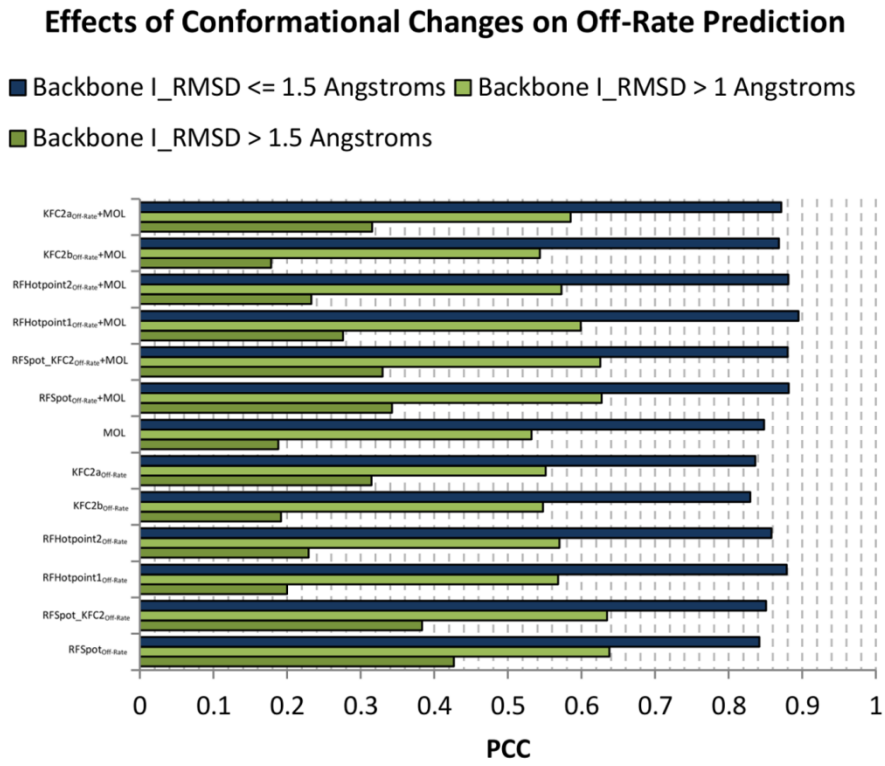



**Figure 7.8: Effects of conformational changes on off-rate prediction.**
From the predictions of the original 13 regression models developed for off-rate prediction. The predictions are assessed separately (PCC with $\Delta\log_{10}(k_{off})$) for mutations on complexes which undergo significant backbone conformational changes of I_RMSD > 1.5 Å (dark green), notable conformational changes of I_RMSD > 1 Å (light green) and little to no conformational changes I_RMSD < 1 Å (dark blue). Predicted accuracy is directly related to the magnitude of conformational change and becomes highly dependent on the model at high conformational changes. I_RMSD values extracted from (Kastritis et al., 2011).

The PCCs for the off-rate model predictions with $\Delta\log10(k_{off})$ are shown under three conformational change categories (Figure 7.8). The PCC, for complexes which show little to no conformational change (I_RMSD < 1.5 Å), averaged over all prediction models, shows a correlation of R=0.86, which decreases to R=0.58 at (I_RMSD >1 Å) and R= 0.28 at (I_RMSD > 1.5 Å). Though for the latter category, *RFSpot*$_{Off\text{-}Rate}$ achieves a correlation of R=0.43. Changes in the different models are more apparent for complexes with higher conformational changes, most notably is the discrepancy in PCC between Molecular and *RFSpot*$_{OFF\text{-}Rate}$ Off-Rate prediction models. This discrepancy is minimal at complexes with little conformational changes, $\Delta R = 0.01_{I\_RMSD < 1.5\ Å}$ and increases to $\Delta R_{I\_RMSD > 1\ Å} = 0.11$ and $\Delta R_{I\_RMSD > 1.5\ Å} = 0.24$ for complexes with significant conformational changes.

## 7.5 Effects of Cross-Validation Routine on Off-Rate Prediction Performance

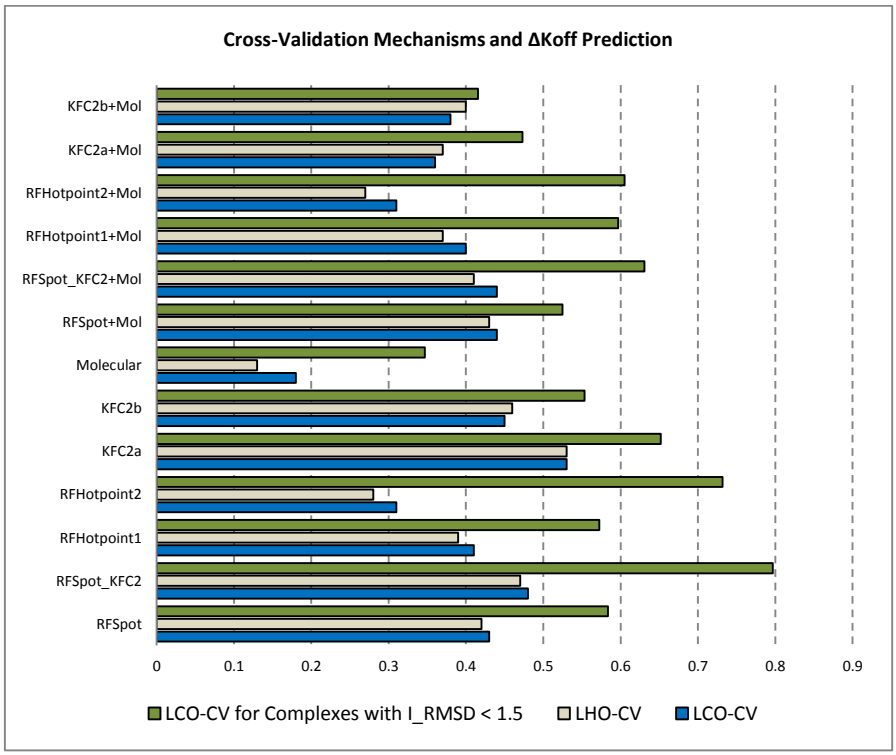



**Figure 7.9: PCCs for off-rate prediction models using the 713 off-rate mutant dataset from SKEMPI**.
Leave-Complex-Out CV(LCO-CV), Leave-Homology-Out CV (LHO-CV) and LCO-CV for complex which undergo minimal to no conformational changes with I_RMSD < 1.5 Å as defined in (Kastritis et al., 2011). The models differ by their features sets. First six use hotspot descriptor sets, followed by a molecular descriptor set model (Molecular), and models that combine both (+Mol). Degradation in performance is observed when using both LHO (blue bars) and LCO-CV (beige bars) routines. This degradation is less evident once controlled for conformational changes (green bars).

The results presented in previous sections which use 20-fold cross validation for the generation of test predictions, are an estimate of predictive power given one already has some mutant information on the complex in question to train upon. This type of cross validation is not a valid estimate of a model's general ability to predict on an unseen complex. Therefore, two additional cross-validation mechanisms were also applied; Leave-Complex-Out CV (LCO-CV), where all mutations of a complex are left out as a test set and Leave-Homology-Out CV

(LHO-CV); a more stringent form of cross-validation which accounts for homology and interface similarity as devised in the work of (Moal and Fernandez-Recio, 2012). The proteins held out in each LHO-CV fold are listed in the appendices Table 10.1. The PCCs of the test predictions with $\Delta\log10(k_{off})$, of the two CV routines, are shown Figure 7.9. Given that for 20-Fold CV, R > 0.7, for LCO-CV and LHO-CV, the models severely over-fit. In essence, the predictive ability of the hotspot descriptors such as $HSEner\_PosCoop_{RFSpot}$ (|R|=0.57), $Int\_HS\_Energy_{Hotpoint1}$ (|R|=0.57) and $SuppHSEnergy_{KFC2a}$ (|R|=0.62) is being impeded by the learning model and noise from other features. It is important to note that the LHO-CV might not be well suited for certain practical purposes. For example, if one wishes to be able to predict mutations on an enzyme inhibitor complex, it would be natural to have such complexes in the training set, unlike what is actually done here for LHO-CV. The largest amount of over-fitting is observed for the molecular descriptor model, which is alleviated with the hotspot descriptor models and in both CV mechanisms, the correlations achieved by the hotspot descriptor models, is higher than that achieved by the molecular descriptor set model. LCO-CV was also performed on the subset of 14 complexes and 265 mutations, which show little to no conformational change. It is observed that the reduction in ability to model the effects of mutations on unseen/unrelated is largely affected by conformational changes. For example, for $RFSpot\_KFC2_{Off\text{-}Rate,}$ the correlation achieved is as high as 0.8 when limited to rigid complexes, even when LCO-CV is being performed.

## 7.6 Discrepancy in Prominent Features Across LHO Folds

The low prediction accuracy across LHO folds suggests that descriptors responsible for one fold are not generalizable to others. To investigate this, models are built for mutations only within a fold and the most important features are highlighted and compared to the features from models built on other LHO folds. Genetic Algorithm Feature Selection (GA-FS) is used to build such specialized off-rate prediction models and both linear and non-linear models are investigated.

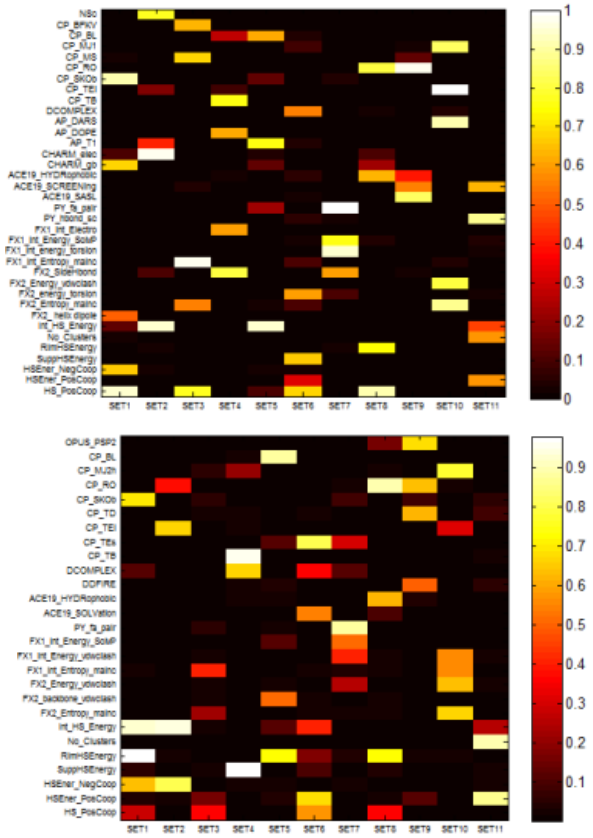



**Figure 7.10: Heterogeneity across different protein families.**
Left: GA-FS (LR), Right: GA-FS (SVM). The colour bar indicates the percentage number of times the given feature made it to the feature set of the final model after a GA-FS run. Features shown are those which make it to the final model more than 50% of the time for at least one set on the x-axis. As observed in both heat maps, different protein-families need to employ different descriptors in order to be accurately predicted.

The Features that make it to the final models (Figure 7.10 left for LR and Figure 7.10 right for SVM) indicate heterogeneity in the features selected across folds, and no one-feature-fits-all may be identified. This again may contribute to the reduction in PCCs when using LHO-CV mechanisms, as mutations on unseen complexes may be better predicted using features that were not prominent in the training set mutations. Biases related to different experimental methods from which the $\Delta\log10(k_{off})$ of the mutations where calculated are also known to have significant effects on the prediction of binding free energies (Kastritis and Bonvin, 2010, Moal et al., 2011) and may also play a role in the reduction of accuracy when using LHO- and LCO-CV mechanisms.

## 7.7 Dicussion

To assess the predictive abilities of hotspot descriptors when combined in learning models, several machine learning models trained on $\Delta\log_{10}(k_{off})$ are investigated in this chapter. Hotspot descriptor models provide more accurate predictions for $\Delta\log_{10}(k_{off})$ than molecular descriptor models and the best regression model, which combines both molecular and hotspot descriptors, *RFSpot_KFC2*$_{Off\text{-}Rate}$*+Mol*, achieves a PCC of R=0.79 with experimental off-rates.

### 7.7.1 Dataset Heterogeneity, Descriptors and Learning Models

The assessment of predictor performance on different subsets of the dataset termed as 'data regions' and the subsequent generation of 'descriptor-data region networks', proved to be an insightful exercise. Firstly, it puts to light the interconnected relationship between, dataset heterogeneity, descriptors and learning models and secondly it highlights regions that still require attention.

An example, which is representative of both these points, is the prediction of mutations in rim regions. When using a global off-rate model, there is a striking discrepancy in our ability to model off-rate changes for mutations in rim regions as opposed to the core regions. What is special about rim mutations that render them to act differently than other mutations? What makes them harder to predict? There are several mechanisms at play that must be considered. Firstly, what has been uncovered is that the features, which are able to characterise changes in off-rate for the majority of mutations, fail to do so for the subset of mutations at the rim regions. This does not necessarily mean that there are not descriptors that are able to characterise the rim region mutations well. Rather, it may very well be that there are very specific descriptors which are able to characterise mutations at the rim regions, but the learning model has no way to distinguish between rim mutations and core mutations. Therefore, the learning model cannot apply different descriptors to such regions separately. To assess if the latter is true, models specific to (trained and tested on) rim region mutations only, were built using the (GA-FS SVM/LR) models. Here the GA-FS SVM model

achieves R=0.58. This in comparison to R=0.38 (p=0.04) achieved on the rim region mutations using a global off-rate model. This shows that indeed, there are descriptors within the feature set, which are able to characterise rim regions better, and such descriptors are highly specific to mutations occurring at the rim. However, it is important to note that the increase in correlation to R=0.58 for rim region mutations is only achieved using the non-linear model (SVM). This is not the case with the linear LR model, even though the LR model is also specific to (i.e. trained and tested on) rim region mutations. One can therefore appreciate that in datasets which are heterogeneous, any generalizations about the importance of descriptors, are highly dependent on not only the data-region in question but also on the learning model used.

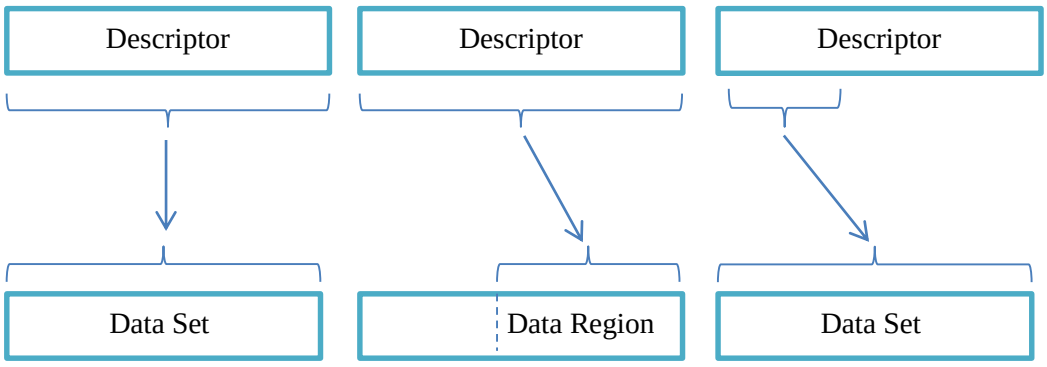



**Figure 7.11**: **Ways in which different learning algorithms link descriptors to a dataset.**
Datasets of protein-protein interactions and mutations at protein interfaces are intrinsically heterogeneous. This renders the 'one-fits-all' assumption of Linear Regression (A) very limiting, even if not considering the limitations brought about by the linear assumption. (B) Learners which are able to distinguish between different 'data regions' within a dataset may apply descriptors which are specific to that region only. These include Look-Ahead Regression Trees and Hierarchical Mixture Models. (C) Certain descriptors might not be only accurate within certain ranges, hence models such as Multi-Adaptive-Regression-Splines (MARS), which are selective on which regions of the descriptor are used in the final model, maybe be used.

Understanding this 3-way relationship between descriptors, data-regions and the learning models used, is an important consideration for the prediction of off-rates on similar datasets to that used in this work. Effectively, the learning model is nothing but a linker of descriptors to data. The various scenarios differing by

the way learning models link descriptors to data are summarized in Figure 7.11. If the dataset is heterogeneous, and data-regions for which certain descriptors are specific to do exist, then the data-regions should be made visible to the model. Introducing categorical features as an indicator to a data region, for example, may do this. In addition, the learning model chosen should be able exploit these categorical features and selectively apply descriptors specifically to the data regions indicated by them. Examples of which are; non-greedy decision trees and hierarchical-mixtures-models, as these models do not necessarily assume features to be ordinal. Effectively, the use of such models has the advantage of still having just one learning model for the whole dataset, but also the added flexibility of having specialized models for different data-regions.

### 7.7.2 Future Endeavors – Conformational Changes

Predicting the effects of mutations of complexes, which undergo significant backbone conformational change, remains a challenge. This is shown to be true both when predicting *wild-type* binding free energies described in Chapter 2, and here in the prediction of off-rate changes upon mutations. Reasons for the reduction in performance can be several-fold, but more than likely stem from the fact that all calculations use only the bound conformations of the receptor-ligand complex. This effectively translates itself into a one-step binding process,

$$R + L \underset{k_2}{\overset{k_1}{\rightleftarrows}} RL \qquad 7.1$$

For complexes which undergo minimal conformational changes, this is indeed a sufficient approximation, as confirmed by the mean correlation of $R_{AVG}$=0.86 for complexes with <1.5A backbone rearrangement. Once conformational changes come into play, then binding is better approximated using a 2-step binding process.

$$R + L \underset{k_2}{\overset{k_1}{\rightleftarrows}} RL \underset{k_4}{\overset{k_3}{\rightleftarrows}} RL^* \qquad 7.2$$

where,

$$K_{off\ 2-step} = \frac{k_2 k_4}{(k_2 + k_3 + k_4)} \quad 7.3$$

Here, the $k_{off}$ can be decreased even if mutations destabilize the bound-state *RL**. Rather, the decrease in off-rate is brought about by an increased stability of the transition state *RL (Lu and Tonge, 2010)*. Therefore, the accurate characterization of the transition state *RL*, which is not trivial, becomes as important as that of the final bound state *RL**. With conformational changes, different binding mechanisms also come in to play. For example Weikl and von Deuster (2009) show that depending on the binding mechanism (conformational selection or induced-fit), mutations that do not affect the stability of the interface, but affect the conformational equilibrium of the receptor *R,* also affect the off-rate. Last but not least, complexes are not static structures, and ideally, a similar conformational sampling mechanism to the one used in Chapter 2 is also employed to the off-rate scenario. This might be particularly important for complexes, which are natively unstructured/disordered in local regions, as these regions may still remain disordered even in the bound state (Xia et al., 2004, Zeth et al., 2002). Binding site variability has also been observed in certain complexes where the variability is not explained by experimental or procedural inaccuracies (Hamp and Rost, 2012).

# Chapter 8

## 8 Distribution of Stability in Protein-Protein Interfaces

Chapters 6 and 7 show how counting the energies of hotspot energies, pre- and post-mutation provides an accurate description of changes in $\Delta\log_{10}(k_{off})$. Here, the focus shifts on understanding to which extent, the off-rate of a complex is affected by the distribution of hotspots. Given that protein-protein interfaces contain a number of hotspots, which may occur in disjointed regions, the central question addressed here is the following; Are certain hotspot regions of the interface more susceptible to destabilizing/stabilizing the interaction upon mutation? For example, are hotspots at the core sufficient for high complex stability? Can rim hotspot residues share a role as important as that of core hotspot residues? Given an interface with a number of hotpots, do hotregions provide an added level of stability which hotspots on their own do not? Knowing

which hotspot regions are more important to the stability of the interaction is critical for inhibiting protein-protein interactions (see section 1.4.1) and designing better protein drugs (see section 1.4.2) and to date, there is no study investigating this. As a basis on which the computational experiments are designed, results from previous work related to the distributional patterns of hotspots are used as initial hypotheses.

In the first part of this work (section 8.1), the role of the core and rim hotspot residues is revisited in the context of the dissociation rate. To do so, the initial assumption taken is that the critical region of stability of a protein-protein interaction emanates from the core hotspots, as evidenced in the work of Bogan and Thorn (1998). With a number of additional computational experiments, our observations suggest that, for off-rates, the above only holds for large complexes. As for small complexes, all regions of the interface are critical for the stability of the interaction i.e. rim hotspot residues are as equally responsible for low dissociation rates. A second property of hotspots related to their distribution, is that hotspots tend to cluster into tightly packed regions known has hotregions (Keskin et al., 2005). The authors report that the conservation of this type of organization suggests that they are important for protein-protein association. However, the aforementioned analysis is not performed in relation to binding free energies or off-rates for protein-protein interactions and the suggestion is somewhat speculative in nature. In section 8.2, the extent to which the presence, number and size of hotregions is advantageous to complex stability, is therefore investigated. In the same work of (Keskin et al., 2005), it is suggested that hotregions are cooperative in nature and future scoring functions should account for this effect so as not to overestimate/underestimate the contribution of hotregions. In the latter sections of this work, when hotregions are tested for potential cooperative effects (section 8.3), no prevalent form of cooperativity is observed. In addition the contribution of hotregions of different sizes towards stability is determined under different cooperativity assumptions (section 8.4).

## 8.1 Critical Regions of Stability in Protein-Protein Complexes

In the previous chapters 6 and 7, it was shown how the change in the off-rate of a complex is directly related to a change in the sum of its hotspot energies. Therefore no distinction is made on which hotspot energies the mutation is modulating. Using the hotspot energies at the core, rim and support regions, it is assessed whether complex stability can be effectively disrupted homogenously across the interface (equally across the three regions) or preferentially in a particular region. *CoreHSEnergy*, *RimHSEnergy* and *SuppHSEnergy* represent the change in total hotspot energies limited to each region upon mutation. Effectively, the PCC of these descriptors with the off-rate expresses how well changes in the given region show themselves as changes in $\log_{10}(k_{off})$ - irrespective of changes in hotspot energies in any other region. Therefore, by assessing the relative PCCs of the three regions we can gauge whether a given region acts independently and dominates in its contribution to complex stability compared to other regions. Given that there are 6 instances of each hotspot descriptor, as generated per each hotspot predictor, the correlations for each descriptor shown are the mean of each descriptor's correlation under the six-hotspot predictors. Hence results can be considered to be independent of the hotspot predictor generating the hotspot descriptors.

From the PCCs of the three-hotspot region specific descriptors (*CoreHSEnergy* |R| = 0.48, *RimHSEnergy* |R|= 0.20 and *SuppHSEnergy* |R| = 0.38), it is observed that changes in the hotspot energies at the core affect the off-rate more significantly than the rim ($p<<0.01$) and support region ($p<0.01$). Given that 355 mutations affect hotspot energies in the core region compared to 148 and 182 for rim and support regions respectively, results may however be biased. For example, if fewer events are observed at the rim region, there is less chance of the rim region playing a significant role in off-rate changes, when looking at it globally over a population of complexes as is done presently. To remove this potential bias, the subset of mutations, which affect all three regions simultaneously, is extracted and PCC recalculated. The PCCs still suggest

dominance from the core region (|R|= 0.53), more significantly than the rim region (|R| = 0.22 p<<0.01).

### 8.1.1 Stability Regions in Small and Large Interfaces

To investigate whether the relative importance of these three regions of stability changes when considering complexes of different interface areas, the dataset is divided into small interface area (SIA) complexes (< 1600 Å$^2$ buried surface area) and large interface area (LIA) complexes (> 1600 Å$^2$ buried surface area). The threshold of 1600 Å$^2$ is such that both subsets are of similar number of examples. The mean PCC for the *CoreHSEnergy*, *SuppHSEnergy* and *RimHSEnergy* for LIA and SIA complexes is calculated and shown in Figure 8.1a and Figure 8.1b respectively.

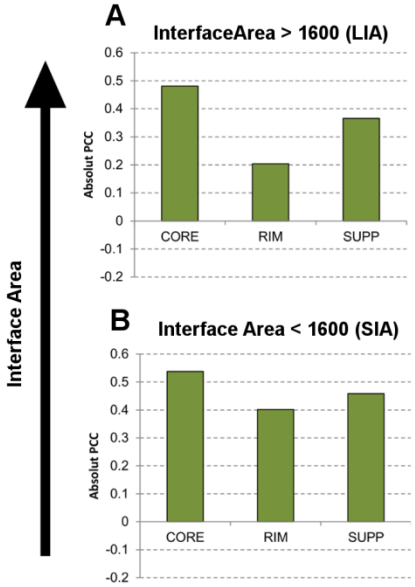



**Figure 8.1: Critical Regions of Stability as a function of Complex Interface Area.**
(A) The absolute PCC of the sum of changes in hotspot energies at the Core, Rim and Support Interface Regions with experimental $log_{10}(k_{off})$ - for complexes with large interface areas. (B) Similar to (A) but for complexes with small interface areas. The threshold for large / small interface area of 1600 Å$^2$ is chosen in such a way that divides the dataset into samples of similar size.

For the LIA complexes, a dominant contribution from the changes in core hotspot energies (*CoreHSEnergy* |R|=0.48) and minimal contribution from *SuppHSEnergy* (|R|=0.37) and *RimHSEnergy* (|R|=0.20) is observed (Figure 9A). Therefore, even though a given set of mutations might be affecting support or rim regions, it is the changes in hotspot energies at the core region which show up as the dominant changes in the off-rate |R|=0.48). For SIA complexes (Figure 9B), changes in hotspot energies at the rim regions, show a highly significant 2-fold increase in correlation ($p<<0.01$). This renders all three regions with somewhat similar contributions to complex stability (*CoreHSEnergy* |R|=0.56, *SuppHSEnergy* |R|=0.46, *RimHSEnergy* |R|=0.40). For LIA complexes, mutations applied in positions that affect the core to those that affect the rim is 2:1. On considering SIA complexes the ratios increase to 3:1. Therefore the increase in importance of the rim hotspot energies occurs in spite of a decreasing ratio. As an additional test, which accounts for biases in the number of examples affecting each region, the correlations are calculated for only the mutations, which make changes in the respective region, again taking an average over all 6 hotspot predictors' descriptors. Here no significant changes in correlation are observed in LIA and SIA complexes for the core and support region. For LIA complexes, changes in rim hotspot energies have minimal effect on the off-rate with |R|=0.29, whereas for SIA complexes, a 1.75-fold increase ($p<0.01$) in correlation is observed (|R|=0.51). This confirms that hotspots at the rim of the interface can have a role as dominant to that of core region hotspots.

### 8.1.2 Stability Regions in Small and Large Complexes

The dataset is divided into the mutations, which are found on Large-Complex-Size (LCS) (with 231 mutations), and Small-Complex-Size (SCS) complexes (with 482 mutations). The PCC for *CoreHSEnergy*, *RimHSEnergy*, *SuppHSEnergy* averaged over the descriptors from all hotspot predictors is calculated for both LCS and SCS Figure 8.2a and Figure 8.2b respectively.

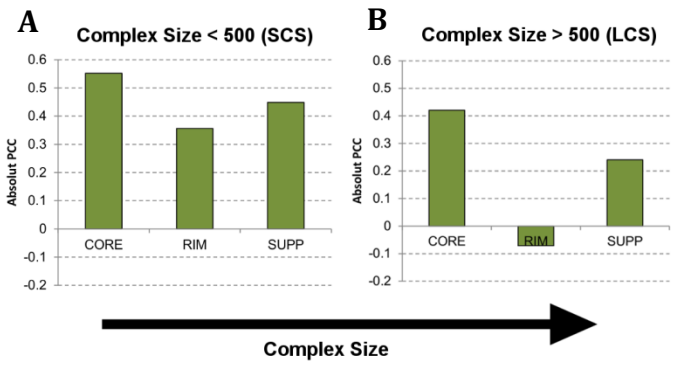



**Figure 8.2: Critical Regions of Stability as a function of Complex Size.**
(A) The absolute PCC of the sum of changes in hotspot energies at the Core, Rim and Support Interface Regions with experimental $log_{10}(k_{off})$ – for large complexes. (B) Similar to (A) but for small complexes. The threshold for large / small complex size of 500 residues is chosen in such a way that divides the dataset into samples of similar size.

Core hotspots are critical to the stability of LCS complexes whereas for SCS complexes, all three regions are important. This effect is synonymous with what is observed in LIA and SIA complexes, though the increase in correlation for *RimHSEnergy* (R=0.07 to R=-0.36 $p<<0.001$) is more pronounced for complex size rather than interface area. Even though fewer mutations are on LCS complexes (231), the percentage of mutants affecting each region in LCS, compared to that for SCS, is similar across the three regions (61%, 52% and 46% for core, rim and support regions respectively) and therefore shows no relationship to the changes seen in the PCC of the three regions from LCS to SCS.

### 8.1.3 The Role of Rim Regions in Small Complex Sizes

On the 50 complexes considered in the 713 off-rate mutant dataset, complex size and interface size show a correlation of R=0.55 (Figure 8.3i). The correlation is higher (R=0.74) for complexes sizes of less than 500 residues, and becomes insignificant (R=0.18) beyond complex sizes of 500 residues and above. The dataset is therefore further divided into four regions (Figure 8.3i), which include: SIA-SCS (191 mutations), SIA-LCS (67 mutations), LIA-SCS (191 mutations), LIA-LCS (164 mutations) and again the PCC for *CoreHSEnergy*, *RimHSEnergy*,

*SuppHSEnergy* averaged over the descriptors of all hotspot predicators is calculated and shown in Figure 8.3g: SIA-SCS, Figure 8.3h: SIA-LCS, Figure 8.3e: LIA-SCS and Figure 8.3f: LIA-LCS.

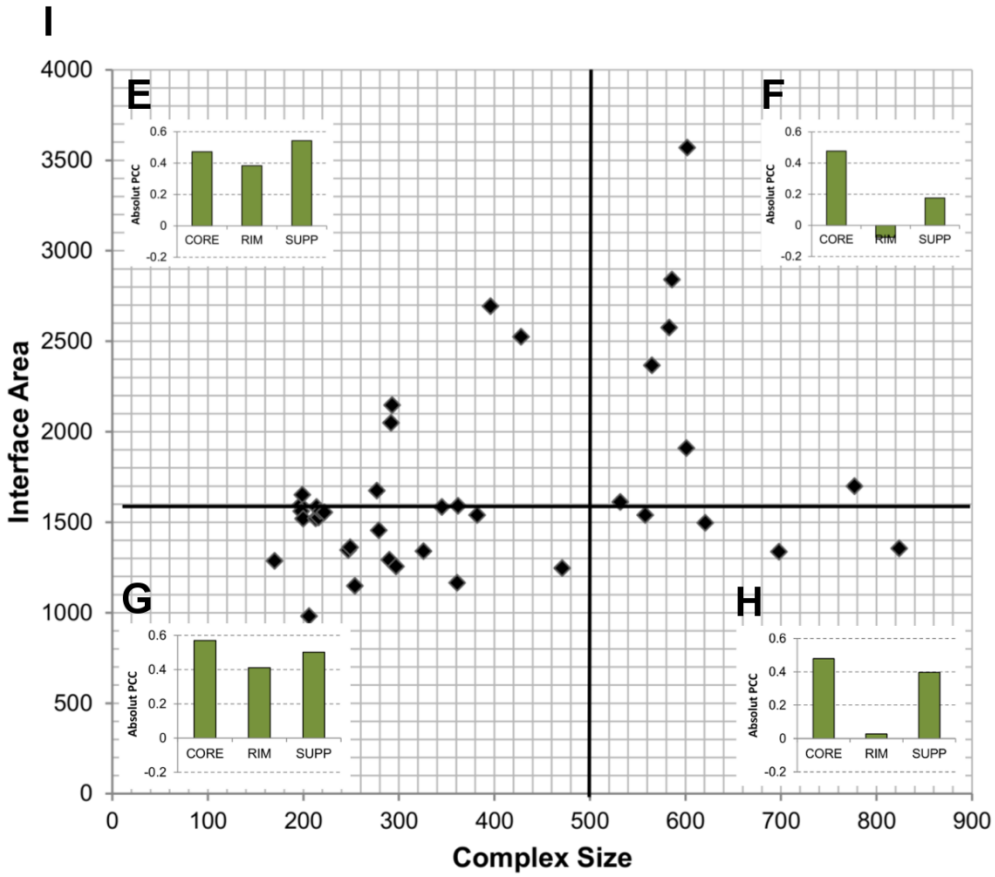



**Figure 8.3: Stability regions, interface-area and complex-size.**
The changes in hotspot energies upon mutation are assessed at three interface regions. This enables exploration of changes in the distribution of stability for complexes of different size and interface-area. CORE, RIM and SUPP represent the PCCs of CoreHSEnergy/RimHSEnergy/SuppHSEnergy averaged for the 6 hotspot prediction algorithms with $\Delta\log_{10}(k_{off})$.(A) PCCs for mutants on Complexes with interface-area >1600 Å$^2$ (LIA). (B) PCCs for mutants on complexes with interface-area <1600 Å$^2$ (SIA). (C) PCCs for mutants on complexes with size <500 residues (SCS). (D) PCCs for mutants on complexes with size >500 residues (LCS). (E) LIA-SCS, (F) LIA-LCS, (G) SIA-SCS, (H) SIA-LCS. (I) Scatter plot of complex size vs. interface area for all complexes in off-rate mutant dataset. Here it is observed that complex stability is distributed across all three regions for small-size complexes (C, E and G), whereas the core becomes a localized region of stability for large-complex sizes (D, F, H). On analysis of the interface-area vs. complex-size subsets (E–H), the distribution of stability regions is affected primarily through complex-size irrespective of interface-area.

Here it is observed that given a fixed complex size (SCS or LCS), moving from small interface areas to larger interface areas, the landscape for the contributions of the Core, Rim and Support regions is unchanging. Therefore, independent of the interface area size, for low complex sizes off-rate has the propensity to be affected equally from all regions of the interface, whereas for high-complex sizes, stability is primarily emanating from core hotspots. More so, when moving from LCS to SCS, rim regions transition from having to an insignificant role to a more primary one – equal to that of core and support regions.

Further analysis of SCS and LCS complexes shows a greater sensitivity in off-rate changes upon mutations for SCS complexes; the mean $|\Delta\log_{10}(k_{off})|$ is 1.4 and 0.69 for SCS and LCS complexes respectively. Though the latter result is intuitive, in that changes on large complexes are less likely to have effects as significant as those on small complexes, the key finding here is that on dissection of the three interface regions, the reduction in the ability to make significant changes in LCS is not equally shared equally on the three regions. Rather, mutations at the core can still have notable effects on the stability of large complexes as in the case of smaller complexes. Conversely, the higher sensitivity of SCS complexes to mutations is due to the increase in importance of role of the rim regions and also possibly the support regions.

## 8.2 Effect of Hotregion Size, Count and Complex Dissociation Rate.

Analysis of the mean PCCs for $No\_Clusters$ (the change in the number of hotregions upon mutation, R = −0.15) and $MaxClusterSize$ (the change in size of the largest hotregion R = −0.09), show no notable contribution to changes in the off-rate (See Table 6.1). Both the change in interface hotspot energy, and change in the number of hotspots show higher correlations (R = −0.51 and R = −0.44 respectively). For the hotspot predictor $RFSpotKFC2$, both $No\_Clusters$ $and$ $MaxClusterSize$ show higher correlations than the average (R = −0.29, for both),

and the combination of the two descriptors into one using multiplication increases the PCC with $\Delta\log_{10}(k_{off})$ to R = −0.48. Though this may suggest that, larger and more hotregions in a complex result in a higher dissociation rate, its correlation of R = 0.6 with the change in hotspot energies ($Int\_HS\_Energy$), suggests that the underlying mechanism might still be the change in hotspot energies, irrespective of hotregion size and count. From this analysis, though it cannot be concluded that larger hotregions do not provide added stability to the complex, it is shown that the disruption of the largest hotregions, is not critical to complex stability.

## 8.3 Hotregion Cooperativity and Complex stability.

Probing the importance of the tendency for hotspots to cluster into hotregions, and for that matter, the importance of both size and number of hotregions for complex stability, has also to be done in the light of hotspot cooperativity. Cooperativity within hotregions has been suggested to be a natural consequence of the tight packing ratios found for hotspot residues in hotregions (Keskin et al., 2005). This adds another layer of complexity in validating the role of hotregions, as under cooperativity, larger hotregions do not necessarily contribute more to complex stability. In turn, this knowledge is critical in order not to overestimate or underestimate the contribution of hotspot energies within hotregions. There are two caveats to this, firstly it is not obvious what type of cooperativity exists within the hotregions and complexes in the dataset, and secondly, if present, this cooperativity has to be accounted for with a function. To our knowledge, this is the first work to include energetic descriptors which account for potential cooperative effects in an empirical scoring function.

The approach taken here is that no assumption is made before hand for any type of cooperativity prevalent in the complexes and hotregions within our dataset. Rather the two hypotheses of positive cooperativity ($HSEner\_PosCoop$) and negative cooperativity ($HSEner\_NegCoop$) are investigated and compared to the baseline hypothesis of additive hotspot energies ($Int\_HS\_Energy$). For ease of

reference, these three descriptors are referred to as the cooperativity descriptors and the motivation behind their functional forms are detailed in the methods section 2.3.7.3. Effectively, the higher the PCC of these descriptors with the off-rate, the more likely it is that hotregions on the complexes of the 713 off-rate mutant dataset, show the given type of cooperative/additive effect.

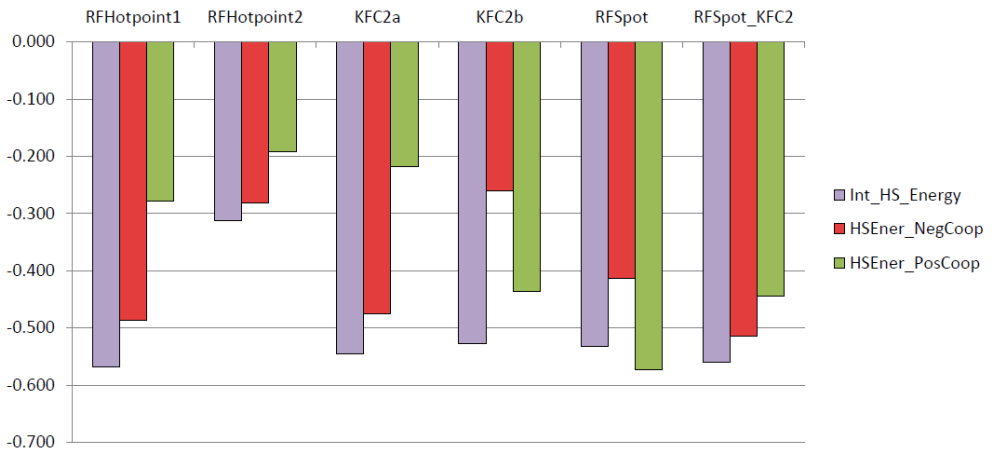



**Figure 8.4: PCCs of Hotspot Cooperativity Descriptors with experimental $\Delta\log_{10}(k_{off})$.**
*Int_HS_Energy* assumes no cooperativity within hotregions and therefore all hotspot energies are additive within a hotregion. *HSEner_PosCoop* assumes positive cooperativity within hotregions, where its total energy is greater than the sum of its parts. *HSEner_PosCoop* assumes negative cooperativity within hotregions and therefore the total energy of a hotregion is less than the sum of its parts. The design and functional forms of each descriptor are detailed in methods section 2.3.7.3. From the correlations of the descriptors generated by the predictions of different hotspot predictor algorithms, there is no prevalent form of cooperativity observed within hotregions. Rather, the data suggests that the additivity assumption is the safest one to take.

In Figure 8.4, the PCCs of cooperativity descriptors with $\Delta\log_{10}(k_{off})$ are highlighted for every hotspot predictor investigated. From these results, no evidence is found for a prevalent form of cooperativity in hotregions, as the additivity assumption works generally better than positive or negative cooperativity assumption. Several alanine scanning experiments on protein-protein interactions indicate that mutations are, to a large extent, naturally additive (Pal et al., 2005, Horovitz, 1996, Gregoret and Sauer, 1993). Eleven residues in the helix-turn-helix motif of the N-terminal domain of Gamma

repressor, found in a region important for DNA binding, were substituted to alanine using binomial mutagenesis (Gregoret and Sauer, 1993). The authors confirmed their nature to be additive and a model assuming additive interactions was able to predict the activity class of mutants with 90% accuracy (Gregoret and Sauer, 1993). In similar fashion, nineteen residues within the hGH site 1 for binding to the hGHR were randomized using a combinatorial, shotgun alanine-scanning library (Pal et al., 2005). On comparison of the counts of double alanine-mutations in hGH site 1 variants selected for binding to the hGHR , from the 144 pairwise combinations, only 15 pairs (10%) behave in a cooperative manner. Still, the experiments mentioned above are not specific to only hotspot residues, and therefore their results are not directly comparable to ours which are specific to hotregions; for example combinatorial mutant analysis of the TEM1-BLIP complex which is performed on residues in tight packed modules, and hence more akin to hotregions, shows that residues within a cluster tend to show strong positive cooperativity (Reichmann et al., 2005). The inclusivity of these results are discussed further in section 8.5.

## 8.4 Effects of Cooperativity on the Effective Energetic Contribution of Hotregions.

As highlighted in the previous section, understanding hotspot cooperativity within hotregions is necessary so as not to overestimate or underestimate the importance of the given hotregion. In order to understand better the effects of the cooperativity descriptors, the average hotspot and hotregion energies of different hotregions sizes is plotted for each of the three cooperativity assumptions.

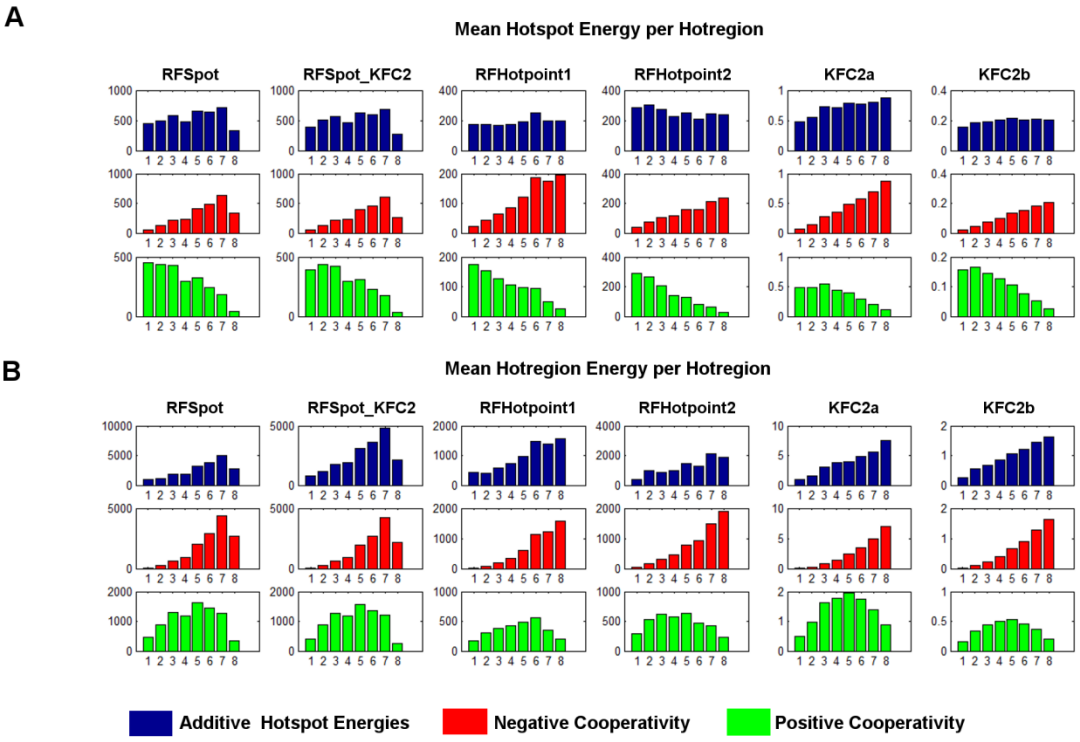



**Figure 8.5: The summation of single-point alanine ΔΔGs of a hotregion may underestimate/overestimate its contribution if negative/positive cooperative effects are at play respectively.**

Here, the effects of accounting for cooperative/additive effects on the predicted hotspot and hotregions energies on all mutated complexes used in this work, is shown. (A) The mean hotspot energies for hotregion sizes of 1 to 8 hotspot residues. Each column shows the predictions of different hotspot predictors. (A) First row (blue), shows the raw mean hotspot energies, which essentially assumes all hotspots are additive within a hotregion. (A) Second row (red), assumes negative cooperativity within hotregions. To account for negative cooperativity, a linearly increasing weight is applied to the hotspot energies according to the size of the hotregion they are in (see methods section 2.3.7.3). (A) Third row (green), assumes positive cooperativity within hotregions and a linearly decreasing weight is applied to the hotspot energies according to the size of hotregion (see methods section 2.3.7.3). (B) is similar to (A) but values are now the mean of the total hotregion energy of the given size. Effectively, the additive hotspot energy assumption results in hotregions contributing in a linearly increasing manner according to their size, the negative cooperativity assumption results in hotregions contributing in an increasing exponential-like manner as the hotregions increase in size, and the positive cooperativity assumption results in hotregions reaching a maximum contribution at around a hotregion size of 5, with their contribution decreasing beyond.

Analysis of the mean hotspot energies predicted by each hotspot predictor (first row in Figure 8.5a) shows a constant mean energy profile of hotspot energies within different hotregions. For the additive energy assumption (first row in Figure 8.5b) and the negative cooperativity assumption (second row in Figure 8.5b), a linear and exponential-like increase of energetic contribution from larger hotregions is shown respectively. For the positive cooperativity assumption, application of a linearly decreasing function on increasing hotregion sizes which have constant hotspot energies to start off with, results in a bell-shape contribution from hotregions. This suggests that maximum stability is provided by hotregion sizes of around 5; therefore, a saturation of hotregion contribution is achieved, beyond which larger hotregions do not necessarily increase complex stability.

## 8.5 Discussion

After confirming in the previous chapters of 6 and 7 that the change in the sum of the hotspot energies across an interface correlates to a change in off-rate, in this chapter, several aspects of hotspot distribution were studied in relation to the off-rate of a complex. Two previously reported properties of hotspots motivated this work; their tendency to occur at core interface regions (Bogan and Thorn, 1998), and their tendency to cluster into hotregions (Keskin et al., 2005). The advantage, if any, of these distributional properties to complex stability is therefore examined.

The off-rate of small complexes is more sensitive to mutations than that in large complexes. In section 8.1.3 it is shown that the higher sensitivity observed in small complexes, is a result of the increased importance of hotspots at the rim regions. Namely, mutations affecting these regions can have effects on the off-rate as significant as those in the core region. From the data available, in section 8.2, no evidence is found that complexes with more or larger hotregions have lower dissociation rates. In order to understand better the role of hotregions, the effects of cooperativity between hotspots in a hotregion are considered in

section 8.3. No prevalent form of cooperativity, positive or negative, is observed. Rather, interactions in hotregions are best described by the additivity rule. Finally in section 8.4, the contribution towards stability of hotregions of different sizes, is presented. Under a negative cooperativity assumption, hotregion contribution increases exponentially with its size. Under a positive cooperativity assumption, a bell-like contribution towards stability is observed. In this case, maximal stability is reached with hotregions of size five and reduces with sizes larger or smaller than this optimal size.

Our results in section 8.1 are best discussed in the light of the 'O-Ring' hypothesis. Hotspots are preferentially found in regions at the interface of low solvent accessibility (Bogan and Thorn, 1998). With this in mind, low solvent accessibility is necessary but not sufficient for high energy hotspots, as a number of residues with low to zero solvent accessibility may still not contribute significantly to binding. The O-Ring hypothesis, describes the protein interface as one where the stability critical residues are found at the core and are surrounded by a ring of energetically unimportant rim residues. The role of the ring is suggested to be secondary one; namely its purpose is to occlude the bulk solvent from the interactions at the core. This provides a lower effective dielectric constant for stronger electrostatic and hydrogen bonding interactions at the core. In our analysis presented in section 8.1, in line with this hypothesis, it is observed that mutations effecting hotspot at the core region have a significant correlative effect on the off-rate. This is in contrast to the other regions of the interface i.e. the rim and support regions. Under the O-ring hypothesis, it might be natural to think that if the distance between solvent and the core hotspots becomes smaller, then hotspot residues at the rim might have an increased role. Therefore, additional investigations were performed on small vs. large interfaces and on small vs. large interfaces. As shown in Figure 8.3, irrespective of interface area, mutations affecting hotspot regions at the rim are able to modulate the off-rate equally as those affecting core hotspot regions. This is in contrast to what is observed in large complexes, where mutations affecting rim hotspot energies are inconsequential to the off-rate. Therefore as it seems, rim residues do have a significant role for small complexes. Rim region residues are generally exposed

both in the unbound and bound states, but form inter-protein contacts in the complex state. Therefore it is unclear, through which underlying mechanism; mutations affecting rim hotspot energies are affecting the off-rate significantly. One mechanism might simply be the disruption of a strong inter-protein contact as a result of the mutation, irrespective of any effect on solvent shielding. The other might be that the inter-protein disruption increases the susceptibility to solvent entry. For additional validation, analysis of the rim regions using MD simulations simulating complex unbinding, for large and small complexes both with small-interface-areas may give further insights.

In section 8.2, no evidence is found that the disruption of large hotregions is critical to complex stability. Nor is it observed that having more hotregions increase complex stability. In the context of protein drug design such as the one presented in Chapter 4, finding mutations which are able to increase further the stability of the interaction is a daunting task. The hotspot representation may facilitate this process if hotspot distribution is completely understood. For example given a hotregion at an interface is it best to make mutations such that the existing hotregion grows larger in size, or is it more advantageous to create a new hotregion. Understanding if there is any advantage, when attempting to increase complex stability, in having larger hotregions, or more hotregions, would ultimately require analysis which controls for the number of hotspots, varies the number of hotregions or their size and assesses changes in the off-rate. However, current experimental data is limited in size and diversity for this to be performed comprehensively.

Counting of the ΔΔGs of all hotspots in hotregion may overestimate or underestimate the hotregions contribution if cooperativity between the hotspots exists. By designing functions which account for potential positive or negative cooperative effect, in section 8.3 no prevalent form of cooperativity is observed. No conclusion could be made as the results are highly dependent on the hotspot predictor generating the hotspots. With this in mind, the additivity assumption showed consistently higher correlations across different hotspot predictors, suggesting that hotregions are most additive in nature. This result, contrasts to that of Keskin et al. (2005), where it is proposed that hotregions are cooperative

in nature. One should note that this proposition is only based on the high-packing ratios of residues in a hotregion to those that are not. With this in mind, the results in of Figure 8.4 in section 8.3 are dependent on both the definition of a contact and that of a hotregion. There is no rule of thumb on how to define a contact or hotregion; in the work of Keskin et al. (2005), the distance between radii balls, with origins set on each C-α atom of the residues in question, is used to define a contact between two hotspot residues. A hotspot residue is added to a hotregion cluster if it is in contact with at least two existing hotspot residues. In this work, the definition uses a more fine-grain approach as a contact between two hotspot residues is created if any of their atoms are at a distance less than their van der Waals radii +0.5 Å (see methods section 2.3.7.5). Though for hotspot residues to be added in an existing hotregion, it only needs to be in contact with any other of the hotspot residues, and therefore might be a more lenient way of adding hotspot residues to a hotregion cluster, which in turn may render less packed hotregions. Other contact methods also include weighted contacts according to whether side-chain or backbone atoms are in contact (Reichmann et al., 2005). Most importantly, these different definitions generate different clusters of different packing ratios depending on their leniency and stringency, and therefore may affect the levels of cooperativity observed. Another factor which may account for the inconclusiveness regarding the more prevalent form of cooperativity is the modelling of cooperativity functions itself. Finding the right weights to apply to hotregions to account for cooperativity is not trivial, as experimental data (such as that found in Reichmann et al. (2005)) is not common enough to be able to learn cooperativity functions from experimental data. Last but not least, the diversity of interactions within the dataset may be better characterised with different cooperativity functions. Interestingly, this diversity of cooperative effects is also observed in the GA-FS runs performed on subsets of related complexes in Figure 7.6 of Chapter 7. Namely we observe that *HSEner_PosCoop*, *HSEner_NegCoop* and *Int_HS_Energy* tend to be important for different sets of related complexes in a mutually exclusive manner. This re-stresses the importance of detecting when a given type of cooperativity is present as much as it is important to model or account for it accurately.

# Chapter 9

## 9 Epilogue

In this thesis, the stability of protein-protein interactions is studied at different levels. Predictive models for the binding free energy and dissociation rate are built and the effect of mutations on both, characterised.

A number of themes reverberate throughout this thesis; firstly, that of conformational changes upon complex formation. Modelling the stability of such complexes remains a major challenge, as is that of characterizing mutations on such complexes. Hopefully the research presented in this thesis, sets a precedent for future models to come. For example, energetics calculated on a single 'snap-shot' of a bound complex can neither account for the conformational changes upon complex formation, nor for the delicate balance between enthalpic and entropic contributions involved. Unappreciated still in modelling stability, and somewhat related to conformational changes, is the binding and unbinding mechanisms at play. Transition states or encounter complexes play a critical role

in the stability of a complex. For example, mutations which stabilize an interaction might be doing so through increasing the stability of the transition state rather than the final bound state. Therefore characterization of binding and unbinding funnels, and the dynamics involved, will undoubtedly play a critical role in predicting the effects of mutations on protein-protein stability.

The machine learning framework is one which is consistently used throughout this work. The ease with which certain machine learning models can be used in 'black-box' fashion has unfortunately sometimes resulted in very dubious procedures and results throughout the years – some of which are mentioned in section 3.5. With powerful tools, comes greater responsibility and the hope is that this work highlights clearly these potential pitfalls, and responsibly avoids them. Be it with adapting cross-validation routines with domain knowledge, for example by using leave-complex-out validation routines; or by making sure the data on which predictor performance is stated, is not at any moment seen during any parameter optimization or feature selection routines. One aspect of machine learning modelling which is still unappreciated is their potential for understanding the mechanisms at play. This is not limited to just global feature importance measures. For example, the random forest algorithm may output descriptors which work hand-in-hand in the prediction. The use of such routines can help us understand the interplay between different determinants of stability. As attractive as linear models remain to the community, the inaccuracies and approximations in stability descriptors, the non-additivity of the physical determinants of affinity, and the diversity of protein-protein interactions, cannot be accounted for using linear modelling. Therefore, I believe that future efforts should shift towards exploiting the advantages of machine learning modelling, where learner choice is made with respect to the descriptors and data at hand and where visibility and interpretability share equal priority to that of predictive power.

I do hope that the work presented in this thesis, at the least improves upon the inaccuracies of previous methods, and at the most makes clear where the current limitations are, and which challenges must be addressed for further advances to be made. I always like to think of this work in the context of developments being

made in other important topics related to structural computational biology; including, protein structure prediction, docking and binding partner prediction. All of these share many similarities both in the underlying physical processes, and sometimes, in where we fail. Nevertheless, what is certain is that as these methods become more precise and efficient, their potential is nothing short of becoming an enabling technology for interpreting disease mutations, designing better drugs and uncovering further the nodes and links of the molecular networks governing our life processes.

# 10 Appendices

**Table 10.1. Hold out Proteins in Leave-Homology-OUT (LHO) Cross Validtion.**

For more stringent cross-validation mechanism, proteins which are from the same complex category (enzyme-inhibitor/antibody-antigen) or which share a common binding site, are put in the same test fold. Categories taken according to Moal and Fernandez-Recio (2012).

| **LHO Cross Validation Folds** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fold Category** | | Share binding site on same/homologous protein | Share binding site on same/homologous protein | Enzyme-Inhibitor | | Share binding site on same/homologous protein | Share binding site on same/homologous protein | | Antibody-Antigen | | Share binding site on same/homologous protein |
| **Mutation Count** | 58 | 62 | 79 | 39 | 74 | 87 | 63 | 36 | 84 | 100 | 31 |
| **PDB_IDs** | 1A22_A_B | 1A4Y_A_B | 1B2S_A_D | 1CBW_FGH_I | 1DAN_HL_UT | 1EMV_A_B | 1FC2_C_D | 1IAR_A_B | 1JRH_LH_I | 1JTG_A_B | 1KTZ_A_B |
| | | 1Z7X_W_X | 1B2U_A_D | 1GL0_E_I | | 1FR2_A_B | 1LFD_A_B | | 1NMB_N_LH | | 1REW_AB_C |
| | | | 1B3S_A_D | 1GL1_A_I | | 2GYK_A_B | 1MAH_A_F | | 2I26_N_L | | 2QJ9_AB_C |
| | | | 1BRS_A_D | 1TM1_E_I | | 2VLN_A_B | 1MQ8_A_B | | 2VIR_AB_C | | 2QJA_AB_C |
| | | | 1X1W_A_D | 2FTL_E_I | | 2VLO_A_B | 2AJF_A_E | | 2VIS_AB_C | | 2QJB_AB_C |
| | | | 1X1X_A_D | 2SIC_E_I | | 2VLP_A_B | 2B42_A_B | | 2VLJ_ABC_DE | | |
| | | | | | | 2VLQ_A_B | 2GOX_A_B | | 2VLR_ABC_DE | | |
| | | | | | | 2WPT_A_B | 3D5R_A_C | | 3HFM_HL_Y | | |
| | | | | | | | 3D5S_A_C | | | | |
| | | | | | | | 3BP8_A_C | | | | |
| | | | | | | | 3BK3_A_C | | | | |

## Table 10.2. ΔG Dataset

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 1A2K_C:AB | OG | 1QG4_A | Ran GTPase-GDP | 1OUN_AB | Nuclear transport factor 2 | 10681579 | 9.31 | 1.11 | ITC |
| 1ACB_E:I | EI | 4CHA_ABC | Chymotrypsin | 1EGL_A | Eglin C | 3071573 | 13.05 | 1.08 | Spectrophotometric inhibition assay |
| 1AHW_AB:C | A | 1FGN_LH | Fab 5g9 | 1TFH_A | Tissue factor | 9480775 | 11.55 | 0.69 | Competitive Inhibition assay |
| 1AK4_A:D | OX | 2CPL_A | Cyclophilin | 1E6J_P | HIV capsid | 9223641 | 6.43 | 1.33 | ITC |
| 1AKJ_AB:DE | OX | 2CLR_DE | MHC class 1 HLA-A2 | 1CD8_AB | T-cell CD8 coreceptor | 10072074 | 5.32 | 1.14 | SPR |
| 1ATN_A:D | OX | 1IJJ_B | Actin | 3DNI_A | Dnase I | 6244947 | 12.07 | 3.28 | Spectrophotometric inhibition assay |
| 1AVX_A:B | EI | 1QQU_A | Porcine trypsin | 1BA7_B | Soybean trypsin inhibitor | | 12.5 | 0.47 | Potentiometric |
| 1AVZ_B:C | OX | 1AVV_A | HIV-1-NEF protein | 1FYN_A | Fyn kinase SH3 domain | 9778343 | 6.55 | 0.73 | ITC |
| 1AY7_A:B | EI | 1RGH_B | Rnase SA | 1A19_B | Barstar | | 13.23 | 0.54 | Fluorescence inhibition assay |
| 1B6C_A:B | OX | 1D6O_A | FKBP binding protein | 1IAS_A | TGFbeta receptor | 11583628 | 8.94 | 1.96 | SPR |
| 1BJ1_HL:VW | AB | 1BJ1_HL | Fab - vEGF | 2VPF_GH | vEGF | 9753694 | 11.55 | 0.5 | SPR |
| 1BRS_A:D | EI | 1A2P_A | Barnase | 1A19_B | Barstar | 8507637 | 17.32 | 0.42 | Fluorescence inhibition assay |
| 1BUH_A:B | EI | 1HCL_A | CDK2 kinase | 1DKS_A | Ckshs1 | 8601310 | 9.7 | 0.75 | SPR |
| 1BVK_DE:F | A | 1BVL_BA | Fv Hulys11 | 3LZT_A | HEW lysozyme | 1560463 | 10.53 | 1.24 | Stopped-flow inhibition |
| 1BVN_P:T | EI | 1PIG_A | Alpha-amylase | 1HOE_A | Tendamistat | 14715318 | 15.06 | 0.87 | SPR |
| 1CBW_ABC:D | EI | 4CHA_ABC | Chymotrypsin | 9PTI_A | BPTI | 8784199 | 10.75 | 0.74 | Spectrophotometric inhibition assay |
| 1DE4_AB:CF | OX | 1A6Z_AB | Hemochromatosis protein HFE | 1CX8_AB | Transferrin receptor ectodom. | 11800564 | 9.78 | 2.59 | SPR |
| 1DFJ_E:I | EI | 9RSA_B | Ribonuclease A | 2BNH_A | Rnase inhibitor | 2271559 | 18.05 | 1.02 | Inhibition assay (indirect-Upa Hydrolysis) |
| 1DQJ_AB:C | A | 1DQQ_CD | Fab Hyhel63 | 3LZT_A | HEW lysozyme | 10828942 | 11.67 | 0.75 | SPR |
| 1E4K_AB:C | OR | 2DTQ_AB | FC fragment of human IgG 1 | 1FNL_A | Human FCGR III | 11544262 | 7.87 | 2.59 | SPR |
| 1E6E_A:B | ES | 1E1N_A | Adrenoxin reductase | 1CJE_D | Adrenoxin | 15181009 | 8.28 | 1.33 | SPR |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 1E6J_HL:P | A | 1E6O_HL | Fab 13B5 | 1A43_A | HIV-1 capsid protein p24 | 11080628 | 10.28 | 1.05 | SPR |
| 1E96_A:B | OG | 1MH1_A | Rac GTPase | 1HH8_A | p67 Phox | 11090627 | 7.42 | 0.71 | ITC |
| 1EFN_B:A | OX | 1AVV_A | HIV-1-NEF protein | 1FYN_A | Fyn kinase SH3 domain | 7588629 | 10.12 | 0.9 | SPR |
| 1EMV_A:B | EI | 1FSJ_B | Colicin E9 nuclease | 1IMQ_A | Im9 immunity protein | 7577967 | 18.58 | 1.28 | Stopped-flow fluormetry |
| 1EWY_A:C | ES | 1GJR_A | Ferredoxin reductase | 1CZP_A | Ferredoxin | 1910307 | 7.43 | 0.8 | Spectroscopic assay |
| 1EZU_C:AB | EI | 1TRM_A | D102N Trypsin | 1ECZ_AB | Y69F D70P Ecotin | 9642073 | 13.77 | 1.21 | Spectroscopic inhibition assay |
| 1F34_A:B | EI | 4PEP_A | Porcine pepsin | 1F32_A | Ascaris inhibitor 3 | 4594130 | 14.19 | 0.93 | Spectroscopic inhibition assay |
| 1F6M_A:C | ES | 1CL0_A | Thioredoxin reductase | 2TIR_A | Thioredoxin 1 | 19933368 | 7.6 | 4.9 | |
| 1FC2_C:D | OX | 1BDD_A | Staphylococcus Protein A | 1FC1_AB | Human Fc fragment | 7646442 | 10.43 | 1.69 | Stopped-flow fluorescence |
| 1FFW_A:B | OX | 3CHY_A | Chemotaxis protein CheY | 1FWP_A | Chemotaxis protein CheA | 8377825 | 8.09 | 1.43 | ITC |
| 1FLE_E:I | EI | 9EST_A | Elastase | 2REL_A(4) | Elafin | 2394696 | 12.28 | 1.02 | Inhibition assay |
| 1FQJ_A:B | OG | 1TND_C | Gt-alpha | 1FQI_A | RGS9 | 10085118 | 9.79 | 0.91 | Fluorescence spectroscopy |
| 1FSK_BC:A | AB | 1FSK_BC | Fab - Birch pollen antigen Bet V1 | 1BV1_A | Birch pollen antigen Bet V1 | | 13.12 | 0.45 | SPR |
| 1GCQ_B:C | OX | 1GRI_B | GRB2 C-ter SH3 domain | 1GCP_B | Vav N-ter SH3 domain | 11406576 | 6.51 | 0.92 | SPR |
| 1GL1_A:I | EI | 4CHA_ABC | Chymotrypsin | 1PMC_A(6) | PMP-C (LCMI II) | 7592720 | 13.23 | 1.2 | Inhibition assay |
| 1GLA_G:F | ER | 1BU6_0 | Glycerol Kinase | 1F3Z_A | Glucose specific IIIGlc | 9538005 | 6.76 | 0.98 | Spectroscopy |
| 1GPW_A:B | OX | 1THF_D | HISF protein | 1K9V_F | Amidotransferase HISH | | 11.32 | 0.65 | Fluorescence Titration |
| 1GRN_A:B | OG | 1A4R_A | CDC42 GTPase | 1RGP_A | CDC42 GAP | 9468490 | 9.03 | 1.22 | Fluorescence Spectroscopy |
| 1GXD_A:C | EI | 1CK7_A | ProMMP2 type IV collagenase | 1BR9_A | Metalloproteinase inhibitor 2 | 9368077 | 11.3 | 1.39 | SPR |
| 1H1V_A:G | OX | 1IJJ_B | Actin | 1P8X_A | Gelsolin precursor C-term | 2836434 | 10.2 | 1.05 | Fluorescence spectroscopy |
| 1H9D_A:B | OX | 1EAN_A | Runx1 domain of CBFalpha1 | 1ILF_A(1) | Dimerisation domain of CBF-beta | 10984496 | 9.18 | 1.32 | Electrophoretic mobility shift assays |
| 1HCF_AB:X | OR | 1B98_AM | Neurotrophin-4 | 1WWB_X | TrkB-d5 growth factor receptor | 11855816 | 13.08 | 0.88 | SPR |
| 1HE8_B:A | OG | 821P_A | Ras GTPase | 1E8Z_A | PIP3 kinase | 11136978 | 7.37 | 0.92 | Stopped-flow fluometry |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 1HIA_AB:I | EI | 2PKA_XY | Kallikrein | 1BX8_A | Hirustatin | 8112345 | 10.76 | 1.4 | Inhibition assay |
| 1I2M_A:B | OG | 1QG4_A | Ran GTPase-GDP | 1A12_A | RCC1 | 7548002 | 15.83 | 2.12 | Stopped-flow fluometry |
| 1I4D_D:AB | OG | 1MH1_A | Rac GTPase | 1I49_AB | Arfaptin | 11346801 | 7.46 | 1.41 | ITC |
| 1IB1_AB:E | OX | 1QJB_AB | 14-3-3 protein | 1KUY_A | Serotonin N-acteylase | 11336675 | 9.76 | 2.09 | Sedimentation equilibrium |
| 1IQD_AB:C | AB | 1IQD_AB | Fab - Factor VIII domain C2 | 1D7P_M | Factor VIII domain C2 | 9657749 | 15 | 0.48 | SPR |
| 1J2J_A:B | OG | 1O3Y_A | Arf1 GTPase.GNP-RanBD1 | 1OXZ_A | GAT domain of GGA1 | 12679809 | 8.13 | 0.63 | SPR |
| 1JIW_P:I | EI | 1AKL_A | Alkaline metallo-proteinase | 2RN4_A(1) | Proteinase inhibitor | 10770939 | 15.55 | 2.07 | Inhibition assay |
| 1JMO_A:HL | ER | 1JMJ_A | Heparin cofactor | 2CN0_HL | Thrombin | 9162031 | 9.47 | 3.21 | Inhibition assay |
| 1JPS_HL:T | A | 1JPT_HL | Fab D3H44 | 1TFH_B | Tissue factor | 11307801 | 13.64 | 0.51 | SPR |
| 1JTG_B:A | EI | 3GMU_B | beta-lactamase inhibitor protein | 1ZG4_A | beta-lactamase TEM-1 | 9890878 | 12.82 | 0.49 | SPR |
| 1JWH_CD:A | ER | 3EED_AB | Casein kinase II beta chain | 3C13_A | Casein kinase II alpha chain | 18824508 | 11.14 | 1.27 | ITC |
| 1K5D_AB:C | OG | 1RRP_AB | Ran GTPase | 1YRG_B | Ran GAP | 14585972 | 12.77 | 1.19 | Stopped-flow fluorescence |
| 1KAC_A:B | OR | 1NOB_F | Adenovirus fiber knob protein | 1F5W_B | Adenovirus receptor | 10684297 | 10.68 | 0.95 | SPR |
| 1KKL_ABC:H | ES | 1JB1_ABC | HPr kinase C-ter domain | 2HPR_A | HPr | 12009882 | 10.02 | 2.2 | SPR |
| 1KLU_AB:D | OX | 1H15_AB | MHC class 2 HLA-DR1 | 1STE_A | Staphylococcus enterotoxin C3 | 10229190 | 7.28 | 0.43 | SPR |
| 1KTZ_A:B | OR | 1TGK_A | TGF-beta | 1M9Z_A | TGF-beta receptor | 16300789 | 8.92 | 0.39 | SPR |
| 1KXP_A:D | OX | 1IJJ_B | Actin | 1KW2_B | Vitamin D binding protein | 2910852 | 12.34 | 1.12 | Inhibition assay |
| 1KXQ_H:A | AB | 1KXQ_H | Camel VHH - Pancreatic alpha-amylase | 1PPI_A | Pancreatic alpha-amylase | 9649422 | 11.54 | 0.72 | SPR |
| 1LFD_B:A | OG | 5P21_A | Ras.GNP | 1LXD_A | RalGDS Ras-interacting domain | 15197281 | 7.79 | 1.79 | Stopped-flow fluorescence |
| 1M10_A:B | ER | 1AUQ_A | Von Willebrand Factor dom. A1 | 1M0Z_B | Glycoprotein IB-alpha | 12183630 | 11.24 | 2.1 | SPR |
| 1MAH_A:F | EI | 1J06_B | Acetylcholinesterase | 1FSC_A | Fasciculin | 8509385 | 14.51 | 0.61 | Inhibition assay |
| 1MLC_AB:E | A | 1MLB_AB | Fab44.1 | 3LZT_A | HEW lysozyme | 10229844 | 9.61 | 0.6 | SPR |
| 1MQ8_A:B | OX | 1IAM_A | ICAM-1 domain 1-2 | 1MQ9_A | Integrin alpha-L I domain | 12526797 | 7.53 | 1.76 | SPR |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 1NB5_AP:I | EI | 8PCH_A | Cathepsin H | 1DVC_A | Stefin A | 8898076 | 13.86 | 1.58 | Inhibition assay |
| 1NCA_HL:N | AB | 1NCA_HL | Fab - Flu virus neuraminidase N9 | 7NN9_A | Flu virus neuraminidase N9 | 9692956 | 11.02 | 0.24 | Fluorescence inhibition assay |
| 1NSN_HL:S | AB | 1NSN_HL | Fab N10 - Staphylococcal nuclease | 1KDC_A | Staphylococcal nuclease | 1704035 | 14 | 0.35 | ELISA inhibiton assay |
| 1NVU_Q:S | OG | 1LF0_A | Ras GTPase.GTP | 2II0_B | Son of sevenless | 15507210 | 7.43 | 1.98 | Fluorescence anisotropy |
| 1NVU_R:S | OG | 1LF0_A | Ras GTPase.GTP | 2II0_B | Son of sevenless | 15507210 | 7.8 | 3.09 | |
| 1OPH_A:B | EI | 1QLP_A | Alpha-1-antitrypsin | 2PTN_A | Trypsin | 9012804 | 11.32 | 1.2 | Fluorescence inhibition assay |
| 1P2C_AB:C | A | 2Q76_AB | FabF10.6.6 | 3LZT_A | HEW lysozyme | 14988501 | 13.63 | 0.46 | SPR |
| 1PPE_E:I | EI | 2PTN_A | Trypsin | 1LU0_A | CMTI-1 squash inhibitor | 8543044 | 15.56 | 0.34 | Spectrophotometric inhibition assay |
| 1PVH_A:B | OR | 1BQU_A | IL6 receptor beta chain D2-D3 domains | 1EMR_A | Leukemia inhibitory factor | 14527405 | 9.52 | 0.34 | ITC |
| 1PXV_A:C | EI | 1X9Y_A | Staphylococcus aureus cystein protease | 1NYC_A | Cystein protease inhibitor | 17261086 | 12.97 | 2.63 | Inhibition assay |
| 1QA9_A:B | OX | 1HNF_A | CD2 | 1CCZ_A | CD58 | 7520278 | 7.16 | 0.73 | SPR |
| 1R0R_E:I | EI | 1SCN_E | Subtilisin carlsberg | 2GKR_I | OMTKY | 7046785 | 14.17 | 0.45 | Spectrophotometry |
| 1R6Q_A:C | ER | 1R6C_X | Clp protease subunit ClpA | 2W9R_A | Clp protease adaptor protein ClpS | 12426582 | 8.84 | 1.67 | SPR |
| 1RLB_ABCD:E | OX | 2PAB_ABCD | Transthyretin | 1HBP_A | Retinol binding protein | 8639713 | 8.18 | 0.66 | Fluorescence anisotropy |
| 1RV6_VW:X | OR | 1FZV_AB | PIGF receptor binding domain | 1QSZ_A | Flt1 protein domain 2 | 8822205 | 13.86 | 1.09 | Inhibition assay |
| 1S1Q_A:B | OX | 2F0R_A | UEV domain | 1YJ1_A | Ubiquitin | 12006492 | 4.29 | 0.98 | SPR |
| 1T6B_X:Y | OR | 1ACC_A | Anthrax protective antigen | 1SHU_X | Anthrax toxin receptor | 15044490 | 13.1 | 0.62 | Stopped-flow fluorescence |
| 1US7_A:B | ER | 2FXS_A | Heat shock protein 82 N-ter domain | 2W0G_A | HSP90 co-chaperone CDC37 C-ter domain | 14718169 | 8.09 | 1.06 | ITC |
| 1UUG_A:B | EI | 3EUG_A | Uracyl-DNA glycosylase | 2UGI_B | Glycosylase inhibitor | 8262921 | 18 | 0.77 | Stopped-flow fluorescence |
| 1VFB_AB:C | A | 1VFA_AB | Fv D1.3 | 8LYZ_A | HEW lysozyme | 8302837 | 11.46 | 1.02 | ITC |
| 1WDW_BD:A | ER | 1V8Z_AB | Tryptophan synthase beta chain 1 | 1GEQ_A | Tryptophan synthase alpha chain | 12643278 | 12.72 | 1.29 | ITC |
| 1WEJ_HL:F | A | 1QBL_HL | Fab E8 | 1HRC_A | Cytochrome C | 2993413 | 12.48 | 0.31 | Spectroscopic inhibition assay |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 1WQ1_R:G | OG | 6Q21_D | Ras GTPase.GDP | 1WER_A | Ras GAP | 8262937 | 6.62 | 1.16 | Fluorescence |
| 1XD3_A:B | OX | 1UCH_A | UCH-L3 | 1YJ1_A | Ubiquitin | 9485312 | 8.9 | 1.24 | Fluorescence spectrophotometry |
| 1XQS_A:C | OX | 1XQR_A | HspBP1 | 1S3X_A | Hsp70 ATPase domain | 15694338 | 7.08 | 1.77 | SPR |
| 1XU1_ABD:T | OR | 1U5Y_ABD | TNF domain of APRIL | 1XUT_A(11) | TNF receptor superfamily member 13B TACI CRD2 domain | 10956646 | 11.18 | 1.3 | SPR |
| 1YVB_A:I | EI | 2GHU_A | Falcipain 2 | 1CEW_I | Cystatin | 17502099 | 11.17 | 0.51 | Inhibition assay |
| 1Z0K_A:B | OG | 2BME_A | Rab4A GTPase.GNP | 1YZM_A | RAB4 binding domain of Rabenosyn | 16034420 | 6.98 | 0.53 | SPR |
| 1ZM4_A:B | ES | 1N0V_C | Elongation factor 2 | 1XK9_A | Diphtheria toxin A catalytic domain | 12270928 | 8.03 | 2.94 | Flourescence |
| 2A9K_A:B | ES | 1U8Z_A | Ral-A.GDP | 2C8B_X | Mono-ADP-ribosyltransferase C3 | 16177825 | 10.25 | 0.85 | ITC |
| 2ABZ_B:E | EI | 3I1U_A | Carboxypeptidase A1 | 1ZFI_A(1) | Leech carboxypeptidase inhibitor | 16126224 | 11.67 | 0.9 | Spectroscopic inbition assay |
| 2AJF_A:E | OR | 1R42_A | Angiotensin-converting enzyme 2 | 2GHV_E | SARS spike protein receptor binding domain | 15791205 | 10.63 | 0.65 | SPR |
| 2AQ3_A:B | OX | 1BEC_A | TCR Vbeta8.2 | 1CK1_A | SEC3 | 20836565 | 6.71 | 1.82 | ITC |
| 2B42_A:B | EI | 2DCY_A | Xylanase | 1T6E_X | Xylanase inhibitor | 16279951 | 12.11 | 0.72 | SPR |
| 2B4J_AB:C | OX | 1BIZ_AB | Integrase (HIV-1) | 1Z9E_A(1) | PC4 and SFRS1 interacting protein | 19801648 | 10.86 | 0.99 | Fluorescence inhibition assay |
| 2BTF_A:P | OX | 1IJJ_B | Actin | 1PNE_A | Profilin | 9788869 | 7.69 | 0.75 | Inhibition assay |
| 2C0L_A:B | OX | 1FCH_A | TRP region of PEX5 | 1C44_A | Sterol carrier protein 2 | 17157249 | 9.82 | 2.62 | ITC |
| 2FJU_B:A | OG | 2ZKM_X | Phospholipase beta 2 | 1MH1_A | Rac GTPase | 12657629 | 7.2 | 1.04 | SPR |
| 2GOX_A:B | OX | 1C3D_A | Complement C3d fragment | 2GOM_A | Staphylococcus aureus Efb-C | 18687868 | 12.08 | 0.6 | SPR |
| 2HLE_A:B | OR | 2BBA_A | Ephrin B4 receptor | 1IKO_P | Ephrin B2 ectodomain | 16472751 | 10.09 | 1.4 | ITC |
| 2HQS_A:H | OX | 1CRZ_A | TolB | 1OAP_A | Pal | 17375930 | 10.15 | 1.14 | ITC |
| 2HRK_A:B | OX | 2HRA_A | Glutamyl-t-RNA synthetase | 2HQT_A | GU-4 nucleic binding protein | 17976650 | 10.98 | 2.03 | SPR |
| 2I25_N:L | A | 2I24_N | Shark single domain antigen receptor | 3LZT_A | HEW lysozyme | 16446445 | 12.28 | 1.21 | SPR |
| 2I9B_E:A | OR | 1YWH_A | Urokinase plasminogen activator surface receptor | 2I9A_A | Urokinase-type plasminogen activator | 15003263 | 12.93 | 3.79 | SPR |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 2J0T_A:D | EI | 966C_A | MMP1 Intersitial collagenase | 1D2B_A(20) | Metalloproteinase inhibitor 1 | 12515831 | 13.34 | 1.23 | Fluorescence inhibition assay |
| 2JEL_HL:P | AB | 2JEL_HL | Fab Jel42 - HPr | 1POH_A | HPr | 9671548 | 11.59 | 0.17 | Fluorescence inhibition assay |
| 2MTA_HL:A | ES | 2BBK_JM | Methylamine dehydrogenase | 2RAC_A | Amicyanin | 8347660 | 7.42 | 0.41 | Spectroscopic inhibition assay |
| 2NYZ_AB:D | OR | 1MKF_AB | Viral chemokine binding p. M3 | 1J9O_A | Chemokine XCL1 | 18070938 | 12.69 | 2.09 | SPR |
| 2O3B_A:B | EI | 1ZM8_A | NucA nuclease | 1J57_A | NuiA nuclease inhibitor | 17138564 | 15.68 | 3.13 | Inhibition assay |
| 2OOB_A:B | ES | 2OOA_A | E3 ubiquitin-protein ligase CBL-B UBA domain | 1YJ1_A | Ubiquitin | 17897937 | 5.66 | 0.85 | ITC |
| 2OOR_AB:C | ER | 1L7E_AB | NAD(P) transhydrogenase subunit alpha part 1 | 1E3T_A | NAD(P) transhydrogenase subunit beta | 8898902 | 10.65 | 1.42 | Fluorescence |
| 2PCB_A:B | ES | 1CCP_A | Cyt C peroxidase | 1HRC_A | Cytochrome C | 9092837 | 6.82 | 0.45 | ITC |
| 2PCC_A:B | ES | 1CCP_A | Cyt C peroxidase | 1YCC_A | Cytochrome C, yeast | 11148036 | 7.91 | 0.39 | ITC |
| 2PTC_E:I | EI | 2PTN_A | Trypsin | 9PTI_A | BPTI | 5041905 | 18.04 | 0.28 | Inhibition assay |
| 2SIC_E:I | EI | 1SUP_A | Subtilisin | 3SSI_A | Streptomyces subtilisin inhibitor | 32173 | 13.84 | 0.36 | Fluorescence titration |
| 2SNI_E:I | EI | 1UBN_A | Subtilisin | 2CI2_I | Chymotrypsin inhibitor 2 | 10065709 | 15.96 | 0.35 | Inhibition assay |
| 2TGP_Z:I | EI | 1TGB_A | Trypsinogen | 9PTI_A | BPTI | 311834 | 7.54 | 0.57 | Spectroscopic inhibition assay |
| 2UUY_A:B | EI | 2PTN_A | Trypsin | 2UUX_A | Tryptase inhibitor from tick | 17391695 | 11.26 | 0.44 | Inhibition assay |
| 2VDB_A:B | OX | 3CX9_A | Serum albumin | 2J5Y_A | Peptostreptococcalalbumin-binding protein | 8900134 | 13.4 | 0.47 | Radioligand inhibition assay |
| 2VIR_AB:C | A | 1GIG_LH | Fab | 2HMG_AB | Flu virus hemagglutinin | 9461077 | 12.28 | 0.8 | SPR |
| 2VIS_AB:C | A | 1GIG_LH | Fab | 2VIU_ACE | Flu virus hemagglutinin | 9461077 | 7.36 | 0.8 | SPR |
| 2WPT_A:B | EI | 1FSJ_B | Colicin E9 nuclease | 2NO8_A | Im2 immunity protein | 9718299 | 10.67 | 1.61 | Stopped-flow fluormetry |
| 3BP8_AB:C | OX | 1Z6R_AB | Mlc transcription regulator | 3BP3_A | PTS glucose-specific enzyme EIICB | 18319344 | 11.44 | 0.45 | SPR |
| 3BZD_A:B | OX | 1BEC_A | TCR Vbeta8.2 | 3BVZ_A | SEC3-1A4 | 20836565 | 9.57 | 1.08 | ITC |
| 3CPH_G:A | OG | 1G16_A | Ras-related protein Sec4 | 3CPI_G | Rab GDP-dissociation inhibitor | 18426803 | 8.84 | 2.12 | ITC |
| 3SGB_E:I | EI | 2QA9_E | Streptogrisin B | 2OVO_A | Ovomucoid inhibitor third domain | 3555488 | 14.51 | 0.36 | Spectrophotometric inhibition assay |

| Complex PDB / Chains | Type | Unbound PDB | Protein 1 | Unbound PDB | Protein 2 | Pubmed ID | -ΔG | I-RMSD | Method |
|---|---|---|---|---|---|---|---|---|---|
| 4CPA_A:I | EI | 8CPA_A | Carboxypeptidase A | 1H20_A(9) | Potato carboxypeptidase inhibitor | 4415398 | 11.32 | 1.52 | Spectrophotometric inhibition assay |

## Table 10.3. $\Delta k_{off}$ Dataset

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1TM1_E_I | YI61A | 2.56E-04 | 1.06E-05 | 1.38 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | YI61G | 1.88E-02 | 1.06E-05 | 3.25 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI65A | 3.47E-04 | 6.10E-06 | 1.75 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI67A | 3.15E-04 | 6.10E-06 | 1.71 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI67C | 3.03E-04 | 6.10E-06 | 1.7 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI67A,RI65A | 7.92E-03 | 1.06E-05 | 2.87 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | TI58D | 3.40E-05 | 3.90E-06 | 0.94 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | TI58A | 2.06E-04 | 3.90E-06 | 1.72 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | TI58D,EI60A | 1.07E-05 | 3.90E-06 | 0.44 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | VI70A | 3.40E-06 | 3.90E-06 | -0.06 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 2SIC_E_I | MI73E | 1.10E-04 | 9.00E-05 | 0.09 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73D | 2.60E-04 | 9.00E-05 | 0.46 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73H | 4.10E-04 | 9.00E-05 | 0.66 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73G | 1.30E-04 | 9.00E-05 | 0.16 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73A | 1.40E-04 | 9.00E-05 | 0.19 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73L | 2.10E-04 | 9.00E-05 | 0.37 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73V | 3.50E-04 | 9.00E-05 | 0.59 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2SIC_E_I | MI73I | 1.70E-03 | 9.00E-05 | 1.28 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 1IAR_A_B | IA5A | 1.40E-02 | 2.10E-03 | 0.82 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | IA5R | 8.70E-03 | 2.10E-03 | 0.62 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | TA6A | 1.90E-03 | 2.10E-03 | -0.04 | 9050834 | Interleukin-4 | Interleukin-4 receptor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1IAR_A_B | TA6D | 1.50E-02 | 2.10E-03 | 0.85 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA8A | 2.50E-03 | 2.10E-03 | 0.08 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA8R | 1.90E-03 | 2.10E-03 | -0.04 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | EA9Q | 2.70E-01 | 2.10E-03 | 2.11 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | IA11A | 2.00E-03 | 2.10E-03 | -0.02 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA12S | 1.90E-03 | 2.10E-03 | -0.04 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA12E | 1.50E-03 | 2.10E-03 | -0.15 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | TA13A | 7.10E-03 | 2.10E-03 | 0.53 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | TA13D | 8.50E-04 | 2.10E-03 | -0.39 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | NA15A | 2.30E-03 | 2.10E-03 | 0.04 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | NA15D | 1.70E-03 | 2.10E-03 | -0.09 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | SA16A | 1.90E-03 | 2.10E-03 | -0.04 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | SA16D | 1.50E-03 | 2.10E-03 | -0.15 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | EA19A | 1.70E-03 | 2.10E-03 | -0.09 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | EA19R | 1.60E-03 | 2.10E-03 | -0.12 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA53Q | 7.30E-03 | 2.10E-03 | 0.54 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA77A | 2.10E-03 | 2.10E-03 | 0 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA77E | 2.00E-03 | 2.10E-03 | -0.02 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA78A | 2.20E-03 | 2.10E-03 | 0.02 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA78E | 2.70E-03 | 2.10E-03 | 0.11 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA81A | 2.80E-03 | 2.10E-03 | 0.13 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA81E | 6.10E-03 | 2.10E-03 | 0.46 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | FA82A | 2.10E-03 | 2.10E-03 | 0 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | FA82D | 7.30E-04 | 2.10E-03 | -0.46 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA84A | 2.90E-03 | 2.10E-03 | 0.14 | 9050834 | Interleukin-4 | Interleukin-4 receptor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1IAR_A_B | KA84D | 9.30E-03 | 2.10E-03 | 0.65 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA85A | 2.70E-03 | 2.10E-03 | 0.11 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA85E | 4.60E-03 | 2.10E-03 | 0.34 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA88Q | 1.40E-01 | 2.10E-03 | 1.82 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA88A | 7.60E-01 | 2.10E-03 | 2.56 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | NA89A | 2.70E-02 | 2.10E-03 | 1.11 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | WA91A | 6.10E-03 | 2.10E-03 | 0.46 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | WA91D | 8.50E-03 | 2.10E-03 | 0.61 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1BRS_A_D | RA59A | 2.40E-03 | 3.73E-06 | 2.81 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q | 1.00E-02 | 3.73E-06 | 3.43 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A | 1.70E-02 | 3.73E-06 | 3.66 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A | 1.29E-01 | 3.73E-06 | 4.54 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | YD29F | 2.40E-06 | 3.73E-06 | -0.19 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | YD29A | 1.00E-03 | 3.73E-06 | 2.43 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | DD35A | 3.80E-03 | 3.73E-06 | 3.01 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | WD38F | 7.00E-05 | 3.73E-06 | 1.27 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | DD39A | 9.00E-01 | 3.73E-06 | 5.38 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | TD42A | 7.20E-05 | 3.73E-06 | 1.29 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | WD44F | 3.40E-06 | 3.73E-06 | -0.04 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | ED76A | 2.10E-05 | 3.73E-06 | 0.75 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | ED80A | 5.20E-06 | 3.73E-06 | 0.14 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,YD29A | 9.70E-01 | 3.73E-06 | 5.42 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,DD35A | 3.60E+00 | 3.73E-06 | 5.98 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,WD38F | 2.10E-02 | 3.73E-06 | 3.75 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,DD39A | 6.80E-01 | 3.73E-06 | 5.26 | 7739054 | Barnase | Barstar |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1BRS_A_D | KA27A,TD42A | 6.80E-03 | 3.73E-06 | 3.26 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,ED76A | 1.30E-02 | 3.73E-06 | 3.54 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | KA27A,ED80A | 3.50E-03 | 3.73E-06 | 2.97 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,YD29A | 2.50E-01 | 3.73E-06 | 4.83 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,DD35A | 1.40E-02 | 3.73E-06 | 3.57 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,WD38F | 1.30E-02 | 3.73E-06 | 3.54 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,TD42A | 2.30E-02 | 3.73E-06 | 3.79 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,ED76A | 1.60E-03 | 3.73E-06 | 2.63 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A,ED80A | 2.00E-03 | 3.73E-06 | 2.73 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,YD29A | 1.10E+00 | 3.73E-06 | 5.47 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,DD35A | 7.10E+00 | 3.73E-06 | 6.28 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,DD39A | 5.30E-02 | 3.73E-06 | 4.15 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,TD42A | 3.50E-02 | 3.73E-06 | 3.97 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,ED76A | 3.40E-02 | 3.73E-06 | 3.96 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA83Q,ED80A | 1.10E-02 | 3.73E-06 | 3.47 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,YD29A | 1.30E+00 | 3.73E-06 | 5.54 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,WD38F | 2.75E-01 | 3.73E-06 | 4.87 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,DD39A | 3.00E-01 | 3.73E-06 | 4.91 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,TD42A | 3.10E-01 | 3.73E-06 | 4.92 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,ED76A | 7.40E-02 | 3.73E-06 | 4.3 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A,ED80A | 2.40E-02 | 3.73E-06 | 3.81 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,YD29A | 1.50E-01 | 3.73E-06 | 4.6 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,YD29F | 4.50E-02 | 3.73E-06 | 4.08 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,WD38F | 1.28E+00 | 3.73E-06 | 5.54 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,DD39A | 1.70E+01 | 3.73E-06 | 6.66 | 7739054 | Barnase | Barstar |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1BRS_A_D | HA102A,TD42A | 2.40E+00 | 3.73E-06 | 5.81 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,ED76A | 5.90E-01 | 3.73E-06 | 5.2 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A,ED80A | 1.80E-01 | 3.73E-06 | 4.68 | 7739054 | Barnase | Barstar |
| 1B2U_A_D | AA27K | 3.80E-03 | 3.60E+00 | -2.98 | 7739054 | Barnase | Barstar |
| 1B2U_A_D | AD36D | 4.50E-03 | 3.60E+00 | -2.9 | 7739054 | Barnase | Barstar |
| 1B2U_A_D | AA27K,AD36D | 3.70E-06 | 3.60E+00 | -5.99 | 7739054 | Barnase | Barstar |
| 1B2S_A_D | AA27K | 7.20E-05 | 6.80E-03 | -1.98 | 7739054 | Barnase | Barstar |
| 1B2S_A_D | AD43T | 4.50E-03 | 6.80E-03 | -0.18 | 7739054 | Barnase | Barstar |
| 1B2S_A_D | AA27K,AD43T | 3.70E-06 | 6.80E-03 | -3.26 | 7739054 | Barnase | Barstar |
| 1B3S_A_D | AA102H | 2.40E-06 | 4.50E-02 | -4.27 | 7739054 | Barnase | Barstar |
| 1B3S_A_D | FD30Y | 1.29E-01 | 4.50E-02 | 0.46 | 7739054 | Barnase | Barstar |
| 1B3S_A_D | AA102H,FD30Y | 3.70E-06 | 4.50E-02 | -4.09 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | DD35A | 2.73E-02 | 1.15E-04 | 2.38 | | Barnase | Barstar |
| 1BRS_A_D | DD39A | 4.57E-01 | 1.15E-04 | 3.6 | | Barnase | Barstar |
| 1BRS_A_D | ED80A | 2.28E-04 | 1.15E-04 | 0.3 | | Barnase | Barstar |
| 1X1W_A_D | AD80E | 1.15E-04 | 2.28E-04 | -0.3 | | Barnase | Barstar |
| 1X1X_A_D | AD76E | 1.15E-04 | 4.75E-04 | -0.62 | | Barnase | Barstar |
| 1BRS_A_D | KA27A | 6.60E-03 | 8.00E-06 | 2.92 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | WA35F | 8.00E-05 | 8.00E-06 | 1 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | DA54A | 5.30E-06 | 8.00E-06 | -0.18 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | NA58A | 6.40E-04 | 8.00E-06 | 1.9 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA59A | 3.70E-03 | 8.00E-06 | 2.67 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA60A | 3.40E-05 | 8.00E-06 | 0.63 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA73A | 7.40E-04 | 8.00E-06 | 1.97 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA87A | 6.70E-02 | 8.00E-06 | 3.92 | 8494892 | Barnase | Barstar |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1BRS_A_D | HA102A | 2.20E-01 | 8.00E-06 | 4.44 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | KA27A | 6.10E-03 | 1.50E-05 | 2.61 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | WA35F | 9.60E-05 | 1.50E-05 | 0.81 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | DA54A | 1.80E-05 | 1.50E-05 | 0.08 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | NA58A | 9.10E-04 | 1.50E-05 | 1.78 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA59A | 6.50E-03 | 1.50E-05 | 2.64 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA60A | 1.30E-04 | 1.50E-05 | 0.94 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA73A | 2.60E-03 | 1.50E-05 | 2.24 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | HA102A | 3.50E-01 | 1.50E-05 | 4.37 | 8494892 | Barnase | Barstar |
| 2B42_A_B | HA374A | 1.59E-03 | 3.60E-04 | 0.65 | 16279951 | TAXI-I | B. subtilis endoxylanase |
| 2B42_A_B | HA374Q | 1.18E-03 | 3.60E-04 | 0.52 | 16279951 | TAXI-I | B. subtilis endoxylanase |
| 2B42_A_B | HA374K | 3.44E-03 | 3.60E-04 | 0.98 | 16279951 | TAXI-I | B. subtilis endoxylanase |
| 2I26_N_L | AN30V | 1.60E-03 | 2.00E-03 | -0.1 | 16446445 | Type II IgNAR | HEW Lysozyme |
| 2I26_N_L | SN61R | 1.20E-03 | 2.00E-03 | -0.22 | 16446445 | Type II IgNAR | HEW Lysozyme |
| 2GOX_A_B | RB131A | 6.87E-02 | 5.63E-04 | 2.09 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 2GOX_A_B | NB138A | 2.14E-02 | 5.63E-04 | 1.58 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 3D5S_A_C | AC41R | 5.63E-04 | 6.87E-02 | -2.09 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 3BP8_A_C | FA136A | 1.30E-02 | 3.85E-03 | 0.53 | 18319344 | Mlc transcription regulator | PTS glucose-specific enzyme EIICB |
| 3BP8_A_C | AC63F | 8.79E-03 | 3.85E-03 | 0.36 | 18319344 | Mlc transcription regulator | PTS glucose-specific enzyme EIICB |
| 3BP8_A_C | FA136A,AC63F | 2.25E-03 | 3.85E-03 | -0.23 | 18319344 | Mlc transcription regulator | PTS glucose-specific enzyme EIICB |
| 2VIS_AB_C | IC131T | 1.10E-04 | 2.16E-03 | -1.29 | 9461077 | IgG1 lambda FAB | Flu virus hemagglutinin |
| 2VIR_AB_C | TC131I | 2.16E-03 | 1.10E-04 | 1.29 | 9461077 | IgG1 lambda FAB | Flu virus hemagglutinin |
| 2WPT_A_B | DA33L | 3.80E-03 | 7.30E-01 | -2.28 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 2WPT_A_B | NA34V | 1.60E-01 | 7.30E-01 | -0.66 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RA38T | 2.90E-01 | 7.30E-01 | -0.4 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | EA39H | 8.80E-01 | 7.30E-01 | 0.08 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RA42E | 5.00E-01 | 7.30E-01 | -0.16 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NA34V,RA38T | 1.80E-02 | 7.30E-01 | -1.61 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | DA33L,NA34V,RA38T | 3.70E-05 | 7.30E-01 | -4.3 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NA34V,RA38T,RA42E | 1.20E-02 | 7.30E-01 | -1.78 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NA34V,RA38T,EA39H,RA42E | 1.30E-02 | 7.30E-01 | -1.75 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | EA30A | 2.80E-07 | 7.30E-01 | -6.42 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | DA33A | 1.20E-08 | 7.30E-01 | -7.78 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NA34A | 7.90E-09 | 7.30E-01 | -7.97 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | VA37A | 9.30E-06 | 7.30E-01 | -4.89 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RA38A | 2.30E-09 | 7.30E-01 | -8.5 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | EA41A | 3.00E-05 | 7.30E-01 | -4.39 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RA42A | 1.00E-08 | 7.30E-01 | -7.86 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SA50A | 9.00E-07 | 7.30E-01 | -5.91 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | PA56A | 2.10E-06 | 7.30E-01 | -5.54 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB72A | 1.08E+01 | 9.00E-01 | 1.08 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB74A | 2.70E-01 | 9.00E-01 | -0.52 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB75A | 3.76E+01 | 9.00E-01 | 1.62 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 2WPT_A_B | SB77A | 3.30E-01 | 9.00E-01 | -0.44 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB78A | 6.50E-01 | 9.00E-01 | -0.14 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB84A | 8.00E-01 | 9.00E-01 | -0.05 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | FB86A | 5.49E+01 | 9.00E-01 | 1.79 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | TB87A | 1.03E+00 | 9.00E-01 | 0.06 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | QB92A | 4.10E-01 | 9.00E-01 | -0.34 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | KB97A | 6.80E-01 | 9.00E-01 | -0.12 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | VB98A | 1.05E+00 | 9.00E-01 | 0.07 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2VLP_A_B | AB54R | 2.36E-06 | 3.54E-05 | -1.18 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 2VLN_A_B | AB75N | 2.36E-06 | 1.68E-04 | -1.85 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 2VLQ_A_B | AB86F | 2.36E-06 | 1.80E-03 | -2.88 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 2VLO_A_B | AB97K | 2.36E-06 | 3.05E-05 | -1.11 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | RB54A | 3.54E-05 | 2.36E-06 | 1.18 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NB72A | 1.28E-05 | 2.36E-06 | 0.73 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB74A | 1.75E-06 | 2.36E-06 | -0.13 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NB75A | 1.68E-04 | 2.36E-06 | 1.85 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB77A | 1.58E-06 | 2.36E-06 | -0.17 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB78A | 1.02E-06 | 2.36E-06 | -0.36 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB84A | 1.87E-06 | 2.36E-06 | -0.1 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A | 1.80E-03 | 2.36E-06 | 2.88 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1EMV_A_B | QB92A | 1.59E-06 | 2.36E-06 | -0.17 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | KB97A | 3.05E-05 | 2.36E-06 | 1.11 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VB98A | 1.45E-05 | 2.36E-06 | 0.79 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A,LA33A | 1.50E-01 | 2.36E-06 | 4.8 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A,VA34A | 4.97E-03 | 2.36E-06 | 3.32 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A,VA37A | 1.70E-03 | 2.36E-06 | 2.86 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A,YA54A | 3.18E+00 | 2.36E-06 | 6.13 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A,YA55A | 3.00E+00 | 2.36E-06 | 6.1 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | CA23A | 6.09E-06 | 1.83E-06 | 0.52 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NA24A | 1.98E-06 | 1.83E-06 | 0.03 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | DA26A | 2.95E-06 | 1.83E-06 | 0.21 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TA27A | 3.37E-06 | 1.83E-06 | 0.27 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA28A | 1.91E-06 | 1.83E-06 | 0.02 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA29A | 7.13E-06 | 1.83E-06 | 0.59 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA30A | 1.48E-05 | 1.83E-06 | 0.91 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA31A | 2.19E-06 | 1.83E-06 | 0.08 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA32A | 2.44E-06 | 1.83E-06 | 0.13 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | LA33A | 4.00E-04 | 1.83E-06 | 2.34 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VA34A | 9.54E-05 | 1.83E-06 | 1.72 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | KA35A | 2.20E-06 | 1.83E-06 | 0.08 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1EMV_A_B | LA36A | 4.75E-06 | 1.83E-06 | 0.41 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VA37A | 2.09E-05 | 1.83E-06 | 1.06 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TA38A | 5.13E-06 | 1.83E-06 | 0.45 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA42A | 3.25E-06 | 1.83E-06 | 0.25 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TA44A | 2.70E-06 | 1.83E-06 | 0.17 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA45A | 2.04E-06 | 1.83E-06 | 0.05 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | HA46A | 3.98E-06 | 1.83E-06 | 0.34 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | PA47A | 2.26E-06 | 1.83E-06 | 0.09 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA48A | 1.77E-06 | 1.83E-06 | -0.01 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | GA49A | 1.22E-05 | 1.83E-06 | 0.82 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA50A | 5.14E-05 | 1.83E-06 | 1.45 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | DA51A | 6.11E-03 | 1.83E-06 | 3.52 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | LA52A | 2.87E-06 | 1.83E-06 | 0.2 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | IA53A | 6.06E-06 | 1.83E-06 | 0.52 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | YA54A | 2.55E-03 | 1.83E-06 | 3.14 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | YA55A | 2.75E-03 | 1.83E-06 | 3.18 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | PA56A | 1.30E-05 | 1.83E-06 | 0.85 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | DA60A | 2.69E-06 | 1.83E-06 | 0.17 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA63A | 4.54E-06 | 1.83E-06 | 0.4 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VA68A | 1.59E-05 | 1.83E-06 | 0.94 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1EMV_A_B | NA69A | 2.46E-06 | 1.83E-06 | 0.13 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1FR2_A_B | AA41E | 1.83E-06 | 1.42E-05 | -0.89 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 2GYK_A_B | AA51D | 1.83E-06 | 6.11E-03 | -3.52 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 2AJF_A_E | KE344R | 1.04E-03 | 1.16E-03 | -0.05 | 15791205 | Human Angiotensin-converting enzyme 2 | SARS spike protein receptor binding domain |
| 2AJF_A_E | FE360S | 8.80E-04 | 1.16E-03 | -0.12 | 15791205 | Human Angiotensin-converting enzyme 2 | SARS spike protein receptor binding domain |
| 2AJF_A_E | NE479K | 2.77E-02 | 1.16E-03 | 1.38 | 15791205 | Human Angiotensin-converting enzyme 2 | SARS spike protein receptor binding domain |
| 2AJF_A_E | TE487S | 1.32E-02 | 1.16E-03 | 1.06 | 15791205 | Human Angiotensin-converting enzyme 2 | SARS spike protein receptor binding domain |
| 1MQ8_A_B | CB161L,CB299F | 4.60E+00 | 4.30E-01 | 1.03 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB161L,CB299F,KB287C,KB294C | 1.40E-02 | 4.30E-01 | -1.49 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB161L,CB299F,EB284C,EB301C | 4.50E-02 | 4.30E-01 | -0.98 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB161L,CB299F,KB160C,TB300C | 1.20E+00 | 4.30E-01 | 0.45 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB161L,CB299F,LB289C,KB294C | 3.60E+00 | 4.30E-01 | 0.92 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB161L,KB160C | 7.70E-01 | 4.30E-01 | 0.25 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MQ8_A_B | CB299F,TB300C | 7.60E-01 | 4.30E-01 | 0.25 | 12526797 | Intercellular adhesion molecule I | Integrin alpha-L |
| 1MAH_A_F | FA295L | 3.70E-03 | 4.40E-03 | -0.08 | 8157652 | Acetylcholinesterase | Fasciculin |
| 1MAH_A_F | FA297I | 6.00E-03 | 4.40E-03 | 0.14 | 8157652 | Acetylcholinesterase | Fasciculin |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1MAH_A_F | FA297Y | 3.10E-03 | 4.40E-03 | -0.15 | 8157652 | Acetylcholinesterase | Fasciculin |
| 1MAH_A_F | YA337A | 6.80E-03 | 4.40E-03 | 0.19 | 8157652 | Acetylcholinesterase | Fasciculin |
| 1MAH_A_F | DA74N | 4.00E-02 | 4.40E-03 | 0.96 | 8157652 | Acetylcholinesterase | Fasciculin |
| 1MAH_A_F | YA124Q | 2.90E-01 | 4.40E-03 | 1.82 | 8157652 | Acetylcholinesterase | Fasciculin |
| 1LFD_A_B | RA20A | 3.30E+01 | 1.49E+01 | 0.35 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | SA22K | 2.54E+01 | 1.49E+01 | 0.23 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | LA23K | 1.68E+01 | 1.49E+01 | 0.05 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | NA27K | 4.46E+01 | 1.49E+01 | 0.48 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | MA30K | 6.70E+00 | 1.49E+01 | -0.35 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | KA32A | 4.00E+01 | 1.49E+01 | 0.43 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | KA48A | 1.68E+01 | 1.49E+01 | 0.05 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA51K | 9.30E+00 | 1.49E+01 | -0.21 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | KA52A | 2.83E+01 | 1.49E+01 | 0.28 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | NA54K | 4.20E+00 | 1.49E+01 | -0.55 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | LA55K | 1.11E+01 | 1.49E+01 | -0.13 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA56A | 2.23E+01 | 1.49E+01 | 0.18 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | EA57A | 2.09E+01 | 1.49E+01 | 0.15 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | EA57K | 2.30E+01 | 1.49E+01 | 0.19 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA58K | 6.40E+00 | 1.49E+01 | -0.37 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | NA92K | 1.61E+01 | 1.49E+01 | 0.03 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | YA93K | 4.80E+01 | 1.49E+01 | 0.51 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA94K | 7.90E+00 | 1.49E+01 | -0.28 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | MA30K,DA58K | 5.80E+00 | 1.49E+01 | -0.41 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | MA30K,DA94K | 1.99E+01 | 1.49E+01 | 0.13 | 15197281 | RalGSD-RBD | H-Ras1 |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1LFD_A_B | MA30K,DA51K,DA58K | 7.80E+00 | 1.49E+01 | -0.28 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | MA30K,DA51K,DA58K,DA94K | 1.00E+01 | 1.49E+01 | -0.17 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA51K,DA56K,EA57K | 6.00E+00 | 1.49E+01 | -0.4 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA58K,DA94K | 1.99E+01 | 1.49E+01 | 0.13 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1KTZ_A_B | VA92I | 7.20E-02 | 5.40E-02 | 0.13 | 19161338 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | RA25K | 2.00E-01 | 5.40E-02 | 0.57 | 19161338 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1REW_AB_C | DA30A,DB30A | 1.20E-03 | 4.00E-04 | 0.48 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | WA31A,WB31A | 2.28E-03 | 4.00E-04 | 0.76 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | DA30A,WA31A,DB30A,WB31A | 1.20E-02 | 4.00E-04 | 1.48 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | FA49A,FB49A | 4.00E-04 | 4.00E-04 | 0 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | PA50A,PB50A | 3.20E-04 | 4.00E-04 | -0.1 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | HA39D,HB39D | 4.40E-04 | 4.00E-04 | 0.04 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | SA88A,SB88A | 4.40E-04 | 4.00E-04 | 0.04 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | LA100A,LB100A | 4.80E-04 | 4.00E-04 | 0.08 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | HA39D,SA88A,HB39D,SB88A | 4.80E-04 | 4.00E-04 | 0.08 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | HA39D,LA100A,HB39D,LB100A | 4.40E-04 | 4.00E-04 | 0.04 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1REW_AB_C | AA34D,AB34D | 2.24E-04 | 4.00E-04 | -0.25 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | DA53A,DB53A | 4.40E-04 | 4.00E-04 | 0.04 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | EA109R,EB109R | 4.80E-04 | 4.00E-04 | 0.08 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | DA30A,AA34D,DB30A,AB34D | 7.60E-04 | 4.00E-04 | 0.28 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | AA34D,DA53A,AB34D,DB53A | 4.40E-04 | 4.00E-04 | 0.04 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | DA53A,EA109R,DB53A,EB109R | 5.60E-04 | 4.00E-04 | 0.15 | 10880444 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | KC88R,SC90T,KC92I,AC93P,QC94H,LC95Q,TC98D | 9.70E-04 | 2.40E-04 | 0.61 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJ9_AB_C | RC88K,TC90S,IC92K,PC93A,HC94Q,QC95L,SC98T | 2.40E-04 | 9.70E-04 | -0.61 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | KC88R,SC90T,KC92I,AC93P,QC94H,LC95Q,TC98D,AC74T,MC78L,KC79G,YC80L | 3.40E-04 | 2.40E-04 | 0.15 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJ9_AB_C | AC74T,MC78L,KC79G,YC80L | 3.40E-04 | 9.70E-04 | -0.46 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJA_AB_C | RC88K,TC90S,IC92K,PC93A,HC94Q,QC95L,SC98T,TC74A,LC78M,GC79K,LC80Y | 2.40E-04 | 3.40E-04 | -0.15 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1REW_AB_C | KC88R,SC90T,KC92I,AC93P,QC94H,LC95Q,TC98D,AC74T,MC78L,KC79G,YC80L,GC42H,DC46E,AC61T,IC62M | 1.11E-03 | 2.40E-04 | 0.67 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJ9_AB_C | AC74T,MC78L,KC79G,YC80L,GC42H,DC46E,AC61T,IC62M | 1.11E-03 | 9.70E-04 | 0.06 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 2QJA_AB_C | GC42H,DC46E,AC61T,IC62M | 1.11E-03 | 3.40E-04 | 0.51 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJB_AB_C | HC42G,EC46D,TC61A,MC62I | 3.40E-04 | 1.11E-03 | -0.51 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJB_AB_C | TC74A,LC78M,GC79K,LC80Y, HC42G,EC46D,TC61A,MC62I | 9.70E-04 | 1.11E-03 | -0.06 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 2QJB_AB_C | RC88K,TC90S,IC92K,PC93A,H C94Q,QC95L,SC98T,TC74A,L C78M,GC79K,LC80Y,HC42G,E C46D,TC61A,MC62I | 2.40E-04 | 1.11E-03 | -0.67 | 18160401 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 3BK3_A_C | LC1A | 2.80E-02 | 2.60E-02 | 0.03 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | LC1R | 2.80E-02 | 2.60E-02 | 0.03 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | TC3P | 4.80E-02 | 2.60E-02 | 0.27 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | TC5P | 7.50E-02 | 2.60E-02 | 0.46 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC18A | 4.40E-02 | 2.60E-02 | 0.23 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC18R | 8.90E-02 | 2.60E-02 | 0.53 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 1JTG_A_B | DB49A | 7.89E-03 | 2.80E-04 | 1.45 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,DB49A | 6.22E-03 | 2.80E-04 | 1.35 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,SA130A,DB49A | 2.47E-03 | 2.80E-04 | 0.95 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,DB49A | 2.05E-03 | 2.80E-04 | 0.87 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,DB49A | 5.27E-03 | 2.80E-04 | 1.27 | 10772866 | TEM-1 beta- | BLIP |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| | | | | | | lactamase | |
| 1JTG_A_B | RA243A,SA235A,DB49A | 6.40E-04 | 2.80E-04 | 0.36 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA130A,DB49A | 2.27E-03 | 2.80E-04 | 0.91 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,KA234A,DB49A | 3.19E-03 | 2.80E-04 | 1.06 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA130A,DB49A | 1.58E-03 | 2.80E-04 | 0.75 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA235A,SA130A,DB49A | 1.05E-03 | 2.80E-04 | 0.57 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,SA130A,KA234A,DB49A | 1.09E-03 | 2.80E-04 | 0.59 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,KA234A,DB49A | 5.10E-03 | 2.80E-04 | 1.26 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA235A,KA234A,DB49A | 4.86E-03 | 2.80E-04 | 1.24 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA130A,RA243A,DB49A | 3.13E-03 | 2.80E-04 | 1.05 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA235A,SA130A,RA243A,DB49A | 3.90E-03 | 2.80E-04 | 1.14 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A | 2.60E-03 | 2.80E-04 | 0.97 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA130A | 6.50E-04 | 2.80E-04 | 0.37 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,SA130A | 1.10E-03 | 2.80E-04 | 0.59 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A | 7.20E-04 | 2.80E-04 | 0.41 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A | 1.12E-03 | 2.80E-04 | 0.6 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA235A | 3.10E-04 | 2.80E-04 | 0.04 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA130A | 1.55E-03 | 2.80E-04 | 0.74 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A,KA234A | 1.87E-03 | 2.80E-04 | 0.83 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA130A | 1.16E-03 | 2.80E-04 | 0.62 | 10772866 | TEM-1 beta-lactamase | BLIP |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1JTG_A_B | RA243A,SA235A,SA130A | 1.06E-03 | 2.80E-04 | 0.58 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,KA234A | 2.74E-03 | 2.80E-04 | 0.99 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A,SA235A,KA234A | 2.68E-03 | 2.80E-04 | 0.98 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA130A,RA243A | 1.89E-03 | 2.80E-04 | 0.83 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KA234A,SA235A,SA130A,RA243A | 3.78E-03 | 2.80E-04 | 1.13 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A | 3.19E-04 | 1.50E-04 | 0.33 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A | 1.16E-04 | 1.50E-04 | -0.11 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA103A | 4.03E-03 | 1.50E-04 | 1.43 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA104A | 8.67E-03 | 1.50E-04 | 1.76 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | PA107A | 2.01E-04 | 1.50E-04 | 0.13 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A | 3.35E-04 | 1.50E-04 | 0.35 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A | 1.78E-04 | 1.50E-04 | 0.07 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA216A | 2.96E-05 | 1.50E-04 | -0.71 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB36A | 1.52E-02 | 1.50E-04 | 2.01 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | HB41A | 3.70E-02 | 1.50E-04 | 2.39 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB49A | 2.90E-03 | 1.50E-04 | 1.29 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YB50A | 3.92E-05 | 1.50E-04 | -0.58 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YB53A | 8.27E-03 | 1.50E-04 | 1.74 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SB71A | 4.05E-04 | 1.50E-04 | 0.43 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB112A | 3.22E-02 | 1.50E-04 | 2.33 | 17070843 | TEM-1 beta-lactamase | BLIP |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1JTG_A_B | SB113A | 1.74E-04 | 1.50E-04 | 0.06 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB142A | 9.80E-03 | 1.50E-04 | 1.82 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | HB148A | 1.40E-02 | 1.50E-04 | 1.97 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB150A | 9.08E-02 | 1.50E-04 | 2.78 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB162A | 9.40E-03 | 1.50E-04 | 1.8 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,HB148A | 8.00E-02 | 1.50E-04 | 2.73 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,WB150A | 3.81E-02 | 1.50E-04 | 2.4 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,RB160A | 1.14E-02 | 1.50E-04 | 1.88 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,HB148A | 8.11E-03 | 1.50E-04 | 1.73 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,WB150A | 1.10E-01 | 1.50E-04 | 2.87 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,RB160A | 1.77E-03 | 1.50E-04 | 1.07 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,WB162A | 5.17E-03 | 1.50E-04 | 1.54 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA103A,WB162A | 9.00E-02 | 1.50E-04 | 2.78 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A,YB53A | 2.99E-02 | 1.50E-04 | 2.3 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A,YB50A | 2.27E-04 | 1.50E-04 | 0.18 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A,FB36A | 1.60E-02 | 1.50E-04 | 2.03 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | PA107A,HB41A | 1.03E-02 | 1.50E-04 | 1.84 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | PA107A,YB50A | 8.75E-05 | 1.50E-04 | -0.23 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | PA107A,YB53A | 4.34E-03 | 1.50E-04 | 1.46 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA216A,YB50A | 1.16E-05 | 1.50E-04 | -1.11 | 17070843 | TEM-1 beta-lactamase | BLIP |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1JTG_A_B | EA110A,SB113A,SB71A | 1.25E-03 | 1.50E-04 | 0.92 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SB71A,SB113A | 6.08E-04 | 1.50E-04 | 0.61 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA104A,SB113A | 5.29E-03 | 1.50E-04 | 1.55 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,WB112A | 5.19E-02 | 1.50E-04 | 2.54 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,WB162A | 2.87E-02 | 1.50E-04 | 2.28 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,KB74A | 9.33E-02 | 1.50E-04 | 2.79 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A,FB142A | 2.81E-02 | 1.50E-04 | 2.27 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,WB112A | 1.46E-02 | 1.50E-04 | 1.99 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,WB162A | 3.59E-03 | 1.50E-04 | 1.38 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,RB160A | 1.77E-03 | 1.50E-04 | 1.07 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,KB74A | 4.12E-02 | 1.50E-04 | 2.44 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,FB142A | 7.31E-03 | 1.50E-04 | 1.69 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A,DB49A | 1.57E-03 | 1.50E-04 | 1.02 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,WB112A | 1.65E-02 | 1.50E-04 | 2.04 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,WB150A | 7.70E-02 | 1.50E-04 | 2.71 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,RB160A | 2.00E-03 | 1.50E-04 | 1.12 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,KB74A | 6.83E-02 | 1.50E-04 | 2.66 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,FB142A | 1.10E-02 | 1.50E-04 | 1.87 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA168A,DB49A | 2.29E-03 | 1.50E-04 | 1.18 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A,SB113A,SB71A | 1.58E-03 | 1.50E-04 | 1.02 | 17070843 | TEM-1 beta-lactamase | BLIP |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1JTG_A_B | VA216A,FB142A | 2.17E-03 | 1.50E-04 | 1.16 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA104A,SB113A,SB71A | 2.58E-02 | 1.50E-04 | 2.24 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | TB32K | 6.00E-04 | 2.80E-04 | 0.33 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VB93K | 3.00E-04 | 2.80E-04 | 0.03 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | TB140K | 6.60E-04 | 2.80E-04 | 0.37 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NB89K | 3.60E-04 | 2.80E-04 | 0.11 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB163A | 2.40E-04 | 2.80E-04 | -0.07 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB163K | 2.20E-04 | 2.80E-04 | -0.11 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | TB140K,QB157K | 6.70E-04 | 2.80E-04 | 0.38 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VB165K,DB163K,NB89K | 2.70E-04 | 2.80E-04 | -0.02 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VB165K,DB163K,DB135K,NB89K | 2.30E-04 | 2.80E-04 | -0.09 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1GL1_A_I | KI31M,AI32G | 7.40E-05 | 1.62E-04 | -0.34 | 7592720 | Bovine alpha-chymotrypsin | PMP-C insect inhibitor |
| 1GL0_E_I | MI30K | 1.10E-04 | 2.10E-04 | -0.28 | 7592720 | Bovine alpha-chymotrypsin | PMP-D2v insect inhibitor |
| 1FC2_C_D | LC136D | 3.60E-03 | 3.20E-03 | 0.05 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | NC147A | 8.40E-03 | 3.20E-03 | 0.42 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | FC149A | 4.30E-03 | 3.20E-03 | 0.13 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | IC150A | 6.20E-03 | 3.20E-03 | 0.29 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | KC154A | 3.10E-02 | 3.20E-03 | 0.99 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 2FTL_E_I | GI12A | 1.90E-05 | 5.00E-08 | 2.58 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | KI15A | 4.20E-05 | 5.00E-08 | 2.92 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | II18A | 9.20E-05 | 5.00E-08 | 3.26 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | GI36A | 1.10E-04 | 5.00E-08 | 3.34 | 8784199 | Bovine trypsin | BPTI |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1CBW_FGH_I | TI11A | 2.30E-03 | 1.80E-03 | 0.11 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI12A | 2.00E-03 | 1.80E-03 | 0.05 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | PI13A | 2.30E-03 | 1.80E-03 | 0.11 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | KI15A | 2.00E-02 | 1.80E-03 | 1.05 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | RI17A | 4.90E-03 | 1.80E-03 | 0.44 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | II18A | 1.70E-02 | 1.80E-03 | 0.98 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | II19A | 2.50E-03 | 1.80E-03 | 0.14 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | RI20A | 3.90E-03 | 1.80E-03 | 0.34 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | FI33A | 2.40E-03 | 1.80E-03 | 0.13 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | VI34A | 2.80E-03 | 1.80E-03 | 0.19 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | YI35A | 8.90E-03 | 1.80E-03 | 0.69 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI36A | 5.60E-03 | 1.80E-03 | 0.49 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI37A | 5.00E-03 | 1.80E-03 | 0.44 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | KI46A | 1.20E-03 | 1.80E-03 | -0.18 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1A4Y_A_B | HB8A | 4.50E-07 | 1.40E-07 | 0.51 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | QB12A | 2.00E-07 | 1.40E-07 | 0.16 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | NB68A | 2.30E-07 | 1.40E-07 | 0.22 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EB108A | 1.60E-07 | 1.40E-07 | 0.06 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA434A | 7.30E-06 | 1.10E-07 | 1.82 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | DA435A | 1.40E-05 | 1.10E-07 | 2.1 | 9050852 | Ribonuclease inhibitor | Angiogenin |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A4Y_A_B | YA437A | 3.10E-07 | 1.10E-07 | 0.45 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1Z7X_W_X | YW437A | 7.40E-04 | 1.20E-05 | 1.79 | 9050852 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | QW430A,VW432A | 4.00E-05 | 1.20E-05 | 0.52 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW438A,SW439A,EW440A | 2.50E-04 | 1.20E-05 | 1.32 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | RW457A | 3.20E-05 | 1.20E-05 | 0.43 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | IW459A | 1.20E-05 | 1.20E-05 | 0 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1A4Y_A_B | QA430A,VA432A | 8.40E-08 | 1.10E-07 | -0.12 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA438A,SA439A,EA440A | 1.70E-06 | 1.10E-07 | 1.19 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RA457A | 5.80E-08 | 1.10E-07 | -0.28 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | IA459A | 2.00E-07 | 1.10E-07 | 0.26 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1Z7X_W_X | EW206A | 7.60E-05 | 1.20E-05 | 0.8 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW261A | 8.00E-05 | 1.20E-05 | 0.82 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW263A | 4.60E-04 | 1.20E-05 | 1.58 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW287A | 5.20E-05 | 1.20E-05 | 0.64 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | SW289A | 3.40E-05 | 1.20E-05 | 0.45 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | KW320A | 5.60E-05 | 1.20E-05 | 0.67 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW344A | 1.10E-04 | 1.20E-05 | 0.96 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW375A | 9.90E-05 | 1.20E-05 | 0.92 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW401A | 4.30E-05 | 1.20E-05 | 0.55 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1A4Y_A_B | HB84A | 1.40E-07 | 1.10E-07 | 0.11 | 10970748 | Ribonuclease inhibitor | Angiogenin |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A4Y_A_B | WB89A | 1.60E-07 | 1.10E-07 | 0.16 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA261A | 1.00E-07 | 1.10E-07 | -0.04 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA263A | 6.90E-07 | 1.10E-07 | 0.8 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EA287A | 8.80E-08 | 1.10E-07 | -0.1 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | SA289A | 6.30E-08 | 1.10E-07 | -0.24 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA318A | 1.30E-06 | 1.10E-07 | 1.07 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KA320A | 9.40E-08 | 1.10E-07 | -0.07 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EA344A | 9.80E-08 | 1.10E-07 | -0.05 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA375A | 3.10E-07 | 1.10E-07 | 0.45 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EA401A | 3.20E-07 | 1.10E-07 | 0.46 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A | 3.90E-06 | 1.30E-07 | 1.48 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB32A | 7.80E-07 | 1.30E-07 | 0.78 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB66A | 1.80E-07 | 1.30E-07 | 0.14 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB70A | 1.10E-07 | 1.30E-07 | -0.07 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB31A | 1.60E-07 | 1.50E-07 | 0.03 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB33A | 2.20E-07 | 1.50E-07 | 0.17 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40Q | 5.70E-05 | 1.30E-07 | 2.64 | 2742853 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | HB13A | 8.10E-08 | 1.50E-07 | -0.27 | 2479414 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA434F | 1.60E-07 | 1.10E-07 | 0.16 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA437F | 9.00E-08 | 1.10E-07 | -0.09 | 10413501 | Ribonuclease inhibitor | Angiogenin |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A4Y_A_B | YA434A,DA435A | 2.70E-03 | 1.10E-07 | 4.39 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA434A,YA437A | 1.70E-03 | 1.10E-07 | 4.19 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A,YA434A | 2.80E-03 | 1.10E-07 | 4.41 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A,DA435A | 2.70E-03 | 1.10E-07 | 4.39 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A,YA434A,DA435A | 6.20E-01 | 1.10E-07 | 6.75 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A,YA434A,YA437A | 2.80E-02 | 1.10E-07 | 5.41 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G | 1.20E-05 | 1.10E-07 | 2.04 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G,YA434F | 9.30E-04 | 1.10E-07 | 3.93 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G,DA435A | 1.90E-05 | 1.10E-07 | 2.24 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G,YA437A | 5.60E-04 | 1.10E-07 | 3.71 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G,YA434A,DA435A | 1.90E-03 | 1.10E-07 | 4.24 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KB40G,YA434A,YA437A | 1.90E-02 | 1.10E-07 | 5.24 | 10413501 | Ribonuclease inhibitor | Angiogenin |
| 1Z7X_W_X | YW434F | 1.00E-05 | 1.20E-05 | -0.08 | 10413501 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | YW437F | 2.90E-04 | 1.20E-05 | 1.38 | 10413501 | Ribonuclease inhibitor | RNase A |
| 1A22_A_B | MA14A | 3.24E-04 | 2.70E-04 | 0.08 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | HA18A | 1.11E-04 | 2.70E-04 | -0.39 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | HA21A | 3.51E-04 | 2.70E-04 | 0.11 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA22A | 1.67E-04 | 2.70E-04 | -0.21 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA25A | 1.27E-04 | 2.70E-04 | -0.33 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DA26A | 2.13E-04 | 2.70E-04 | -0.1 | 7504735 | Human growth hormone | hGH binding protein |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A22_A_B | QA29A | 1.03E-04 | 2.70E-04 | -0.42 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | LA45A | 1.16E-03 | 2.70E-04 | 0.63 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA46A | 2.43E-04 | 2.70E-04 | -0.05 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PA48A | 3.24E-04 | 2.70E-04 | 0.08 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SA51A | 3.24E-04 | 2.70E-04 | 0.08 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA56A | 5.67E-04 | 2.70E-04 | 0.32 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PA61A | 1.94E-03 | 2.70E-04 | 0.86 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SA62A | 4.32E-04 | 2.70E-04 | 0.2 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NA63A | 3.24E-04 | 2.70E-04 | 0.08 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA64A | 2.13E-03 | 2.70E-04 | 0.9 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA65A | 1.86E-04 | 2.70E-04 | -0.16 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA68A | 8.91E-04 | 2.70E-04 | 0.52 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | YA164A | 5.67E-04 | 2.70E-04 | 0.32 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA167A | 1.32E-04 | 2.70E-04 | -0.31 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KA168A | 1.73E-04 | 2.70E-04 | -0.19 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DA171A | 1.24E-03 | 2.70E-04 | 0.66 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KA172A | 5.40E-03 | 2.70E-04 | 1.3 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA174A | 8.91E-05 | 2.70E-04 | -0.48 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TA175A | 6.75E-03 | 2.70E-04 | 1.4 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA176A | 5.94E-03 | 2.70E-04 | 1.34 | 7504735 | Human growth hormone | hGH binding protein |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A22_A_B | RA178A | 6.48E-03 | 2.70E-04 | 1.38 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IA179A | 7.83E-04 | 2.70E-04 | 0.46 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA183A | 3.78E-04 | 2.70E-04 | 0.15 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA186A | 2.62E-04 | 2.70E-04 | -0.01 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB244A | 1.42E-03 | 5.10E-05 | 1.44 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB270A | 1.72E-04 | 5.10E-05 | 0.53 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB276A | 7.65E-05 | 5.10E-05 | 0.18 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB298A | 3.39E-05 | 5.10E-05 | -0.18 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB302A | 2.81E-05 | 5.10E-05 | -0.26 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB303A | 1.15E-03 | 5.10E-05 | 1.35 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB305A | 1.08E-03 | 5.10E-05 | 1.33 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PB306A | 5.14E-04 | 5.10E-05 | 1 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB320A | 4.42E-05 | 5.10E-05 | -0.06 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KB321A | 4.39E-05 | 5.10E-05 | -0.07 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB326A | 4.91E-04 | 5.10E-05 | 0.98 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB327A | 2.86E-04 | 5.10E-05 | 0.75 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB364A | 6.08E-04 | 5.10E-05 | 1.08 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB365A | 1.66E-03 | 5.10E-05 | 1.51 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB366A | 3.89E-05 | 5.10E-05 | -0.12 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | VB371A | 1.53E-05 | 5.10E-05 | -0.52 | 9571026 | Human growth hormone | hGH binding protein |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1A22_A_B | QB416A | 1.86E-04 | 5.10E-05 | 0.56 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB417A | 7.05E-05 | 5.10E-05 | 0.14 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NB418A | 8.65E-05 | 5.10E-05 | 0.23 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB419A | 4.05E-05 | 5.10E-05 | -0.1 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NB272A | 6.97E-05 | 5.10E-05 | 0.14 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB277A | 6.41E-05 | 5.10E-05 | 0.1 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KB415A | 2.02E-04 | 5.10E-05 | 0.6 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA25A | 2.80E-04 | 4.90E-04 | -0.24 | 8756685 | Human growth hormone | hGH binding protein |
| 1A22_A_B | YA42A,QA46A | 8.50E-04 | 4.90E-04 | 0.24 | 8756685 | Human growth hormone | hGH binding protein |
| 1JRH_LH_I | NI48A | 4.81E-03 | 8.75E-03 | -0.26 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI48Q | 9.62E-03 | 8.75E-03 | 0.04 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YI49F | 3.88E-02 | 8.75E-03 | 0.65 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | VI51A | 2.06E-01 | 8.75E-03 | 1.37 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI52A | 3.32E-03 | 8.75E-03 | -0.42 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI53A | 7.61E-03 | 8.75E-03 | -0.06 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | SI54A | 1.29E-02 | 8.75E-03 | 0.17 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | EI55A | 4.29E-03 | 8.75E-03 | -0.31 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WI56F | 3.50E-03 | 8.75E-03 | -0.4 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WI56Y | 7.52E-03 | 8.75E-03 | -0.07 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI79A | 4.64E-03 | 8.75E-03 | -0.28 | 9878445 | mAbs A6 | Interferon gamma receptor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1JRH_LH_I | WI82F | 6.37E-02 | 8.75E-03 | 0.86 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WI82Y | 5.01E-02 | 8.75E-03 | 0.76 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | RI84A | 5.34E-03 | 8.75E-03 | -0.21 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI98A | 1.08E-02 | 8.75E-03 | 0.09 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1DAN_HL_UT | RH134A | 1.80E-03 | 5.70E-04 | 0.5 | 8962059 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | MH164A | 1.50E-03 | 5.70E-04 | 0.42 | 8962059 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KH192A | 3.90E-04 | 5.70E-04 | -0.17 | 8962059 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LH144A | 5.70E-04 | 5.70E-04 | 0 | 8962059 | Factor VIIa | Tissue factor |
| 1NMB_N_LH | DH56N | 2.90E-03 | 5.20E-03 | -0.25 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 1NMB_N_LH | YH99A | 5.88E-02 | 5.20E-03 | 1.05 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 1NMB_N_LH | YH100aF | 5.60E-03 | 5.20E-03 | 0.03 | 9579662 | Subtype N9 | Antibody NC10 |
| 1NMB_N_LH | TL93F | 4.30E-03 | 5.20E-03 | -0.08 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 1NMB_N_LH | TL93W | 6.30E-03 | 5.20E-03 | 0.08 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 1NMB_N_LH | LL94V | 1.11E-02 | 5.20E-03 | 0.33 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 3HFM_HL_Y | NL31D | 3.68E-04 | 5.40E-05 | 0.83 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YL50F | 2.06E-03 | 5.40E-05 | 1.58 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DH32A | 1.19E-03 | 5.40E-05 | 1.34 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DH32N | 3.30E-05 | 5.40E-05 | -0.21 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DH32A,KY97M | 5.80E-04 | 5.40E-05 | 1.03 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DH32N,KY97M | 1.12E-04 | 5.40E-05 | 0.32 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97M | 2.80E-04 | 5.40E-05 | 0.72 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | 2.71E-04 | 5.40E-05 | 0.7 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | SH31A | 4.80E-05 | 5.40E-05 | -0.05 | 10338006 | HyHEL-10 | HEW Lysozyme |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 3HFM_HL_Y | QL53A | 2.00E-04 | 5.40E-05 | 0.57 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YL96A | 2.30E-03 | 5.40E-05 | 1.63 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YL96F | 4.50E-04 | 5.40E-05 | 0.92 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21G | 5.80E-03 | 1.12E-04 | 1.71 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | 4.40E-04 | 1.12E-04 | 0.59 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21N | 4.50E-03 | 1.12E-04 | 1.6 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21E | 1.80E-03 | 1.12E-04 | 1.21 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21Q | 5.80E-03 | 1.12E-04 | 1.71 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21H | 4.40E-03 | 1.12E-04 | 1.59 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21M | 1.43E-03 | 1.12E-04 | 1.11 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21K | 1.50E-03 | 1.12E-04 | 1.13 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21W | 2.00E-03 | 1.12E-04 | 1.25 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97E | 1.50E-02 | 1.12E-04 | 2.13 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97M | 1.70E-04 | 1.12E-04 | 0.18 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97R | 8.20E-03 | 1.12E-04 | 1.86 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101G | 1.10E-04 | 1.12E-04 | -0.01 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101S | 1.00E-03 | 1.12E-04 | 0.95 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101N | 6.00E-04 | 1.12E-04 | 0.73 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101E | 3.60E-03 | 1.12E-04 | 1.51 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101Q | 1.80E-03 | 1.12E-04 | 1.21 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101K | 1.10E-03 | 1.12E-04 | 0.99 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101F | 9.00E-04 | 1.12E-04 | 0.91 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101R | 1.30E-03 | 1.12E-04 | 1.06 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YY20F | 6.10E-05 | 1.12E-04 | -0.26 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | GY102V | 1.00E-04 | 1.12E-04 | -0.05 | 9761467 | HyHEL-10 | HEW Lysozyme |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 3HFM_HL_Y | HY15A | 6.30E-05 | 1.00E-04 | -0.2 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | 4.40E-04 | 1.00E-04 | 0.64 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | WY63A | 1.77E-04 | 1.00E-04 | 0.25 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY73A | 6.20E-05 | 1.00E-04 | -0.21 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | LY75A | 1.64E-04 | 1.00E-04 | 0.22 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | TY89A | 1.13E-04 | 1.00E-04 | 0.05 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | NY93A | 1.50E-04 | 1.00E-04 | 0.18 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | IY98A | 9.10E-05 | 1.00E-04 | -0.04 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | SY100A | 1.60E-04 | 1.00E-04 | 0.2 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101A | 2.90E-04 | 1.00E-04 | 0.46 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 1DAN_HL_UT | EU208A | 1.82E-03 | 2.10E-03 | -0.06 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | VU207A | 1.76E-03 | 2.10E-03 | -0.08 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU203A | 1.98E-03 | 2.10E-03 | -0.03 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KU201A,DU204A | 2.18E-03 | 2.10E-03 | 0.02 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NU199A,RU200A | 2.86E-03 | 2.10E-03 | 0.13 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU197A,VU198A | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | SU195A | 2.11E-03 | 2.10E-03 | 0 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | SU195A,RU196A | 3.37E-03 | 2.10E-03 | 0.21 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YU185A | 1.78E-03 | 2.10E-03 | -0.07 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KU181A | 2.42E-03 | 2.10E-03 | 0.06 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LU176A | 2.01E-03 | 2.10E-03 | -0.02 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NU173A,EU174A | 2.40E-03 | 2.10E-03 | 0.06 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU172A | 1.90E-03 | 2.10E-03 | -0.04 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KU169A | 2.00E-03 | 2.10E-03 | -0.02 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU167A | 2.71E-03 | 2.10E-03 | 0.11 | 7654692 | Factor VIIa | Tissue factor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1DAN_HL_UT | KU165A,KU166A | 1.69E-03 | 2.10E-03 | -0.09 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | GU164R | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | SU163A | 1.93E-03 | 2.10E-03 | -0.04 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | WU158F | 1.70E-03 | 2.10E-03 | -0.09 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YU156L | 1.95E-03 | 2.10E-03 | -0.03 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | IU152A | 2.58E-03 | 2.10E-03 | 0.09 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KU149A,DU150A | 1.63E-03 | 2.10E-03 | -0.11 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DU145A | 1.37E-03 | 2.10E-03 | -0.19 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU144A | 2.14E-03 | 2.10E-03 | 0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FU140A | 1.86E-02 | 2.10E-03 | 0.95 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU139A | 2.09E-03 | 2.10E-03 | 0 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU136A,NU137A,NU138A | 1.73E-03 | 2.10E-03 | -0.08 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU135A | 3.11E-03 | 2.10E-03 | 0.17 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LU133A | 1.68E-03 | 2.10E-03 | -0.1 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU130A,RU131F | 2.08E-03 | 2.10E-03 | 0 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DU129A | 2.18E-03 | 2.10E-03 | 0.02 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU128A | 2.15E-03 | 2.10E-03 | 0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KU122A | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QU114A,EU117A | 1.74E-03 | 2.10E-03 | -0.08 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NU107A,QU110A | 4.55E-03 | 2.10E-03 | 0.34 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LU104A,EU105A | 1.88E-03 | 2.10E-03 | -0.05 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU99A | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT68A | 2.42E-03 | 2.10E-03 | 0.06 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT68A,QT69A | 1.88E-03 | 2.10E-03 | -0.05 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT65A,DT66A | 6.85E-04 | 2.10E-03 | -0.49 | 7654692 | Factor VIIa | Tissue factor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1DAN_HL_UT | DT61A,ET62A | 2.05E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT58E | 3.08E-02 | 2.10E-03 | 1.17 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT58A | 4.97E-02 | 2.10E-03 | 1.37 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT54A,ET56A | 3.95E-03 | 2.10E-03 | 0.27 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TT52A | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FT50A | 3.99E-03 | 2.10E-03 | 0.28 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT48A | 2.70E-03 | 2.10E-03 | 0.11 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST47A | 2.07E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT46A | 2.16E-03 | 2.10E-03 | 0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT46A,KT48A | 1.51E-02 | 2.10E-03 | 0.86 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | WT45F | 5.45E-02 | 2.10E-03 | 1.41 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT44A | 5.77E-03 | 2.10E-03 | 0.44 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST42A | 1.95E-03 | 2.10E-03 | -0.03 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT41A | 3.01E-03 | 2.10E-03 | 0.16 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT41A,DT44A | 5.25E-02 | 2.10E-03 | 1.4 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QT37A | 4.56E-03 | 2.10E-03 | 0.34 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT28A | 2.08E-03 | 2.10E-03 | 0 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ET26A | 2.09E-03 | 2.10E-03 | 0 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | WT25F | 8.93E-04 | 2.10E-03 | -0.37 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ET24A | 5.51E-03 | 2.10E-03 | 0.42 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | IT22A | 6.53E-03 | 2.10E-03 | 0.49 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TT21A | 2.06E-03 | 2.10E-03 | -0.01 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT20R | 4.23E-02 | 2.10E-03 | 1.3 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT20A | 1.30E-02 | 2.10E-03 | 0.79 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NT18A | 1.76E-03 | 2.10E-03 | -0.08 | 7654692 | Factor VIIa | Tissue factor |

| Complex PDB / Chains | Mutation(s) | $k_{off}$_mut (s^(-1)) | $k_{off}$_wt (s^(-1)) | $\Delta\log_{10}(k_{off})$ | Pubmed ID | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|---|
| 1DAN_HL_UT | KT15A | 1.25E-03 | 2.10E-03 | -0.23 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | WT14F | 1.24E-03 | 2.10E-03 | -0.23 | 7654692 | Factor VIIa | Tissue factor |
| 2VLR_ABC_DE | AE99S | 1.60E-01 | 2.90E-01 | -0.26 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | IE53V | 1.30E-01 | 1.60E-01 | -0.09 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | IE53L | 4.00E-01 | 1.60E-01 | 0.4 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | NE55A | 6.30E-01 | 1.60E-01 | 0.6 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | NE55D | 3.80E-01 | 1.60E-01 | 0.38 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | DE56A | 2.50E-01 | 1.60E-01 | 0.19 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | QE58A | 3.40E-01 | 1.60E-01 | 0.33 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | QE58E | 2.60E-01 | 1.60E-01 | 0.21 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SE99A | 2.90E-01 | 1.60E-01 | 0.26 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | YE101F | 2.10E-01 | 1.60E-01 | 0.12 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SD31A | 2.00E-01 | 1.60E-01 | 0.1 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SD32A | 4.60E-01 | 1.60E-01 | 0.46 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | QD34A | 1.03E+00 | 1.60E-01 | 0.81 | 18275829 | HL-A2-flu | JM22 |

## Table 10.4. SKEMPI Hotspot (ΔΔG) Dataset

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A22_A_B | HA18A | COR | -0.48639 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | HA21A | SUP | 0.155438 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA22A | RIM | -0.21984 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA25A | COR | -0.44731 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DA26A | RIM | -0.21131 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | YA42A | COR | 0.199344 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | LA45A | COR | 1.224517 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA46A | RIM | 0.108017 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PA48A | COR | 0.410656 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SA51A | SUP | 0.348235 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA56A | RIM | 0.410656 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PA61A | SUP | 1.209325 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SA62A | COR | 0.155438 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NA63A | COR | 0.314372 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA64A | COR | 1.642626 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA65A | RIM | -0.47308 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QA68A | RIM | 0.588454 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | YA164A | SUP | 0.348235 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA167A | SUP | 0.278455 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KA168A | COR | -0.15485 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DA171A | COR | 0.790924 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KA172A | SUP | 2.015046 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EA174A | COR | -0.92461 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TA175A | COR | 1.907029 | 7504735 | Human growth hormone | hGH binding protein |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A22_A_B | FA176A | SUP | 0.410656 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA178A | COR | 2.425703 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IA179A | SUP | 0.806313 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RA183A | RIM | 0.542858 | 7504735 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB243A | SUP | 2.116247 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB244A | RIM | 1.692722 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB270A | SUP | 0.690199 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB271A | COR | 0.535847 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB273A | RIM | 0.110914 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB274A | RIM | 0 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB275A | RIM | -0.09424 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB276A | COR | 0.514301 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB280A | RIM | -0.01769 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB298A | RIM | -0.05473 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB302A | SUP | -0.18217 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB303A | SUP | 1.607865 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB305A | SUP | 1.941551 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | PB306A | SUP | 3.305722 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB320A | RIM | -0.18217 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KB321A | COR | 0.081285 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB324A | SUP | 0.274082 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB326A | SUP | 0.993925 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB327A | RIM | 0.970688 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB364A | SUP | 1.486534 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB365A | COR | 2.130757 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB366A | RIM | 0.017174 | 9571026 | Human growth hormone | hGH binding protein |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A22_A_B | KB367A | RIM | -0.01769 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | VB371A | COR | -0.61701 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB395A | RIM | -0.09424 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB416A | SUP | 0.891094 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB417A | COR | 0.274082 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NB418A | COR | 0.295628 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB419A | COR | 0.033864 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB301A | SUP | 1.761612 | 9571026 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB243A | SUP | 0.278455 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB244A | RIM | 0.650875 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB270A | SUP | 0.418016 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB271A | COR | 0.599325 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB275A | RIM | -0.07911 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB326A | SUP | 0.982421 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB327A | RIM | 0.410656 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB417A | COR | 0.331545 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB276A | COR | 0.56037 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB280A | RIM | 0.166723 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB273A | RIM | -0.44105 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB274A | RIM | -0.58109 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB298A | RIM | -0.32785 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB299A | RIM | -0.50694 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB301A | SUP | 1.085717 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB302A | SUP | 0.467123 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB303A | SUP | 0.432467 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB305A | SUP | 0.132202 | 2034689 | Human growth hormone | hGH binding protein |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A22_A_B | PB306A | SUP | 2.626807 | 2034689 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA25A | COR | -0.43145 | 8756685 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB243A | SUP | 2.200368 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB244A | RIM | 1.800119 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB273A | RIM | 0.099964 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB274A | RIM | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB275A | RIM | -0.10032 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB276A | COR | 0.599716 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | WB280A | RIM | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB298A | RIM | -0.10032 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB302A | SUP | -0.20008 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB303A | SUP | 1.800119 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB305A | SUP | 2.000724 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | CB308A | SUP | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB320A | RIM | -0.20008 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KB321A | COR | 0.099964 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | CB322A | COR | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | SB324A | SUP | 0.200134 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB326A | SUP | 1.000226 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | EB327A | RIM | 1.000226 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | DB364A | SUP | 1.598413 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | IB365A | COR | 2.200368 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | QB366A | RIM | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | KB367A | RIM | 0 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | VB371A | COR | -0.70092 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | TB395A | RIM | -0.10032 | 7529940 | Human growth hormone | hGH binding protein |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A22_A_B | QB416A | SUP | 0.900116 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | RB417A | COR | 0.200134 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | NB418A | COR | 0.300791 | 7529940 | Human growth hormone | hGH binding protein |
| 1A22_A_B | CA182A | COR | 1.010373 | 2471267 | Human growth hormone | hGH binding protein |
| 1A22_A_B | FA191A | RIM | 0.191291 | 2471267 | Human growth hormone | hGH binding protein |
| 1A4Y_A_B | HB8A | RIM | 0.904115 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | QB12A | SUP | 0.300265 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | NB68A | RIM | 0.11781 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EB108A | COR | -0.32272 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA434A | COR | 3.262015 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | DA435A | COR | 3.485544 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | YA437A | COR | 0.836312 | 9050852 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RA457A | RIM | -0.22399 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | IA459A | SUP | 0.679432 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | HB84A | COR | 0.170438 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WB89A | RIM | 0.240219 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA261A | COR | 0.100657 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA263A | COR | 1.171374 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | SA289A | SUP | 0.042336 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA318A | SUP | 1.500745 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | KA320A | COR | -0.31 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EA344A | COR | 0.17861 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | WA375A | SUP | 1.035362 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | EA401A | COR | 0.883734 | 10970748 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB5A | COR | 2.309282 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | RB32A | RIM | 0.910251 | 1281426 | Ribonuclease inhibitor | Angiogenin |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1A4Y_A_B | RB31A | COR | 0.250522 | 1281426 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | HB13A | SUP | -0.29669 | 2479414 | Ribonuclease inhibitor | Angiogenin |
| 1A4Y_A_B | HB114A | COR | 0.656829 | 2479414 | Ribonuclease inhibitor | Angiogenin |
| 1AHW_AB_C | YC157A | SUP | -1.88986 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | TC167A | COR | -0.07415 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | TC170A | SUP | 1.106266 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | LC176A | RIM | 0.987378 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | DC178A | RIM | -0.48481 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | TC197A | RIM | 1.346485 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | VC198A | RIM | -0.31437 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AHW_AB_C | NC199A | RIM | 1.078705 | 9480775 | Immunoglobulin FAB 5G9 | Tissue factor |
| 1AK4_A_D | PD485A | RIM | 2.449888 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | VD486A | COR | 2.355922 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | HD487A | RIM | 2.374152 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | GD489A | COR | 3.441638 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | PD490A | COR | 3.537182 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | ID491A | RIM | 1.60439 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1AK4_A_D | PD493A | RIM | 2.047078 | 9223641 | Cyclophilin A | HIV-1 capsid protein |
| 1BRS_A_D | KA27A | COR | 5.380949 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA59A | COR | 5.245372 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | RA87A | SUP | 5.564701 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | HA102A | COR | 6.145795 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | YD29A | RIM | 3.470544 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | DD35A | COR | 4.50317 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | DD39A | COR | 7.650989 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | TD42A | COR | 1.85763 | 7739054 | Barnase | Barstar |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1BRS_A_D | ED76A | RIM | 1.364171 | 7739054 | Barnase | Barstar |
| 1BRS_A_D | EA73A | SUP | 2.347719 | 9126847 | Barnase | Barstar |
| 1BRS_A_D | RA59A | COR | 4.635372 | 8507637 | Barnase | Barstar |
| 1BRS_A_D | HA102A | COR | 6.912182 | 8507637 | Barnase | Barstar |
| 1BRS_A_D | DD35A | COR | 4.069636 | -3, 2004 | Barnase | Barstar |
| 1BRS_A_D | DD39A | COR | 5.935123 | | Barnase | Barstar |
| 1BRS_A_D | ED76A | RIM | 0.823553 | | Barnase | Barstar |
| 1BRS_A_D | KA27A | COR | 5.409263 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | NA58A | SUP | 3.066363 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA59A | COR | 5.183674 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA60A | RIM | -0.32588 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA73A | SUP | 1.897844 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA87A | SUP | 5.952121 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | HA102A | COR | 6.254761 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | KA27A | COR | 4.58795 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | NA58A | SUP | 3.115762 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | RA59A | COR | 4.89059 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA60A | RIM | 0.514863 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | EA73A | SUP | 2.813123 | 8494892 | Barnase | Barstar |
| 1BRS_A_D | HA102A | COR | 6.076371 | 8494892 | Barnase | Barstar |
| 1CBW_FGH_I | TI11A | COR | 0.221988 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI12A | SUP | 0.685735 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | PI13A | COR | -0.05647 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | KI15A | COR | 2.015046 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | RI17A | COR | 0.553533 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | II18A | COR | 1.415721 | 8784199 | Bovine alpha-chymotrypsin | BPTI |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1CBW_FGH_I | II19A | RIM | 0.142877 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | VI34A | RIM | 0.05155 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI36A | SUP | 0.96419 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | GI37A | COR | 0.821313 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | RI39A | RIM | 0.221988 | 8784199 | Bovine alpha-chymotrypsin | BPTI |
| 1CBW_FGH_I | KI15A | COR | 2.210915 | 10339415 | Bovine alpha-chymotrypsin | BPTI |
| 1CHO_EFG_I | LI18A | COR | 4.76622 | 9047374 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | TI17A | COR | 4.158528 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | EI19A | COR | 2.333248 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | YI20A | COR | 2.543463 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | RI21A | RIM | 3.191768 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | PI14A | RIM | 0.380623 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | KI13A | RIM | 0.181043 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | GI32A | SUP | -1.0918 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | NI36A | COR | -1.36417 | 11171964 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | GI32A | SUP | -0.77164 | Stephen Ming-teh Lu, PhD Thesis, Purdue University, 2000 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | TI17A | COR | 4.317521 | Stephen Ming-teh Lu, PhD Thesis, Purdue University, 2000 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |
| 1CHO_EFG_I | LI18A | COR | 4.932125 | Stephen Ming-teh Lu, PhD Thesis, Purdue University, 2000 | Bovine alpha-chymotrypsin | Turkey ovomucoid third domain |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1DAN_HL_UT | WT45A | COR | 1.500008 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST47A | COR | -0.12653 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YT51A | RIM | -0.12653 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FT76A | COR | 1.105437 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YT78A | RIM | 0.627639 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | PU92A | RIM | -0.18583 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YU94A | COR | 0.311685 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QT37A | SUP | 0.729036 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | GT43A | COR | 0.064695 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT48A | SUP | 0.920722 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TT60A | SUP | 2.224076 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT44A | RIM | 1.379911 | 7756258 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RH134A | COR | 0.749025 | 8962059 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | MH164A | COR | 0.744485 | 8962059 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT20A | COR | 2.438602 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT44A | RIM | 2.387218 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | WT45A | COR | 2.371096 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT46A | COR | 0.894808 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QU110A | COR | 1.305465 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU135A | RIM | 0.986135 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FU140A | COR | 2.215453 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | VU207A | COR | 1.569868 | 7947809 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST16A | COR | -0.12981 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT61A | COR | 0.242044 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ET62A | SUP | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | IT63A | SUP | 0 | 8312277 | Factor VIIa | Tissue factor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1DAN_HL_UT | VT64A | COR | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QT69A | RIM | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LT72A | COR | -0.05984 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU105A | RIM | -0.05984 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU106A | COR | -0.05984 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NU107A | RIM | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU131A | RIM | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU132A | SUP | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NU138A | RIM | 0 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | VU146A | COR | 0.19954 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FU147A | SUP | -0.05984 | 8312277 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU208A | RIM | -0.00483 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | VU207A | COR | -0.18867 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TU203A | RIM | 0.135272 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | SU163A | RIM | 0.022644 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FU140A | COR | 1.281753 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | RU135A | RIM | 0.519314 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | LU133A | COR | -0.02756 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU128A | RIM | 0.085805 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | EU99A | RIM | -0.17559 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | YU94A | COR | 1.02415 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT68A | RIM | -0.07038 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT58A | COR | 1.989497 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | FT50A | COR | 0.437911 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT48A | SUP | 0.414491 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST47A | COR | 0.043562 | 7654692 | Factor VIIa | Tissue factor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1DAN_HL_UT | KT46A | COR | 0.231824 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | DT44A | RIM | 0.735014 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ST42A | RIM | -0.0693 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT41A | RIM | 0.321569 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | QT37A | SUP | 0.546693 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | ET24A | RIM | 0.657886 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | IT22A | SUP | 0.645076 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TT21A | SUP | -0.15901 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT20A | COR | 2.587908 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | NT18A | COR | 0.180215 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | TT17A | COR | 0.120547 | 7654692 | Factor VIIa | Tissue factor |
| 1DAN_HL_UT | KT15A | RIM | -0.39734 | 7654692 | Factor VIIa | Tissue factor |
| 1DQJ_AB_C | YC20A | COR | 3.28718 | 10828942 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | RC21A | COR | 1.169074 | 10828942 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | WC63A | SUP | 1.346872 | 10828942 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | KC97A | COR | 3.521282 | 10828942 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | DC101A | COR | 1.453057 | 10828942 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YC20A | COR | 3.28718 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | RC21A | COR | 1.259554 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | WC62A | RIM | 0.758418 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | WC63A | SUP | 1.346872 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | LC75A | COR | 1.453057 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | TC89A | COR | 0.84122 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | NC93A | COR | 0.650164 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | KC96A | SUP | 6.158576 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | KC97A | COR | 3.521282 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1DQJ_AB_C | SC100A | COR | 0.77593 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | DC101A | COR | 1.301276 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | NA31A | COR | 2.014335 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | NA32A | SUP | 4.092513 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YA50A | COR | 2.679498 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | SA91A | SUP | 1.432529 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YA96A | COR | 1.135559 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | DB32A | COR | 2.014335 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YB33A | COR | 5.526939 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YB50A | SUP | 6.89111 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | YB53A | COR | 1.180807 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DQJ_AB_C | WB98A | COR | 4.933734 | 12515535 | HyHEL-63 Fab | HEW Lysozyme |
| 1DVF_AB_CD | HA30A | RIM | 1.650121 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YA32A | COR | 2.031279 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YA49A | COR | 1.629452 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YA50A | RIM | 0.688082 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | WA92A | COR | 0.340876 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | SA93A | COR | 1.162959 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | TB30A | RIM | 0.907919 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YB32A | COR | 1.832291 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | WB52A | COR | 4.134849 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DB54A | RIM | 4.283173 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | NB56A | COR | 1.162959 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DB58A | COR | 1.600721 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | EB98A | SUP | 4.188524 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | RB99A | COR | 1.875485 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1DVF_AB_CD | DB100A | RIM | 2.790763 | 8703938 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | EB98A | SUP | 4.188524 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DB54A | RIM | 4.283173 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DB58A | COR | 1.600721 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YA49A | COR | 1.729232 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YA32A | COR | 2.031279 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | NB56A | COR | 1.162959 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | WB52A | COR | 4.134849 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DB100A | RIM | 2.790763 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YD98A | COR | 4.741557 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | YC49A | COR | 1.861434 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | QD100A | COR | 1.629452 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | ND54A | RIM | 1.861434 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | RD100bA | COR | 4.092513 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | KD30A | RIM | 1.003967 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | HD33A | COR | 1.861434 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | DD52A | SUP | 1.683043 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1DVF_AB_CD | ID97A | COR | 2.682746 | 8993317 | IgG1-kappa D1.3 Fv | E5.2 Fv |
| 1EAW_A_B | QA38A | COR | -0.51891 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | IA41A | SUP | -0.82251 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | IA60A | RIM | -0.19436 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | DA60bA | RIM | 1.502775 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | RA60cA | RIM | 0.587615 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | FA60eA | SUP | -0.42881 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | YA60gA | SUP | -0.07894 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | FA94A | SUP | 0.728594 | 17475279 | Membrane-type serine protease 1 | BPTI |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1EAW_A_B | DA96A | RIM | 0.65444 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | FA97A | RIM | 0.89202 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | HA143A | COR | -0.01448 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | YA146A | RIM | 0.502154 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | TA150A | RIM | 0.089451 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | QA175A | RIM | -0.1331 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EAW_A_B | DA217A | RIM | 2.229134 | 17475279 | Membrane-type serine protease 1 | BPTI |
| 1EFN_A_B | IA96A | COR | 1.451027 | 7588629 | Fyn SH3 domain R96I mutant | HIV-1 Nef |
| 1EMV_A_B | RB54A | COR | 1.666447 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NB72A | COR | 1.165191 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB74A | COR | -0.24113 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NB75A | SUP | 2.335586 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB77A | COR | -0.23299 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB78A | SUP | -0.54014 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SB84A | SUP | -0.10947 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | FB86A | COR | 3.880681 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TB87A | SUP | 0.158506 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | QB92A | SUP | -0.27773 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | KB97A | COR | 1.960515 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VB98A | SUP | 1.08934 | 18471830 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | CA23A | COR | 0.92197 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | NA24A | RIM | 0.139471 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TA27A | COR | 0.72846 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA28A | RIM | 0.173174 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA29A | RIM | 0.956433 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA30A | RIM | 1.416635 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1EMV_A_B | LA33A | SUP | 3.419279 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VA34A | COR | 2.57945 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | VA37A | SUP | 1.664612 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | TA38A | RIM | 0.899966 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | EA41A | COR | 2.084014 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | HA46A | SUP | 0.832184 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | PA47A | RIM | 0.437469 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA48A | COR | 0.007269 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | GA49A | SUP | 1.486044 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | SA50A | COR | 2.18822 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | DA51A | COR | 5.918129 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | IA53A | SUP | 0.848125 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | YA54A | COR | 4.836801 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | YA55A | RIM | 4.636833 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1EMV_A_B | PA56A | RIM | 1.24284 | 9425068 | Colicin E9 immunity protein | Colicin E9 DNase |
| 1F47_A_B | DA4A | RIM | 0.691506 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | YA5A | RIM | 0.869367 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | LA6A | COR | 0.925431 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | DA7A | RIM | 1.733659 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | IA8A | COR | 2.516245 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | PA9A | RIM | -0.05756 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | FA11A | COR | 2.445483 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | LA12A | COR | 2.295326 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | KA14A | RIM | -0.04264 | 10880432 | FtsZ fragment | ZipA |
| 1F47_A_B | QA15A | RIM | -0.0456 | 10880432 | FtsZ fragment | ZipA |
| 1FC2_C_D | NC147A | RIM | 0.605702 | 8588944 | Protein A/Z | IgG1 MO61 Fc |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1FC2_C_D | IC150A | SUP | 5.276767 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | KC154A | RIM | 1.526621 | 8588944 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | NC147A | RIM | 0.582278 | 8332602 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | IC150A | SUP | 2.185484 | 8332602 | Protein A/Z | IgG1 MO61 Fc |
| 1FC2_C_D | KC154A | RIM | 1.231969 | 8332602 | Protein A/Z | IgG1 MO61 Fc |
| 1FCC_A_C | TC25A | COR | 0.240219 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | KC28A | COR | 1.256154 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | KC31A | COR | 3.477555 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | NC35A | COR | 2.365107 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | DC40A | RIM | 0.272251 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | EC42A | RIM | 0.385442 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FCC_A_C | WC43A | COR | 3.773184 | 10452608 | IgG1 MO61 Fc | B domain of Protein G |
| 1FFW_A_B | EB171A | SUP | 0.716771 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | EB178A | COR | 0.639143 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | HB181A | RIM | 0.033864 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | DB202A | RIM | -0.07415 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | DB207A | RIM | 0.096285 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | CB213A | RIM | 0.204301 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | FB214A | COR | 3.64594 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1FFW_A_B | IB216A | SUP | 0.42783 | 21642453 | Chemotaxis protein CheY | Chemotaxis protein CheA |
| 1GC1_G_C | SC23A | RIM | 0.292682 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | QC25A | COR | 0.032032 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | HC27A | COR | 0.282555 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | KC29A | SUP | 0.536239 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | NC32A | RIM | 0.182654 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | QC33A | RIM | 0.105268 | 2402498 | gp120 | CD4 |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1GC1_G_C | KC35A | COR | 0.322066 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | QC40A | COR | -0.41066 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | SC42A | COR | 0 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | LC44A | SUP | 1.05602 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | TC45A | COR | -0.14889 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | KC46A | COR | 1.431019 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | NC52A | COR | 0.708338 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | RC59A | COR | 1.175914 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | SC60A | RIM | -0.08859 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | DC63A | RIM | -0.31933 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | QC64A | RIM | 0.442689 | 2402498 | gp120 | CD4 |
| 1GC1_G_C | EC85A | SUP | 1.323296 | 2402498 | gp120 | CD4 |
| 1GCQ_AB_C | PC595A | COR | 0.767836 | 11406576 | Growth factor receptor-bound protein 2 | VavS |
| 1GCQ_AB_C | PC657A | COR | 1.316456 | 11406576 | Growth factor receptor-bound protein 2 | VavS |
| 1GCQ_AB_C | PC608A | COR | 0.120808 | 11406576 | Growth factor receptor-bound protein 2 | VavS |
| 1GCQ_AB_C | PC609A | SUP | 0.08525 | 11406576 | Growth factor receptor-bound protein 2 | VavS |
| 1H9D_A_B | RB3A | RIM | 1.162188 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1H9D_A_B | VB4A | SUP | 1.402407 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1H9D_A_B | GB61A | COR | 2.077467 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1H9D_A_B | QB67A | COR | 1.364171 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1H9D_A_B | LB103A | SUP | 0.940201 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1H9D_A_B | NB104A | COR | 2.304372 | 10984496 | AML1 Runx1 Runt domain | Core-binding factor beta |
| 1IAR_A_B | IA5A | COR | 1.171374 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | TA6A | SUP | -0.10364 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA8A | RIM | -0.02236 | 9050834 | Interleukin-4 | Interleukin-4 receptor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1IAR_A_B | TA13A | SUP | 0.978577 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | QA78A | COR | 0.124842 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA81A | RIM | 0.479624 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | FA82A | SUP | -0.08647 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | KA84A | COR | 0.344976 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA85A | COR | 0.426889 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | RA88A | COR | 3.754718 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | NA89A | SUP | 1.558794 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1IAR_A_B | WA91A | COR | 0.729529 | 9050834 | Interleukin-4 | Interleukin-4 receptor |
| 1JCK_A_B | TB20A | COR | 1.654842 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | YB26A | COR | 1.774828 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | NB60A | COR | 1.642626 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | YB90A | SUP | 2.59614 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | VB91A | COR | 2.232905 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | KB103A | SUP | 0.676638 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JCK_A_B | FB176A | RIM | 2.133934 | 9500785 | Beta-chain of 14.3.d | Staphylococcal enterotoxin C3 |
| 1JRH_LH_I | EL27A | RIM | 0.542858 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | DL28A | COR | 0.434841 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YL30A | COR | 1.108953 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YL91A | COR | 0.581094 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WL92A | COR | 2.819669 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | SL93A | COR | -0.65088 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | TL94A | COR | 0.385442 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WL96A | COR | 1.666811 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YH32A | RIM | 1.433952 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WH52A | SUP | 2.687467 | 11123892 | mAbs A6 | Interferon gamma receptor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1JRH_LH_I | WH53A | COR | 2.422402 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | DH54A | RIM | 1.886944 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | DH56A | RIM | 1.855479 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YH58A | COR | 1.256154 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | RH95A | SUP | 0.542858 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YH99A | RIM | 1.061531 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | HH100bA | COR | 1.698531 | 11123892 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI47A | SUP | 3.578757 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI48A | SUP | -0.29312 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YI49A | COR | 3.400763 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | GI50A | COR | 4.527355 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | VI51A | COR | 1.88353 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI52A | COR | 2.984793 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI53A | COR | 3.893992 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | SI54A | COR | 0.297682 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | EI55A | RIM | -0.43501 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WI82A | COR | 4.529326 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | RI84A | SUP | -0.24642 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI98A | RIM | -0.04264 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI47A | SUP | 3.852295 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI48A | SUP | 0.634185 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | YI49A | COR | 3.656642 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | GI50A | COR | 4.377107 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | VI51A | COR | 1.476288 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI52A | COR | 3.789874 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | NI53A | COR | 4.709525 | 9878445 | mAbs A6 | Interferon gamma receptor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1JRH_LH_I | SI54A | COR | 0.458078 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | EI55A | RIM | -0.69706 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | WI82A | COR | 4.332732 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | RI84A | SUP | 1.032008 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JRH_LH_I | KI98A | RIM | 0.667108 | 9878445 | mAbs A6 | Interferon gamma receptor |
| 1JTG_A_B | DB49A | COR | 2.561484 | 9891008 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB142A | COR | 3.379217 | 9891008 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB49A | COR | 1.814255 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA235A | SUP | 1.239083 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA130A | SUP | 0.791973 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A | SUP | 1.339371 | 10772866 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB49A | COR | 1.791966 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KB74A | SUP | 3.559509 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB142A | COR | 2.102992 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YB143A | COR | 0.38203 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA104A | COR | 1.552953 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YA105A | COR | -0.16837 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RA243A | SUP | 1.265764 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SA130A | SUP | 0.333848 | 15618400 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | QA99A | COR | 0.429602 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | NA100A | RIM | -0.45581 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA103A | SUP | 1.910744 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA104A | COR | 1.767957 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | PA107A | COR | -0.38276 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | EA110A | COR | 4.061928 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | MA129A | COR | 0.738821 | 17070843 | TEM-1 beta-lactamase | BLIP |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1JTG_A_B | EA168A | RIM | -0.07258 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | VA216A | COR | -0.40694 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB36A | COR | 3.20142 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | HB41A | SUP | 3.2497 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB49A | COR | 1.672743 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YB50A | COR | -0.40694 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | YB53A | SUP | 2.077467 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SB71A | SUP | 0.358089 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | KB74A | SUP | 3.823668 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB112A | COR | 3.010511 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | SB113A | COR | -0.16837 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | FB142A | COR | 2.508271 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | HB148A | SUP | 2.747889 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB150A | COR | 4.253562 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | RB160A | RIM | 2.222204 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | WB162A | COR | 2.340827 | 17070843 | TEM-1 beta-lactamase | BLIP |
| 1JTG_A_B | DB163A | RIM | -1.34093 | 10876236 | TEM-1 beta-lactamase | BLIP |
| 1KTZ_A_B | RA25A | RIM | 1.481581 | 19161338 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | RA94A | RIM | 2.884404 | 19161338 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | LB27A | COR | 2.271497 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | FB30A | COR | 3.426639 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | DB32A | RIM | 1.96819 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | SB49A | RIM | 0.773119 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | IB50A | COR | 2.343055 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | TB51A | COR | 1.96012 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | SB52A | SUP | 0.663091 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1KTZ_A_B | IB53A | COR | 1.816812 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | EB55A | RIM | 1.663096 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | VB77A | SUP | 0.86157 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | DB118A | RIM | 1.261316 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1KTZ_A_B | EB119A | COR | 1.940852 | 16300789 | Transforming growth factor beta 3 | TGF-beta type II receptor |
| 1LFD_A_B | RA20A | RIM | 1.136044 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | KA32A | COR | 1.326342 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA51A | RIM | -0.57906 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | KA52A | COR | 1.179308 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1LFD_A_B | DA56A | RIM | -0.27968 | 15197281 | RalGSD-RBD | H-Ras1 |
| 1NMB_N_LH | YH99A | RIM | 2.141579 | 9579662 | Subtype N9 neuraminidase | Antibody NC10 |
| 1PPF_E_I | LI18A | COR | 1.071087 | 9047374 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | TI17A | COR | 3.484821 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | EI19A | COR | 1.203289 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | YI20A | COR | 3.210383 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | RI21A | RIM | 0.208053 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | PI14A | RIM | -0.12413 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | KI13A | RIM | 0.756479 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | GI32A | COR | 0.235382 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | NI36A | RIM | -1.64552 | 11171964 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | GI32A | COR | 0.279808 | Stephen Ming-teh Lu, PhD Thesis, Purdue University, 2000 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1PPF_E_I | TI17A | COR | 2.894247 | Stephen Ming-teh Lu, PhD Thesis, Purdue | Human leukocyte elastase | Turkey ovomucoid third domain |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| | | | | University, 2000 Stephen Ming-teh Lu, PhD Thesis, Purdue University, | | |
| 1PPF_E_I | LI18A | COR | 0.962834 | 2000 | Human leukocyte elastase | Turkey ovomucoid third domain |
| 1R0R_E_I | LI18A | COR | 0.314609 | 9047374 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | TI17A | COR | 1.173231 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | EI19A | COR | 2.089436 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | YI20A | COR | 5.474549 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | RI21A | RIM | -0.09605 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | PI14A | RIM | -0.63891 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | KI13A | RIM | -0.61 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | GI32A | SUP | 1.298157 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | NI36A | RIM | -0.03315 | 11171964 | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | GI32A | SUP | 1.331018 | | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | TI17A | COR | 0.787094 | | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1R0R_E_I | LI18A | COR | 0.525685 | | Subtilisin Carlsberg | Turkey ovomucoid third domain |
| 1REW_AB_C | QC86A | COR | 2.658561 | 15064755 | Bone morphogenetic protein-2 | BMPR-IA receptor |
| 1S1Q_A_B | VA43A | COR | 0.670753 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1S1Q_A_B | FA44A | RIM | 0.199344 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1S1Q_A_B | NA45A | RIM | 1.231969 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1S1Q_A_B | DA46A | RIM | 0.965521 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1S1Q_A_B | WA75A | SUP | 0.280782 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1S1Q_A_B | FA88A | SUP | 0.77525 | 12006492 | Tumor susceptibility gene 101 protein | Ubiquitin |
| 1SMF_E_I | EI16A | RIM | 1.012809 | 20656696 | Bovine trypsin | Mung bean inhibitor peptide |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1SMF_E_I | SI12A | COR | 1.899972 | 20656696 | Bovine trypsin | Mung bean inhibitor peptide |
| 1SMF_E_I | TI10A | COR | 2.046052 | 20656696 | Bovine trypsin | Mung bean inhibitor peptide |
| 1SMF_E_I | II13A | COR | 3.511419 | 20656696 | Bovine trypsin | Mung bean inhibitor peptide |
| 1TM1_E_I | YI61A | COR | 2.17953 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI65A | SUP | 3.07986 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI67A | SUP | 2.923439 | 10065709 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | TI58A | COR | 2.572396 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | EI60A | COR | 2.924655 | 7947796 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | TI58A | COR | 2.728342 | 15865427 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | EI60A | COR | 3.054218 | 15865427 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI62A | RIM | 1.256154 | 15865427 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI65A | SUP | 3.75601 | 15865427 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | RI67A | SUP | 3.098124 | 15865427 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | MI59A | COR | 1.027668 | 15504027 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1TM1_E_I | YI61A | COR | 2.981582 | 15504027 | Subtilisin BPN | Chymotrypsin inhibitor 2 |
| 1UUZ_A_D | HA62A | RIM | 1.774828 | 17405861 | Inhibitor of vertebrate lysozyme | HEW Lysozyme |
| 1UUZ_A_D | CA64A | COR | 0.650875 | 17405861 | Inhibitor of vertebrate lysozyme | HEW Lysozyme |
| 1VFB_AB_C | HA30A | COR | 0.845261 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YA32A | COR | 1.339986 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YA49A | COR | 0.797899 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YA50A | COR | 0.388012 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | WA92A | COR | 2.728342 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | SA93A | RIM | 0.343278 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | TB30A | RIM | -0.05587 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YB32A | COR | 0.460352 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | WB52A | COR | 0.364468 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1VFB_AB_C | DB54A | RIM | 0.638906 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DB58A | RIM | -0.2071 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | EB98A | SUP | 1.157075 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | RB99A | RIM | -0.09978 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DB100A | COR | 3.07162 | 8703938 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DC18A | COR | 0.340283 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | NC19A | COR | 0.396264 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YC23A | SUP | 0.410656 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | SC24A | COR | 0.851346 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | KC116A | COR | 0.71377 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | TC118A | COR | 0.765438 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DC119A | RIM | 0.953515 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | VC120A | SUP | 0.91736 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | QC121A | COR | 2.878286 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | IC124A | SUP | 1.231969 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | RC125A | RIM | 1.837722 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | LC129A | RIM | 0.171622 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YA32A | COR | 1.717649 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YA50A | COR | 0.524568 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | WA92A | COR | 3.350074 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | YB32A | COR | 1.123715 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | WB52A | COR | 0.91736 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DB54A | RIM | 0.992047 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1VFB_AB_C | DB100A | COR | 2.941075 | 9609690 | IgG1-kappa D1.3 Fv | HEW Lysozyme |
| 1XD3_A_B | KB6A | RIM | 1.344086 | 10518943 | UCH-L3 | Ubiquitin |
| 1XD3_A_B | KB27A | SUP | -0.06242 | 10518943 | UCH-L3 | Ubiquitin |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 1XD3_A_B | DB39A | RIM | -0.41066 | 10518943 | UCH-L3 | Ubiquitin |
| 1XD3_A_B | LB8A | COR | 2.676792 | 10518943 | UCH-L3 | Ubiquitin |
| 1XD3_A_B | IB44A | SUP | 0.265981 | 10518943 | UCH-L3 | Ubiquitin |
| 1Z7X_W_X | YW434A | COR | 5.955931 | 9050852 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | DW435A | COR | 3.662431 | 9050852 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | YW437A | COR | 2.624135 | 9050852 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | RW457A | RIM | 0.848247 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | IW459A | SUP | 0.337421 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW206A | SUP | 1.018685 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW261A | COR | 1.335954 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW263A | SUP | 2.212418 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW287A | RIM | 1.321325 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | SW289A | SUP | 0.814384 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW318A | SUP | 0.99347 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | KW320A | COR | 1.321325 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW344A | SUP | 1.561543 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | WW375A | COR | 1.66956 | 10970748 | Ribonuclease inhibitor | RNase A |
| 1Z7X_W_X | EW401A | RIM | 1.306325 | 10970748 | Ribonuclease inhibitor | RNase A |
| 2B42_A_B | HA374A | COR | 1.638458 | 16279951 | TAXI-I | B. subtilis endoxylanase |
| 2BTF_A_P | FP59A | COR | 1.595741 | 9788869 | Bovine beta-actin | Bovine profilin |
| 2BTF_A_P | KP125A | COR | 0.460056 | 9788869 | Bovine beta-actin | Bovine profilin |
| 2FTL_E_I | GI12A | COR | 4.392778 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | KI15A | COR | 10.1592 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | II18A | COR | 5.021843 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | GI36A | SUP | 2.21439 | 8784199 | Bovine trypsin | BPTI |
| 2FTL_E_I | KI15A | COR | 10.63893 | 10339415 | Bovine trypsin | BPTI |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 2G2U_A_B | EB31A | RIM | 0.650875 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | SB35A | RIM | -0.9509 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | FB36A | COR | 2.763951 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | SB39A | SUP | -0.95614 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | HB41A | SUP | 1.717049 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | GB48A | COR | -0.4266 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | YB50A | COR | -2.07572 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | YB51A | COR | -0.62875 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | YB53A | SUP | 2.301744 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | SB71A | SUP | -0.51221 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | EB73A | COR | -1.97944 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | KB74A | SUP | -0.21734 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | WB112A | COR | 0.958735 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | SB113A | SUP | -0.61231 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | GB141A | SUP | -0.41381 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | FB142A | COR | 0.275827 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | YB143A | COR | -1.84724 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | RB144A | RIM | -0.34239 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | HB148A | SUP | 1.118951 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | WB150A | COR | 1.785222 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | RB160A | RIM | 0.669794 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2G2U_A_B | WB162A | COR | 0.53121 | 15284234 | SHV-1 beta-lactamase | BLIP |
| 2GOX_A_B | RB131A | COR | 2.24674 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 2GOX_A_B | NB138A | COR | 1.56953 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 2GOX_A_B | RB131A | COR | 1.519609 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |
| 2GOX_A_B | NB138A | COR | 1.333782 | 18687868 | Complement C3d | Fibrinogen-binding protein Efb-C |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 2HRK_A_B | KA159A | RIM | 0.953515 | 17976650 | GluRS | Arc1p |
| 2I9B_A_E | RE137A | RIM | -0.28764 | 10864923 | Urokinase-type plasminogen activator | Urokinase plasminogen activator receptor |
| 2I9B_A_E | KE139A | RIM | 0.674278 | 10864923 | Urokinase-type plasminogen activator | Urokinase plasminogen activator receptor |
| 2J0T_A_D | VD4A | RIM | 0 | 12515831 | MMP1 Interstitial collagenase | Metalloproteinase inhibitor 1 |
| 2J0T_A_D | SD68A | COR | 2.106373 | 12515831 | MMP1 Interstitial collagenase | Metalloproteinase inhibitor 1 |
| 2J0T_A_D | TD2A | COR | 4.289029 | 9268350 | MMP1 Interstitial collagenase | Metalloproteinase inhibitor 1 |
| 2J0T_A_D | MD66A | RIM | 1.642626 | 9268350 | MMP1 Interstitial collagenase | Metalloproteinase inhibitor 1 |
| 2J1K_C_T | RC384A | COR | 0.764846 | 16923808 | CAV-2 | CAR D1 domain |
| 2JEL_LH_P | TP62A | SUP | 0 | 1711212 | Jel42 antibody | Histadine-containing protein HPr |
| 2JEL_LH_P | EP68A | RIM | 0.410656 | 1711212 | Jel42 antibody | Histadine-containing protein HPr |
| 2JEL_LH_P | EP70A | SUP | 2.728342 | 1711212 | Jel42 antibody | Histadine-containing protein HPr |
| 2JEL_LH_P | HP76A | RIM | -0.41066 | 1711212 | Jel42 antibody | Histadine-containing protein HPr |
| 2O3B_A_B | EB24A | COR | 5.474915 | 17138564 | NucA nuclease | NuiA nuclease inhibitor |
| 2O3B_A_B | QB74A | RIM | 3.232964 | 17138564 | NucA nuclease | NuiA nuclease inhibitor |
| 2O3B_A_B | WB76A | COR | 4.073704 | 17138564 | NucA nuclease | NuiA nuclease inhibitor |
| 2PCC_A_B | DA34A | COR | -0.89705 | 11148036 | Cytochrome C peroxidase | Cytochrome C |
| 2PCC_A_B | VA197A | COR | 2.102682 | 11148036 | Cytochrome C peroxidase | Cytochrome C |
| 2PCC_A_B | EA290A | RIM | 6.202555 | 11148036 | Cytochrome C peroxidase | Cytochrome C |
| 2PCC_A_B | KB87A | RIM | 0.901871 | 11148036 | Cytochrome C peroxidase | Cytochrome C |
| 2PCC_A_B | KB72A | RIM | 0.303669 | 11148036 | Cytochrome C peroxidase | Cytochrome C |
| 2SIC_E_I | MI73A | COR | 0.217859 | 8276767 | Subtilisin BPN | Streptomyces subtilisin inhibitor |
| 2VLJ_ABC_DE | DE32A | COR | 1.573201 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | NE55A | RIM | 1.129622 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | DE56A | COR | 0.132202 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | QE58A | COR | 0.495437 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SE99A | SUP | -0.03521 | 18275829 | HL-A2-flu | JM22 |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 2VLJ_ABC_DE | YE101A | COR | 0.232574 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SD31A | COR | 0.627639 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | SD32A | COR | 1.038295 | 18275829 | HL-A2-flu | JM22 |
| 2VLJ_ABC_DE | QD34A | SUP | 0.975874 | 18275829 | HL-A2-flu | JM22 |
| 2WPT_A_B | EA30A | RIM | 1.733953 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | DA33A | SUP | -0.1322 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NA34A | COR | -0.37987 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | VA37A | SUP | 3.8093 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RA38A | RIM | -1.11093 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | EA41A | COR | 4.50317 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SA50A | COR | 2.425703 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | PA56A | SUP | 2.927686 | 9718299 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RB54A | COR | 1.134343 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB72A | COR | 0.917825 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB74A | COR | -0.84931 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB75A | SUP | 1.237639 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB77A | COR | -0.61164 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB78A | SUP | -0.14954 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB84A | SUP | -0.09361 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | FB86A | COR | 1.170181 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | TB87A | SUP | 0.52622 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | QB92A | SUP | 0.900139 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | KB97A | COR | 0.497408 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | VB98A | SUP | 0.119302 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | RB54A | COR | 0.61 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB72A | COR | 0.48481 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 2WPT_A_B | SB74A | COR | 0.581094 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | NB75A | SUP | 1.265559 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB77A | COR | -0.30264 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB78A | SUP | -0.04087 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | SB84A | SUP | -0.04087 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | FB86A | COR | 0.945562 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | TB87A | SUP | 0.226905 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | QB92A | SUP | -0.1322 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | KB97A | COR | 0.801228 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 2WPT_A_B | VB98A | SUP | 0.410656 | 18471830 | Colicin E2 immunity protein | Colicin E9 DNase |
| 3BK3_A_C | LC1A | RIM | 0 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC2A | RIM | 1.038654 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC18A | COR | 0.486392 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC21A | COR | 1.307704 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BK3_A_C | IC27A | COR | 1.261516 | 18477456 | Bone morphogenetic protein-2 | Crossveinless 2 |
| 3BN9_B_CD | IB41A | COR | 0 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | IB60A | COR | 0.835589 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | DB60aA | RIM | 0.422577 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | DB60bA | RIM | 0.311247 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | RB60cA | RIM | -0.04502 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | FB94A | SUP | 0.639528 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | NB95A | COR | 0.773691 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | TB98A | COR | 1.13256 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | HB143A | COR | 0.085101 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | YB146A | RIM | 1.085138 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | QB174A | RIM | -0.03471 | 17475279 | Membrane-type serine protease 1 | E2 Fab |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 3BN9_B_CD | QB175A | COR | 2.510108 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | DB217A | COR | 0.566465 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | QB221aA | RIM | 0.706027 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BN9_B_CD | KB224A | COR | 0.78532 | 17475279 | Membrane-type serine protease 1 | E2 Fab |
| 3BP8_A_C | FA136A | SUP | 0.708987 | 18319344 | Mlc transcription regulator | PTS glucose-specific enzyme EIICB |
| 3HFM_HL_Y | YY20A | COR | 4.878217 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | NL32A | SUP | 5.109633 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY96A | SUP | 6.991293 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | NL31A | COR | 5.216466 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YL50A | COR | 4.559636 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97A | COR | 6.169981 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YH33A | COR | 6.037779 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | WH98A | COR | 5.513151 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DH32A | COR | 1.899077 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | COR | 1.027668 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | SH31A | COR | 0.170438 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YH50A | SUP | 7.322839 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | QL53A | COR | 0.953515 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YL96A | COR | 2.708257 | 10338006 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YH53A | COR | 3.198175 | 7629185 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | YH58A | COR | 1.649124 | 7629185 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101A | COR | 1.208733 | 7683415 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | COR | 0.821313 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | KY97A | COR | 5.558082 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101A | COR | 1.52264 | 9761467 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | HY15A | SUP | -0.44552 | 9761468 | HyHEL-10 | HEW Lysozyme |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 3HFM_HL_Y | YY20A | COR | 4.273437 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY21A | COR | 0.862188 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | WY63A | SUP | 0.31933 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | RY73A | RIM | -0.33155 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | LY75A | COR | 0.704771 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | TY89A | RIM | 0 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | NY93A | COR | 0.211313 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | IY98A | SUP | 0 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | SY100A | COR | 0.267779 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3HFM_HL_Y | DY101A | COR | 0.953515 | 9761468 | HyHEL-10 | HEW Lysozyme |
| 3NPS_A_BC | IA41A | COR | 0.64183 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | DA60aA | RIM | 0.340139 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | DA60bA | RIM | 1.067948 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | RA60cA | RIM | -1.06322 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | DA96A | RIM | 1.507709 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | FA97A | COR | 0.463747 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | TA98A | RIM | 0.724136 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | HA143A | COR | 1.875931 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | YA146A | COR | 1.775668 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | TA150A | RIM | 0.175048 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | QA174A | RIM | -0.05925 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | QA175A | COR | 0.741239 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3NPS_A_BC | QA221aA | COR | -0.04094 | 17475279 | Membrane-type serine protease 1 | S4 Fab |
| 3SGB_E_I | LI18A | COR | 2.98916 | 9047374 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | TI17A | COR | 3.591963 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | EI19A | RIM | 1.019236 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |

| Protein | Mutation(s)_PDB | Location(s) | ΔΔG | Reference | Protein 1 | Protein 2 |
|---|---|---|---|---|---|---|
| 3SGB_E_I | YI20A | COR | 1.943608 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | RI21A | RIM | 0.053752 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | PI14A | RIM | -0.1895 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | KI13A | COR | -2.57214 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | GI32A | COR | 1.222797 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | NI36A | RIM | 0.331072 | 11171964 | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | GI32A | COR | 1.364171 | | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | TI17A | COR | 3.229379 | | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 3SGB_E_I | LI18A | COR | 2.956299 | | Streptomyces griseus proteinase B | Turkey ovomucoid third domain |
| 4CPA_A_I | VI38A | COR | 2.325533 | 8063780 | Carboxypeptidase A | Potato carboxypeptidase inhibitor |

# Bibliography


ALBECK, S., UNGER, R. & SCHREIBER, G. 2000. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J Mol Biol,* 298**,** 503-20.

ALEXANDER-BRETT, J. M. & FREMONT, D. H. 2007. Dual GPCR and GAG mimicry by the M3 chemokine decoy receptor. *J Exp Med,* 204**,** 3157-72.

ANDRUSIER, N., NUSSINOV, R. & WOLFSON, H. J. 2007. FireDock: fast interaction refinement in molecular docking. *Proteins,* 69**,** 139-59.

ARKIN, M. R. & WELLS, J. A. 2004. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov,* 3**,** 301-17.

AUDIE, J. & SCARLATA, S. 2007. A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophys Chem,* 129**,** 198-211.

BAI, H., YANG, K., YU, D., ZHANG, C., CHEN, F. & LAI, L. 2011. Predicting kinetic constants of protein-protein interactions based on structural properties. *Proteins,* 79**,** 720-34.

BAIGELMAN, W. & CHODOSH, S. 1977. BRonchodilator action of the anticholinergic drug, ipratropium bromide (sch 1000), as an aerosol in chronic bronchitis and asthma. *Chest,* 71**,** 324-328.

BAKER, N. M. & DER, C. J. 2013. Cancer: Drug for an 'undruggable' protein. *Nature,* 497**,** 577-8.

BAS, D. C., ROGERS, D. M. & JENSEN, J. H. 2008. Very fast prediction and rationalization of pKa values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics,* 73**,** 765-783.

BEN-NAIM, A. 2006. On the driving forces for protein-protein association. *J Chem Phys,* 125**,** 24901.

BISHOP, C. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer.

BOGAN, A. A. & THORN, K. S. 1998. Anatomy of hot spots in protein interfaces. *J Mol Biol,* 280**,** 1-9.

BOLON, D. N., GRANT, R. A., BAKER, T. A. & SAUER, R. T. 2005. Specificity versus stability in computational protein design. *Proc Natl Acad Sci U S A,* 102**,** 12724-9.

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

BRAY, D. 2009. *Wetware: A Computer in Every Living Cell*, Yale University Press.
BREIMAN, L. 2001a. Random Forests. *Mach. Learn.,* 45**,** 5-32.
BREIMAN, L. 2001b. Random Forests. *Machine Learning,* 45**,** 5-32.
BROOKS, B. R., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. & KARPLUS, M. 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30**,** 1545-614.
CAPRIOTTI, E., FARISELLI, P. & CASADIO, R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res,* 33**,** W306-10.
CARRA, C., SAHA, J. & CUCINOTTA, F. A. 2012. Theoretical prediction of the binding free energy for mutants of replication protein A. *J Mol Model,* 18**,** 3035-49.
CARRINGTON, B. J. & MANCERA, R. L. 2004. Comparative estimation of vibrational entropy changes in proteins through normal modes analysis. *J Mol Graph Model,* 23**,** 167-74.
CASTRO, M. J. & ANDERSON, S. 1996. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to beta-trypsin and alpha-chymotrypsin. *Biochemistry,* 35**,** 11435-46.
CHANDRASEKARAN, R. & RAMACHANDRAN, G. N. 1970. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int J Protein Res,* 2**,** 223-33.
CHAUDHURY, S., LYSKOV, S. & GRAY, J. J. 2010. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics,* 26**,** 689-91.
CHEN, W., CHANG, C. E. & GILSON, M. K. 2004. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. *Biophys J,* 87**,** 3035-49.
CHERFILS, J., DUQUERROY, S. & JANIN, J. 1991. Protein-protein recognition analyzed by docking simulation. *Proteins,* 11**,** 271-80.
CHO, K. I., KIM, D. & LEE, D. 2009. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res,* 37**,** 2672-87.
CHO, K. I., LEE, K., LEE, K. H., KIM, D. & LEE, D. 2006. Specificity of molecular interactions in transient protein-protein interaction interfaces. *Proteins,* 65**,** 593-606.
CHOTHIA, C. & JANIN, J. 1975. Principles of protein-protein recognition. *Nature,* 256**,** 705-8.
CHUANG, G. Y., KOZAKOV, D., BRENKE, R., COMEAU, S. R. & VAJDA, S. 2008. DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J,* 95**,** 4217-27.
CLACKSON, T., ULTSCH, M. H., WELLS, J. A. & DE VOS, A. M. 1998. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J Mol Biol,* 277**,** 1111-28.
CLACKSON, T. & WELLS, J. A. 1995. A hot spot of binding energy in a hormone-receptor interface. *Science,* 267**,** 383-6.
COPELAND, R. A., POMPLIANO, D. L. & MEEK, T. D. 2006. Drug-target residence time and its implications for lead optimization. *Nat Rev Drug Discov,* 5**,** 730-9.
CUNNINGHAM, B. C. & WELLS, J. A. 1989. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science,* 244**,** 1081-5.
DARNELL, S. J., PAGE, D. & MITCHELL, J. C. 2007. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins,* 68**,** 813-23.
DAVID, A., RAZALI, R., WASS, M. N. & STERNBERG, M. J. 2012. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat,* 33**,** 359-63.

DE GROOT, B. L., VAN AALTEN, D. M., SCHEEK, R. M., AMADEI, A., VRIEND, G. & BERENDSEN, H. J. 1997. Prediction of protein conformational freedom from distance constraints. *Proteins,* 29**,** 240-51.

DEJACO, C., DUFTNER, C., GRUBECK-LOEBENSTEIN, B. & SCHIRMER, M. 2006. Imbalance of regulatory T cells in human autoimmune diseases. *Immunology,* 117**,** 289-300.

DILL, K. A. 1990. Dominant forces in protein folding. *Biochemistry,* 29**,** 7133-55.

DISSE, B., SPECK, G. A., ROMINGER, K. L., WITEK, T. J., JR. & HAMMER, R. 1999. Tiotropium (Spiriva): mechanistical considerations and clinical profile in obstructive lung disease. *Life Sci,* 64**,** 457-64.

DOLINSKY, T. J., NIELSEN, J. E., MCCAMMON, J. A. & BAKER, N. A. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res,* 32**,** W665-7.

DUNBRACK, R. L., JR. & COHEN, F. E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci,* 6**,** 1661-81.

DUNBRACK, R. L., JR. & KARPLUS, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol,* 230**,** 543-74.

DUNBRACK, R. L., JR. & KARPLUS, M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol,* 1**,** 334-40.

EDGAR, J. D. 2008. T cell immunodeficiency. *J Clin Pathol,* 61**,** 988-93.

FENG, Y., KLOCZKOWSKI, A. & JERNIGAN, R. L. 2010. Potentials 'R' Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics,* 11**,** 92.

FINKELSTEIN, A. V. & JANIN, J. 1989. The price of lost freedom: entropy of bimolecular complex formation. *Protein Eng,* 3**,** 1-3.

FISCHER, T. B., ARUNACHALAM, K. V., BAILEY, D., MANGUAL, V., BAKHRU, S., RUSSO, R., HUANG, D., PACZKOWSKI, M., LALCHANDANI, V., RAMACHANDRA, C., ELLISON, B., GALER, S., SHAPLEY, J., FUENTES, E. & TSAI, J. 2003. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics,* 19**,** 1453-4.

FLEISHMAN, S. J., WHITEHEAD, T. A., EKIERT, D. C., DREYFUS, C., CORN, J. E., STRAUCH, E. M., WILSON, I. A. & BAKER, D. 2011. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science,* 332**,** 816-21.

FRIEDMAN, J. 1991. Multivariate Adaptive Regression Splines. *The Annals of Statistics,* 19**,** 1-67.

FRY, D. C. 2012. Small-molecule inhibitors of protein-protein interactions: how to mimic a protein partner. *Curr Pharm Des,* 18**,** 4679-84.

FULTON, A. B. 1982. How crowded is the cytoplasm? *Cell,* 30**,** 345-347.

GILSON, M. K., GIVEN, J. A., BUSH, B. L. & MCCAMMON, J. A. 1997. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J,* 72**,** 1047-69.

GILSON, M. K. & ZHOU, H. X. 2007. Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct,* 36**,** 21-42.

GREGERSEN, N., BROSS, P., JORGENSEN, M. M., CORYDON, T. J. & ANDRESEN, B. S. 2000. Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders. *J Inherit Metab Dis,* 23**,** 441-7.

GREGORET, L. M. & SAUER, R. T. 1993. Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci U S A,* 90**,** 4246-50.

GROSDIDIER, S. & FERNANDEZ-RECIO, J. 2012. Protein-protein Docking and Hot-spot Prediction for Drug Discovery. *Curr Pharm Des,* 18**,** 4607-18.

HABER, D. A. & SETTLEMAN, J. 2007. Cancer: drivers and passengers. *Nature,* 446**,** 145-6.

HAJDUK, P. J., HUTH, J. R. & TSE, C. 2005. Predicting protein druggability. *Drug Discov Today,* 10**,** 1675-82.

HAMP, T. & ROST, B. 2012. Alternative protein-protein interfaces are frequent exceptions. *PLoS Comput Biol,* 8**,** e1002623.

HILDEBRAND, J. H. 1979. Is there a “hydrophobic effect”? *Proceedings of the National Academy of Sciences,* 76**,** 194.

HOLDGATE, G. A. & GILL, A. L. 2011. Kinetic efficiency: the missing metric for enhancing compound quality? *Drug Discov Today,* 16**,** 910-3.

HONIG, B., SHARP, K. & YANG, A. S. 1993. Macroscopic models of aqueous solutions: biological and chemical applications. *The Journal of Physical Chemistry,* 97**,** 1101-1109.

HOROVITZ, A. 1996. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold Des,* 1**,** R121-6.

HORTON, N. & LEWIS, M. 1992. Calculation of the free energy of association for protein complexes. *Protein Sci,* 1**,** 169-81.

HUANG, S. Y. & ZOU, X. 2010. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J Chem Inf Model,* 50**,** 262-73.

HUSE, M. 2009. The T-cell-receptor signaling network. *J Cell Sci,* 122**,** 1269-73.

JAMESON, S. C. 1998. T cell receptor antagonism in vivo, at last. *Proc Natl Acad Sci U S A,* 95**,** 14001-2.

JHA, R. K., LEAVER-FAY, A., YIN, S., WU, Y., BUTTERFOSS, G. L., SZYPERSKI, T., DOKHOLYAN, N. V. & KUHLMAN, B. 2010. Computational design of a PAK1 binding protein. *J Mol Biol,* 400**,** 257-70.

JIANG, L., GAO, Y., MAO, F., LIU, Z. & LAI, L. 2002. Potential of mean force for protein-protein interaction studies. *Proteins,* 46**,** 190-6.

JIANG, L., KUHLMAN, B., KORTEMME, T. & BAKER, D. 2005. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins,* 58**,** 893-904.

JIN, L., WANG, W. & FANG, G. 2014. Targeting protein-protein interaction by small molecules. *Annu Rev Pharmacol Toxicol,* 54**,** 435-56.

JIN, L. & WELLS, J. A. 1994. Dissecting the energetics of an antibody-antigen interface by alanine shaving and molecular grafting. *Protein Sci,* 3**,** 2351-7.

KASTRITIS, P. L. & BONVIN, A. M. 2010. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res,* 9**,** 2216-25.

KASTRITIS, P. L. & BONVIN, A. M. 2013. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J R Soc Interface,* 10**,** 20120835.

KASTRITIS, P. L., MOAL, I. H., HWANG, H., WENG, Z., BATES, P. A., BONVIN, A. M. & JANIN, J. 2011. A structure-based benchmark for protein-protein binding affinity. *Protein Sci,* 20**,** 482-91.

KATO, M., KOMAMURA, K. & KITAKAZE, M. 2006. Tiotropium, a Novel Muscarinic M3 Receptor Antagonist, Improved Symptoms of Chronic Obstructive Pulmonary Disease Complicated by Chronic Heart Failure. *Circulation Journal,* 70**,** 1658-1660.

KAUZMANN, W. 1959. Some Factors in the Interpretation of Protein Denaturation. Elsevier.

KERSH, G. J. & ALLEN, P. M. 1996. Essential flexibility in the T-cell recognition of antigen. *Nature,* 380**,** 495-498.

KESKIN, O., MA, B. & NUSSINOV, R. 2005. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol,* 345**,** 1281-94.

KIEL, C., SELZER, T., SHAUL, Y., SCHREIBER, G. & HERRMANN, C. 2004. Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate

of association by stabilizing the encounter complex. *Proc Natl Acad Sci U S A,* 101**,** 9223-8.

KORTEMME, T. & BAKER, D. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A,* 99**,** 14116-21.

KORTEMME, T., JOACHIMIAK, L. A., BULLOCK, A. N., SCHULER, A. D., STODDARD, B. L. & BAKER, D. 2004. Computational redesign of protein-protein interaction specificity. *Nat Struct Mol Biol,* 11**,** 371-9.

KRYSTEK, S., STOUCH, T. & NOVOTNY, J. 1993. Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J Mol Biol,* 234**,** 661-79.

LASKOWSKI, R. A. 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph,* 13**,** 323-30, 307-8.

LAVIGNE, P., BAGU, J. R., BOYKO, R., WILLARD, L., HOLMES, C. F. & SYKES, B. D. 2000. Structure-based thermodynamic analysis of the dissociation of protein phosphatase-1 catalytic subunit and microcystin-LR docked complexes. *Protein Sci,* 9**,** 252-64.

LAZARIDIS, T. & KARPLUS, M. 1999. Effective energy function for proteins in solution. *Proteins,* 35**,** 133-52.

LEACH, A. R. 2009. *Molecular Modelling: Principles And Applications, 2/E*, Pearson Education.

LEVY, E. D. 2010. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol,* 403**,** 660-70.

LEWITZKY, M., SIMISTER, P. C. & FELLER, S. M. 2012. Beyond 'furballs' and 'dumpling soups' - towards a molecular architecture of signaling complexes and networks. *FEBS Lett,* 586**,** 2740-50.

LIANG, S., LIU, S., ZHANG, C. & ZHOU, Y. 2007. A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. *Proteins,* 69**,** 244-53.

LIN, J. & WEISS, A. 2001. T cell receptor signalling. *J Cell Sci,* 114**,** 243-4.

LINS, L. & BRASSEUR, R. 1995. The hydrophobic effect in protein folding. *FASEB J,* 9**,** 535-40.

LISE, S., ARCHAMBEAU, C., PONTIL, M. & JONES, D. T. 2009. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics,* 10**,** 365.

LISE, S., BUCHAN, D., PONTIL, M. & JONES, D. T. 2011. Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One,* 6**,** e16774.

LIU, S., LIU, S., ZHU, X., LIANG, H., CAO, A., CHANG, Z. & LAI, L. 2007. Nonnatural protein-protein interaction-pair design by key residues grafting. *Proc Natl Acad Sci U S A,* 104**,** 5330-5.

LIU, S. & VAKSER, I. A. 2011. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics,* 12**,** 280.

LIU, S., ZHANG, C., ZHOU, H. & ZHOU, Y. 2004. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins,* 56**,** 93-101.

LU, H., LU, L. & SKOLNICK, J. 2003. Development of unified statistical potentials describing protein-protein interactions. *Biophys J,* 84**,** 1895-901.

LU, H. & TONGE, P. J. 2010. Drug-target residence time: critical information for lead optimization. *Curr Opin Chem Biol,* 14**,** 467-74.

LU, M., DOUSIS, A. D. & MA, J. 2008. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J Mol Biol,* 376**,** 288-301.

LUBY-PHELPS, K. 2000. Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. *Int Rev Cytol,* 192**,** 189-221.

MA, B. & NUSSINOV, R. 2007. Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. *Curr Top Med Chem,* 7**,** 999-1005.

MA, D., GUO, Y., LUO, J., PU, X. & LI, M. 2014. Prediction of protein–protein binding affinity using diverse protein–protein interface features. *Chemometrics and Intelligent Laboratory Systems,* 138**,** 7-13.

MA, X. H., WANG, C. X., LI, C. H. & CHEN, W. Z. 2002. A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng,* 15**,** 677-81.

MANDELL, D. J. & KORTEMME, T. 2009. Computer-aided design of functional protein interactions. *Nat Chem Biol,* 5**,** 797-807.

MCDONALD, I. K. & THORNTON, J. M. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol,* 238**,** 777-93.

MILLER, S. D., TURLEY, D. M. & PODOJIL, J. R. 2007. Antigen-specific tolerance strategies for the prevention and treatment of autoimmune disease. *Nat Rev Immunol,* 7**,** 665-77.

MISURA, K. M., MOROZOV, A. V. & BAKER, D. 2004. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J Mol Biol,* 342**,** 651-64.

MITCHELL, A., ROMANO, G. H., GROISMAN, B., YONA, A., DEKEL, E., KUPIEC, M., DAHAN, O. & PILPEL, Y. 2009. Adaptive prediction of environmental changes by microorganisms. *Nature,* 460**,** 220-4.

MITRA, P. & PAL, D. 2010. New measures for estimating surface complementarity and packing at protein-protein interfaces. *FEBS Lett,* 584**,** 1163-8.

MIYAZAWA, S. & JERNIGAN, R. L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol,* 256**,** 623-44.

MOAL, I. H., AGIUS, R. & BATES, P. A. 2011. Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics,* 27**,** 3002-9.

MOAL, I. H. & BATES, P. A. 2012. Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS Comput Biol,* 8**,** e1002351.

MOAL, I. H. & FERNANDEZ-RECIO, J. 2012. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics,* 28**,** 2600-7.

MOORE, J. M., PATAPOFF, T. W. & CROMWELL, M. E. 1999. Kinetics and thermodynamics of dimer formation and dissociation for a recombinant humanized monoclonal antibody to vascular endothelial growth factor. *Biochemistry,* 38**,** 13960-7.

MORETTI, R., FLEISHMAN, S. J., AGIUS, R., TORCHALA, M., BATES, P. A., KASTRITIS, P. L., RODRIGUES, J. P., TRELLET, M., BONVIN, A. M., CUI, M., ROOMAN, M., GILLIS, D., DEHOUCK, Y., MOAL, I., ROMERO-DURANA, M., PEREZ-CANO, L., PALLARA, C., JIMENEZ, B., FERNANDEZ-RECIO, J., FLORES, S., PACELLA, M., PRANEETH KILAMBI, K., GRAY, J. J., POPOV, P., GRUDININ, S., ESQUIVEL-RODRIGUEZ, J., KIHARA, D., ZHAO, N., KORKIN, D., ZHU, X., DEMERDASH, O. N., MITCHELL, J. C., KANAMORI, E., TSUCHIYA, Y., NAKAMURA, H., LEE, H., PARK, H., SEOK, C., SARMIENTO, J., LIANG, S., TERAGUCHI, S., STANDLEY, D. M., SHIMOYAMA, H., TERASHI, G., TAKEDA-SHITAKA, M., IWADATE, M., UMEYAMA, H., BEGLOV, D., HALL, D. R., KOZAKOV, D., VAJDA, S., PIERCE, B. G., HWANG, H., VREVEN, T., WENG, Z., HUANG, Y., LI, H., YANG, X., JI, X., LIU, S., XIAO, Y., ZACHARIAS, M., QIN, S., ZHOU, H. X., HUANG, S. Y., ZOU, X., VELANKAR, S., JANIN, J., WODAK, S. J. & BAKER, D. 2013. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins,* 81**,** 1980-7.

MORROW, J. K. & ZHANG, S. 2012. Computational prediction of protein hot spot residues. *Curr Pharm Des,* 18**,** 1255-65.

MULLARD, A. 2012. Protein–protein interaction inhibitors get into the groove. *Nat Rev Drug Discov,* 11**,** 173-175.

NAUCHITEL, V., VILLAVERDE, M. C. & SUSSMAN, F. 1995. Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease. *Protein Sci,* 4**,** 1356-64.

NOOREN, I. M. & THORNTON, J. M. 2003. Diversity of protein-protein interactions. *EMBO J,* 22**,** 3486-92.

NOSKOV, S. Y. & LIM, C. 2001. Free energy decomposition of protein-protein interactions. *Biophys J,* 81**,** 737-50.

NOVOTNY, J., BRUCCOLERI, R. E. & SAUL, F. A. 1989. On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry,* 28**,** 4735-49.

OFRAN, Y. & ROST, B. 2007. Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol,* 3**,** e119.

OLTERSDORF, T., ELMORE, S. W., SHOEMAKER, A. R., ARMSTRONG, R. C., AUGERI, D. J., BELLI, B. A., BRUNCKO, M., DECKWERTH, T. L., DINGES, J., HAJDUK, P. J., JOSEPH, M. K., KITADA, S., KORSMEYER, S. J., KUNZER, A. R., LETAI, A., LI, C., MITTEN, M. J., NETTESHEIM, D. G., NG, S., NIMMER, P. M., O'CONNOR, J. M., OLEKSIJEW, A., PETROS, A. M., REED, J. C., SHEN, W., TAHIR, S. K., THOMPSON, C. B., TOMASELLI, K. J., WANG, B., WENDT, M. D., ZHANG, H., FESIK, S. W. & ROSENBERG, S. H. 2005. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature,* 435**,** 677-81.

PAL, G., ULTSCH, M. H., CLARK, K. P., CURRELL, B., KOSSIAKOFF, A. A. & SIDHU, S. S. 2005. Intramolecular cooperativity in a protein binding site assessed by combinatorial shotgun scanning mutagenesis. *J Mol Biol,* 347**,** 489-94.

PAN, A. C., BORHANI, D. W., DROR, R. O. & SHAW, D. E. 2013. Molecular determinants of drug-receptor binding kinetics. *Drug Discov Today,* 18**,** 667-73.

QIU, D., SHENKIN, P. S., HOLLINGER, F. P. & STILL, W. C. 1997. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *The Journal of Physical Chemistry A,* 101**,** 3005-3014.

QUINLAN, J. R. Learning with continuous classes. Proceedings of the 5th Australian joint Conference on Artificial Intelligence, 1992. Singapore, 343-348.

RAJGARIA, R., MCALLISTER, S. R. & FLOUDAS, C. A. 2006. A novel high resolution Calpha--Calpha distance dependent force field based on a high quality decoy set. *Proteins,* 65**,** 726-41.

RAJGARIA, R., MCALLISTER, S. R. & FLOUDAS, C. A. 2008. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins,* 70**,** 950-70.

REICHMANN, D., RAHAT, O., ALBECK, S., MEGED, R., DYM, O. & SCHREIBER, G. 2005. The modular architecture of protein-protein binding interfaces. *Proc Natl Acad Sci U S A,* 102**,** 57-62.

REYNOLDS, C., DAMERELL, D. & JONES, S. 2009. ProtorP: a protein-protein interaction analysis server. *Bioinformatics,* 25**,** 413-4.

RICCI, F., ROKACH, L. & SHAPIRA, B. 2011. Introduction to Recommender Systems Handbook. *In:* RICCI, F., ROKACH, L., SHAPIRA, B. & KANTOR, P. (eds.) *Recommender Systems Handbook.* Springer US.

ROSENBLATT, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev,* 65**,** 386-408.

ROSETTE, C., WERLEN, G., DANIELS, M. A., HOLMAN, P. O., ALAM, S. M., TRAVERS, P. J., GASCOIGNE, N. R., PALMER, E. & JAMESON, S. C. 2001. The impact of duration

versus extent of TCR occupancy on T cell activation: a revision of the kinetic proofreading model. *Immunity,* 15**,** 59-70.

RYKUNOV, D. & FISER, A. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics,* 11**,** 128.

SCHAEFER, M. & KARPLUS, M. 1996. A Comprehensive Analytical Treatment of Continuum Electrostatics. *The Journal of Physical Chemistry,* 100**,** 1578-1599.

SCHYMKOWITZ, J., BORG, J., STRICHER, F., NYS, R., ROUSSEAU, F. & SERRANO, L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res,* 33**,** W382-8.

SHEN, M. Y. & SALI, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci,* 15**,** 2507-24.

SHULTZABERGER, R. K., ROBERTS, L. R., LYAKHOV, I. G., SIDOROV, I. A., STEPHEN, A. G., FISHER, R. J. & SCHNEIDER, T. D. 2007. Correlation between binding rate constants and individual information of E. coli Fis binding sites. *Nucleic Acids Res,* 35**,** 5275-83.

SIMONS, K. T., RUCZINSKI, I., KOOPERBERG, C., FOX, B. A., BYSTROFF, C. & BAKER, D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins,* 34**,** 82-95.

SITKOFF, D., SHARP, K. A. & HONIG, B. 1994. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *The Journal of Physical Chemistry,* 98**,** 1978-1988.

STONE, J. D., CHERVIN, A. S. & KRANZ, D. M. 2009. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology,* 126**,** 165-76.

SU, Y., ZHOU, A., XIA, X., LI, W. & SUN, Z. 2009. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Sci,* 18**,** 2550-8.

SWINNEY, D. C. 2004. Biochemical mechanisms of drug action: what does it take for success? *Nat Rev Drug Discov,* 3**,** 801-8.

THANGUDU, R. R., BRYANT, S. H., PANCHENKO, A. R. & MADEJ, T. 2012. Modulating protein-protein interactions with small molecules: the importance of binding hotspots. *J Mol Biol,* 415**,** 443-53.

TOBI, D. 2010. Designing coarse grained-and atom based-potentials for protein-protein docking. *BMC Struct Biol,* 10**,** 40.

TOBI, D. & BAHAR, I. 2006. Optimal design of protein docking potentials: efficiency and limitations. *Proteins,* 62**,** 970-81.

TONG, W., LI, L. & WENG, Z. 2004. Computational prediction of binding hotspots. *Conf Proc IEEE Eng Med Biol Soc,* 4**,** 2980-3.

TSAI, C. J., LIN, S. L., WOLFSON, H. J. & NUSSINOV, R. 1997. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci,* 6**,** 53-64.

TUNCBAG, N., GURSOY, A. & KESKIN, O. 2009. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics,* 25**,** 1513-20.

TUNCBAG, N., KESKIN, O. & GURSOY, A. 2010. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res,* 38**,** W402-6.

TURK, M. A. & PENTLAND, A. P. Face recognition using eigenfaces. Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on, 3-6 Jun 1991 1991. 586-591.

VAJDA, S., WENG, Z., ROSENFELD, R. & DELISI, C. 1994. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry,* 33**,** 13977-88.

VALITUTTI, S., MULLER, S., CELLA, M., PADOVAN, E. & LANZAVECCHIA, A. 1995. Serial triggering of many T-cell receptors by a few peptide-MHC complexes. *Nature,* 375**,** 148-51.

VASSILEV, L. T., VU, B. T., GRAVES, B., CARVAJAL, D., PODLASKI, F., FILIPOVIC, Z., KONG, N., KAMMLOTT, U., LUKACS, C., KLEIN, C., FOTOUHI, N. & LIU, E. A. 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science,* 303**,** 844-8.

VREVEN, T., HWANG, H., PIERCE, B. G. & WENG, Z. 2012. Prediction of protein-protein binding free energies. *Protein Sci,* 21**,** 396-404.

WAINREB, G., WOLF, L., ASHKENAZY, H., DEHOUCK, Y. & BEN-TAL, N. 2011. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics,* 27**,** 3286-92.

WALLQVIST, A., JERNIGAN, R. L. & COVELL, D. G. 1995. A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci,* 4**,** 1881-903.

WANG, L., LIU, Z. P., ZHANG, X. S. & CHEN, L. 2012. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng Des Sel,* 25**,** 119-26.

WANG, Y. & WITTEN, I. H. 1996. Induction of model trees for predicting continuous classes.

WEIKL, T. R. & VON DEUSTER, C. 2009. Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. *Proteins,* 75**,** 104-10.

WENG, Z., DELISI, C. & VAJDA, S. 1997. Empirical free energy calculation: comparison to calorimetric data. *Protein Sci,* 6**,** 1976-84.

WHITEHEAD, T. A., CHEVALIER, A., SONG, Y., DREYFUS, C., FLEISHMAN, S. J., DE MATTOS, C., MYERS, C. A., KAMISETTY, H., BLAIR, P., WILSON, I. A. & BAKER, D. 2012. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol,* 30**,** 543-8.

WOLPERT, D. 1992. Stacked Generalization. *Neural Networks,* 5**,** 241-259.

XIA, D., ESSER, L., SINGH, S. K., GUO, F. & MAURIZI, M. R. 2004. Crystallographic investigation of peptide binding sites in the N-domain of the ClpA chaperone. *J Struct Biol,* 146**,** 166-79.

XIA, J. F., ZHAO, X. M., SONG, J. & HUANG, D. S. 2010. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics,* 11**,** 174.

XU, D., LIN, S. L. & NUSSINOV, R. 1997. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J Mol Biol,* 265**,** 68-84.

YAN, Z., GUO, L., HU, L. & WANG, J. 2013. Specificity and affinity quantification of protein-protein interactions. *Bioinformatics,* 29**,** 1127-33.

YANG, Y. & ZHOU, Y. 2008. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci,* 17**,** 1212-9.

YATES, C. M. & STERNBERG, M. J. 2013. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J Mol Biol,* 425**,** 3949-63.

YOUNG, L., JERNIGAN, R. L. & COVELL, D. G. 1994. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci,* 3**,** 717-29.

YU, Y. B., PRIVALOV, P. L. & HODGES, R. S. 2001. Contribution of translational and rotational motions to molecular association in aqueous solution. *Biophys J,* 81**,** 1632-42.

YUE, P., LI, Z. & MOULT, J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol,* 353**,** 459-73.

YURA, K. & HAYWARD, S. 2009. The interwinding nature of protein-protein interfaces and its implication for protein complex formation. *Bioinformatics,* 25**,** 3108-13.

ZETH, K., RAVELLI, R. B., PAAL, K., CUSACK, S., BUKAU, B. & DOUGAN, D. A. 2002. Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA. *Nat Struct Biol,* 9**,** 906-11.

ZHANG, C., LIU, S., ZHOU, H. & ZHOU, Y. 2004. The dependence of all-atom statistical potentials on structural training database. *Biophys J,* 86**,** 3349-58.

ZHANG, C., VASMATZIS, G., CORNETTE, J. L. & DELISI, C. 1997. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol,* 267**,** 707-26.

ZHOU, H.-X. 2001. Loops in Proteins Can Be Modeled as Worm-Like Chains. *The Journal of Physical Chemistry B,* 105**,** 6763-6766.

ZHOU, H.-X. 2004. Polymer Models of Protein Stability, Folding, and Interactions†. *Biochemistry,* 43**,** 2141-2154.

ZHOU, H. & SKOLNICK, J. 2011. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J,* 101**,** 2043-52.

ZHU, X. & MITCHELL, J. C. 2011. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins,* 79**,** 2671-83.