2017

Simulation & Experiment Learning From Kinases In Cancer

E. Joseph Jordan
University of Pennsylvania, e.jjordan12@gmail.com

Simulation & Experiment Learning From Kinases In Cancer

Abstract

The decreasing cost of genome sequencing technology has lead to an explosion of information about which mutations are frequently observed in cancer, demonstrating an important role in cancer progression for kinase domain mutations. Many therapies have been developed that target mutations in kinase proteins that lead to constitutive activation. However, a growing body of evidence points to the serious dangers of many kinase ATP competitive inhibitors leading to paradoxical activation in non-constitutively active proteins. The large number of observed mutations and the critical need to only treat patients harboring activating mutations with targeted therapies raises the question of how to classify the thousands of mutations that have been observed. We start with an in depth look at the state of knowledge of the distribution and effects of kinase mutations. We then report on computational methods to understand and predict the effects of kinase domain mutations. Using molecular dynamics simulations of mutant kinases, we show that there is a switch-like network of labile hydrogen bonds that are often perturbed in activating mutations. This is paired with a description of a software platform that has been developed to streamline the execution and analysis of molecular dynamics simulations. We conclude by examining a machine learning method to demonstrate what kinds information derived from protein sequence alone have the most value in distinguishing activating and non-activating mutations.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
Ravi Radhakrishnan

Subject Categories
Bioinformatics | Biophysics

SIMULATION & EXPERIMENT
LEARNING FROM KINASES IN CANCER

E. Joseph Jordan

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Signature ______________________________
Ravi Radhakrishnan, Professor

Graduate Group Chairperson

Signature ______________________________
Kim Sharp, Professor

Dissertation Committee

Arjun Raj, Associate Professor

Roland Dunbrack, Professor

Kate Ferguson, Associate Professor

Mark Lemmon, David A. Sackler Professor of Pharmacology, FRS

Ronen Marmorstein, George W. Raiziss Professor

Dedicated to all the other me's who could have written this thesis but were unable to because poverty or simple misfortune prevented them from pursuing their education.

ACKNOWLEDGMENTS

I would like to acknowledge the futility of such a project in the face of our civilization collapsing.

ABSTRACT

SIMULATION & EXPERIMENT
LEARNING FROM KINASES IN CANCER
E. Joseph Jordan
Ravi Radhakrishnan

The decreasing cost of genome sequencing technology has lead to an explosion of information about which mutations are frequently observed in cancer, demonstrating an important role in cancer progression for kinase domain mutations. Many therapies have been developed that target mutations in kinase proteins that lead to constitutive activation. However, a growing body of evidence points to the serious dangers of many kinase ATP competitive inhibitors leading to paradoxical activation in non-constitutively active proteins. The large number of observed mutations and the critical need to only treat patients harboring activating mutations with targeted therapies raises the question of how to classify the thousands of mutations that have been observed. We start with an in depth look at the state of knowledge of the distribution and effects of kinase mutations. We then report on computational methods to understand and predict the effects of kinase domain mutations. Using molecular dynamics simulations of mutant kinases, we show that there is a switch-like network of labile hydrogen bonds that are often perturbed in activating mutations. This is paired with a description of a software platform that has been developed to streamline the execution and analysis of molecular dynamics simulations. We conclude by examining a machine learning method to demonstrate what kinds information derived from protein sequence alone have the most value in distinguishing activating and non-activating mutations.

Contents

ACKNOWLEDGMENT **iii**

ABSTRACT **iv**

LIST OF TABLES **viii**

LIST OF ILLUSTRATIONS **ix**

1 Introduction **1**

1.1 Cancer 1

1.2 Kinase Biology 3

1.3 Kinase Structure 4

1.3.1 Activation Loop 5

1.3.2 KE salt bridge 6

1.3.3 αC-in vs αC-out 6

1.3.4 DFG-in vs DFG-out 6

1.3.5 R-spine 6

1.4 Philosophical Introduction 8

2 Mutations **11**

2.1 Introduction 11

2.1.1 Analysis of COSMIC 11

2.2 Results 11

2.2.1 Mutations in all proteins 11

2.2.2 Frequently mutated residues 12

2.2.3 Frequently mutated PFAM domains 13

2.2.4 Kinase subdomain mutation distribution 13

2.3 Discussion 14

2.3.1 Description of Kinase Mutations 14

2.3.2 Similarities of kinase mutations 19

3 Kinase MD **21**

3.1 Introduction 21

3.1.1 Kinase Activity 21

3.1.2 Kinase Structure 22

3.1.3 Computational studies 22

3.2 Methods . . . 26
3.2.1 Simulations . . . 26
3.2.2 Analysis . . . 27
3.3 Results . . . 28
3.3.1 Analysis of kinase activity . . . 28
3.3.2 H-bonds analysis . . . 28
3.3.3 Solvent accessible surface area analysis . . . 37
3.4 Discussion . . . 40
3.4.1 H-bonds false positives . . . 40
3.4.2 Utility of SASA . . . 40
3.4.3 Importance of R-spines and H-bonds . . . 41

4 Biophyscode 42
4.1 Introduction . . . 42
4.1.1 Existing MD tools . . . 42
4.1.2 Need for a GROMACS framework . . . 44
4.1.3 Need for "big green button" . . . 45
4.2 Methods . . . 45
4.2.1 Inefficiency in standard methods . . . 45
4.2.2 Modernizing methods . . . 46
4.2.3 Our objective . . . 47
4.2.4 The BioPhysCode method . . . 47
4.3 Results . . . 50
4.3.1 System setup . . . 50
4.3.2 Simulation . . . 52
4.3.3 Analysis . . . 53
4.4 Conclusion . . . 54

5 Svm 57
5.1 Introduction . . . 57
5.1.1 Application of machine learning to protein sequences . . . 57
5.1.2 Support vector machines . . . 59
5.2 Methods . . . 60
5.2.1 Construction of training set . . . 60
5.2.2 Creation of feature vectors . . . 60
5.2.3 SVM parameter search . . . 61
5.2.4 SVM model evaluation . . . 62
5.3 Results . . . 62
5.3.1 Selection of model parameters . . . 62
5.3.2 SVM test set performance . . . 63
5.3.3 SVM performance on synthetic datasets . . . 64
5.4 Future Directions . . . 64

6 Conclusion 66
6.1 Overview of Thesis . . . 66

Appendices **67**

A Kinase mutation experimental measurements **67**

List of Tables

1.1 Major kinase families 4

2.1 COSMIC frequently mutated proteins 12
2.2 COSMIC frequently mutated residues 13
2.3 COSMIC mutations mapped to PFAM domains 14

3.1 Experimental characterizations of mutant kinases 29
3.2 H-bonds occupancy classificatory power by subdomain(s) 32
3.4 Measures of binary classifiers 33
3.5 Labile H-bonds 36
3.7 H-bonds at the same position in kinase domain 36
3.6 H-bonds occupancy classificatory power by residue(s) 37

5.1 Results for separate iterations with different SVM model parameters. 63
5.2 **Prediction results for a number of different methods.** 63
5.3 Prediction results on synthetic datasets. 64

A.1 Experimental characterizations of mutant kinases 67

List of Figures

1.1 ALK conformations and subdomains 5
1.2 Nuances among BRAF structures 7

2.1 Several kinases with large number of mutations 15

3.1 ALK mutation catalytic and transformation assays 30
3.2 Inactive Hbonds contact maps 31
3.3 H-bonds occupancy plots 34
3.4 H-bonds occupancy plots 39

4.1 Schematic of the framework. 56

5.1 Structural Bioinformatics 65

Chapter 1

Introduction

> I insist on the fact that there is generally no growth but only a luxurious squandering of energy in every form!
>
> Georges Bataille *The Accursed Share*

1.1 Cancer

In 2014 there were 2,626,418 deaths recorded in the United States. Of these, 591,700 were recorded as being the result of malignant neoplasm, *cancer*, accounting for 22.5% of deaths and second only to heart disease as the leading cause of death in the most recent year with available data from the Centers for Disease Control and Prevention. Interestingly, many more Americans under the age of 35 die from accidents, murder, or suicide, than from cancer. It is not until the age of 35 that cancer becomes the leading cause of death for Americans, a figure which holds true until the age of 80, when heart disease becomes the leading cause of death. Over the last 60 years, the proportion of deaths from heart disease, stroke, and lower respiratory diseases (e.g. pneumonia) have decreased significantly while cancer has caused a relatively constant fraction of mortality [134] This information, when taken together, suggests both that cancer is a very challenging problem and that if the goal of public health policy is to increase life expectancy, resources may be better allocated to other public health initiatives such as calming traffic or reducing access to firearms. The importance of political interventions in public health is especially highlighted by the fact that the two decades of the twentieth century that saw the largest increases in life expectancy in Britain were the 1910s and 1940s, *despite the massive loss of life incurred by the world wars during these decades*, due to significant increases in public support for social services introduced to support production for the war efforts [222]. Nonetheless, over the past 30 years, mortality rates of lung, prostate, breast, and colorectal cancers have declined and 5 year survival has increased by 20% in blacks and 24% among whites [1]. This progress should not be gainsaid, nor should the concomitant increase in our collective understanding of the underlying biological processes driving cancer progression, which this thesis will hopefully be a small contribution to.

Tumorigenesis was first posited to be an evolutionary process by Nowell in 1976 [178]. Since then, the idea that cancer cells undergo selection on the path from normal to cancerous cell has only gained traction [241]. This idea is predicated upon the knowledge that tumors are composed of a heterogeneous population of cells, in terms of mutations, expression levels, somatic copy number, and epigenetic factors [88, 89]. These factors are then selected upon for robustness and ability to proliferate, alter the tumor microenvironment, and invade neighboring tissues [241]. Of all the functional alterations that a cancer cell undergoes, one of the easiest ones to understand conceptually, and also to measure unambiguously, is that of mutations which may alter protein function. Mutations that ablate a protein's regular function are often observed in cancer cell lines, especially among tumor suppressors such as TP53 and RB1 [88, 89]. The transformation from normal to cancerous cell is often marked by a gradual accumulation of mutations over time that eventually increase the ability of the cancer cell to sustain itself and reproduce [241]. Mutations that confer selective advantage on the cancer cell line are known as driver mutations while passenger mutations are neutral in terms of selective advantage [84]. All cells acquire mutations over their life cycle from differentiation to senescence, with an average of 3 mutations occurring during each round of cell division and no known increase in the rate of mutation or proliferation over the life cycle of a healthy cell. A correlation between total number of cell divisions at age of cancer onset and likelihood of cancer diagnosis has even been demonstrated [243, 242]. It has also been shown that cancer cells acquire mutations at a faster rate than normal cells and that this rate may increase over time [151]. Although some of these mutations may affect protein splicing or regulation, these are not currently thought to contribute greatly to cancer progression as most mutations will fall in intergenic regions or within introns of the coding sequences of proteins [83].

One of the grand challenges of the understanding of cancer progression is to find mechanistic links between molecular alterations and the hallmarks of cancers such as increased proliferation and survival, aggressive invasion and metastasis, evasion of cell death, and increased metabolism [88, 89]. This challenge is also of clinical importance because patient outcome to therapy (both in terms of initial response to therapy and subsequent development of resistance to therapy) is now shown to depend on the genetic alterations (primary or acquired) in the individual patients [11, 43, 137, 176]. Many targets for therapeutic intervention/inhibition have been evaluated in the past few years on the basis of a strong promise provided by preclinical investigations. Nevertheless, experience has shown that the clinical trials are often unsuccessful when the drugs are administered to un-cohorted patient populations. There is thus a growing consensus around the need to employ targeted therapies on select populations of patients classified into cohorts based on molecular/genetic alterations [176]. Rapid genotyping platforms and advances in sequencing cancer genomes allow detection of genetic aberrations in clinical samples. This allows the identification of molecular targets in each individual patient, and also the tracking of acquired molecular changes [45] (expression [53, 105], mutation [252, 101, 233, 173], epigenetic changes [238], post-translational modifications [208], etc.) during the progression of the disease or during treatment. Even with quantitative patient data involving protein expression using immune histochemistry, gene copy number and mutations using sequencing and DNA mutational analysis, and gene expression using florescence *in situ* hybridization, polymerase chain reaction or microarray technology, single-cell imaging [132], mapping this high-dimensional data to a set of viable cellular mechanisms and using them to infer treatment options is a daunting undertaking. A further problem is the heterogeneity of tumors [30], that might show differential expression, copy number, or mutation patterns within a tumor [172], in different tumor areas within an individual [64], or in different individuals. The question then is: how precisely can a tumor be characterized by these techniques? That is, to relate the molecular profile of a given patient to disease prognosis and potentially to the efficacy of therapy is a grand challenge in clinical oncology. This represents a promising opportunity for *in silico* modeling approaches, and the more specific question of how to relate molecular profile to disease prognosis is the subject of this thesis.

Recent large-scale sequencing projects have generated copious data on somatic mutations in cancer. Tumor resequencing efforts have lead to a proliferation of data on cancer somatic mutations [72]. This in turn has lead to efforts to computationally assess which of these mutations are drivers and which are passengers. Most of these efforts have been adaptations of methods developed for predicting whether a single nucleotide polymorphism (SNP), not necessarily cancer related, is deleterious to protein structure and function. Among these methods, the most popular are sequence alignment or structure based, machine learning, and statistical. The sequence- and structure-based methods are fairly accurate and sensitive over the whole genome, but are generally less accurate than protein family specific methods [117]. The statistical methods generally try to assess deleteriousness by calculating the difference between expected and observed mutation rates and locations (e.g. [83]), but give no insight into why a specific mutation is deleterious. Of the protein family specific methods, the machine learning technique support vector machines (SVM) is the most widely used (for a more detailed discussion see chapter 5).

Alternatively, *ab initio* (physics-based) methods such as molecular dynamics (MD) simulations have also been employed to interrogate the effects of mutations on structure, dynamics, and drug interactions at the molecular level. These efforts generally reveal a more detailed picture of how a mutation specifically alters the dynamics of a protein but are generally reserved to study mutations that are observed quite often, as they are much more computationally costly than statistical or machine learning based methods. There are three MD methods that are often used in this domain: (1) nanosecond conventional MD, (2) microsecond conventional MD, (3) enhanced sampling or free energy calculation methods (for references on the application of these techniques to kinases, see the chapter 3). In addition to the computational resources needed to perform MD analyses of mutations, another limitation has been the absence of tools that could allow for easy set-up, completion, and analysis of MD simulation for large numbers of mutations. This issue is addressed in chapter 4 on the BioPhysCode software suite.

1.2 Kinase Biology

There are 518 proteins in the human genome that have been designated as containing kinase domains, of which 478 are not classified as atypical [156]. There are several major families of kinases, the largest of which are listed in Table 1.1 [90].

Most though not all of these catalyze the transfer of the γ phosphate from ATP to a substrate molecule and all share a distinctive fold [109] to be described in the next section. In this work, we will be primarily focused on the TK and TKL families, which generally have as their substrate a serine, threonine, or tyrosine residue, often on another kinase protein. Given the conserved nature of the kinase domain fold, the results should be applicable to

Table 1.1: Major kinase families

Kinase family	kinase family description
ACG	regulate cyclic nucleotides, phospholipids, and calcium
CMGC	cyclin dependent kinases and mitogen associated protein kinases (MAPK)
CAMK	calcium/calmodulin regulated kinases
STE	more MAPKs, homologous with yeast STE signaling pathway
TK	tyrosine kinases
TKL	tyrosine kinase like, usually phosporylate serine or threonine

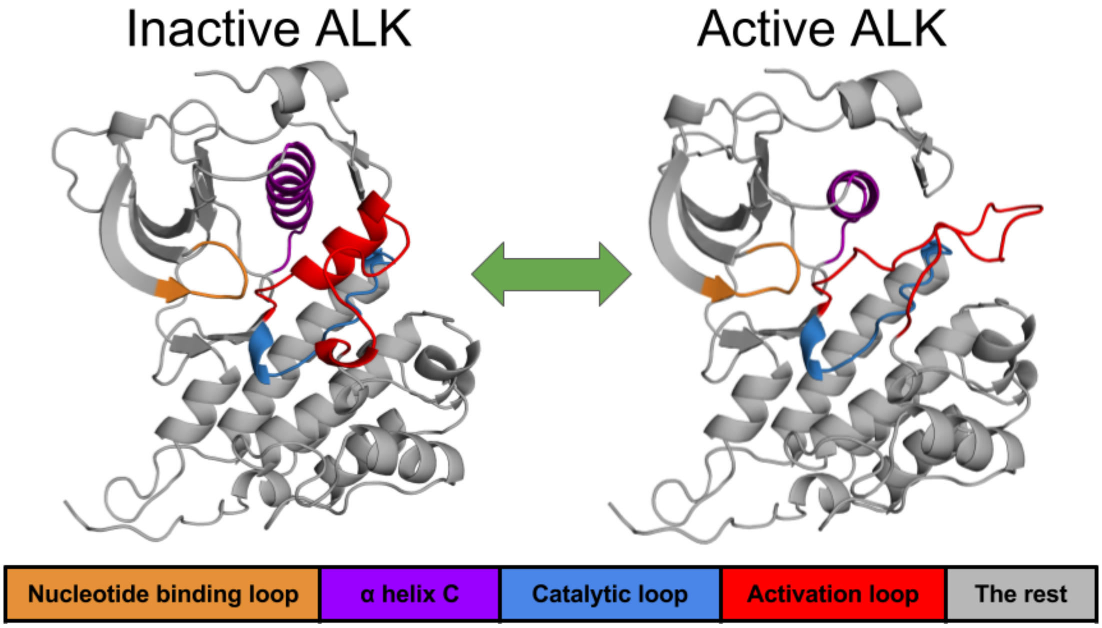
most kinase families, especially CMGC and STE. In addition to the kinase domain, each kinase protein may have domains involved in diverse functions related to ligand binding, scaffolding downstream targets, determining the location of the protein in the cell, and promoting substrate specificity [156, 205].

Kinase proteins are involved in many cellular processes such as signaling, differentiation, and proliferation [147]. Given the role of kinases in cell signaling processes, their role in cancers should not be surprising [89, 147, 156]. For many kinases, constitutive activation upregulates cell proliferation, and in these cases activating mutations will be driver mutations. Many kinase activating mutations have been clinically observed [72], but determining which mutations are activating can still be quite challenging. This is due to the fact that to establish a kinase domain mutation as a driver mutation it is usually necessary to show both that the mutation leads to increased kinase activity in the mutant relative to the wild type protein, as well as to show that this increased activity can lead to increased phosphorylation of downstream targets or increased capacity for transformation (ability to proliferate in the absence of growth factors) in the appropriate cellular context [254, 24, 21]. Even this may be insufficient as there are cases where mutations lead to loss of kinase activity but still lead to increases in phosphorylation of downstream targets [94]. While it is unlikely that any computational method could capture all the nuance of the full biological system, the time and expense required to investigate kinase domain mutations *in vitro* and *in vivo* should be and has been a call to arms for modelers seeking to understand these complex processes.

1.3 Kinase Structure

Kinases are composed of an N-terminal lobe composed mostly of β sheets and a C-terminal lobe composed mostly of α helices. They also have large conformational differences between their active and inactive forms [109] (see Figure 1.1). The largest differences between the active and inactive conformations of kinases are found in the activation loop and αC helix, though other differences also exist. These differences have lead to extensive investigation of how clinically observed mutations disrupt the conformational equilibrium between the active and inactive states, potentially impacting catalysis (see chapter 2 for information on biochemical analysis).

Figure 1.1: ALK conformations and subdomains



1.3.1 Activation Loop

In the active conformation the activation loop (or A-loop) exists in an extended conformation, allowing ATP to bind in the nucleotide binding pocket, with the aid of the glycine rich nucleotide binding loop (also referred to as the P-loop). In the inactive conformation the activation loop sits more or less on top of the catalytic loop (or C-loop), which generally prevents nucleotide binding, though this is not the case in ALK which has been shown to bind ADP in an inactive conformation [146]. At any rate, correct substrate binding is prevented in the inactive conformation [109]. The ALK inactive conformation in Figure 1.1 displays a short α helix in the activation loop. This is not the case in all inactive kinase structures (e.g. in some other members of the insulin receptor kinase family, of which ALK is a member [12]), but is a common feature of inactive conformations (e.g. EGFR [271], SRC, and CDK [109]). Another common feature of the active conformation is phosphorylation of residues in the activation loop. In ALK there are 3 activation loop tyrosines that can be phosphorylated, though evidence supports the claim that only one of these residues is required for activation [61] while in EGFR there is only one tyrosine in the activation loop and it has been shown to be dispensable for full EGFR activation [267]. As a final example, BRAF, a serine/threonine kinase, has an activation loop serine which must be phosphorylated for full activation and an activation loop threonine which is dispensable for activation [269]. Without burying the lede in a flurry of examples, kinases have subtle differences in activation loop in the inactive conformation which are important for specificity and regulation while they generally have similar activation loop structures in the active conformation, generally accompanied by phosphorylation of activation loop residues [109, 147].

1.3.2 KE salt bridge

For a kinase to be active, it must bind ATP. Attendant upon this binding is the formation of a conserved salt bridge between a glutamate in the αC helix and an N-lobe lysine (KE salt bridge) which is requisite for ATP binding. In the inactive conformation, the KE salt bridge is generally not formed. In the absence of this non-covalent bond, the N-lobe is free to separate from the C-lobe, disrupting the ability to form a catalytically competent conformation. While a conformation can be characterized as active despite the lack of the KE salt bridge or inactive despite its presence, catalysis is though to occur only when this conserved salt bridge is formed [109, 136].

1.3.3 αC-in vs αC-out

Another feature that has been used to distinguish active from inactive conformations is the rotation and placement of the αC helix. The active conformation is commonly denoted αC-in due to the above discussed KE salt bridge constraining the orientation of the αC helix. In the αC-out conformation the αC helix rotates by several degrees away from the the the C-lobe, accompanied by separation of N- and C-lobes [109]. Interestingly, recent hydrogen/deuterium exchange and molecular dynamics experiments have shown that some kinases may have an intermediate state with a partially disordered αC helix [255, 256, 223]. The αC-in conformation is shown in Figure 1.2 (a) and (c), which are active and inactive respectively, demonstrating that this feature by itself does not tell the whole story of kinase activation. The αC-out conformation is shown in Figure 1.2 (b), which is an inactive conformation.

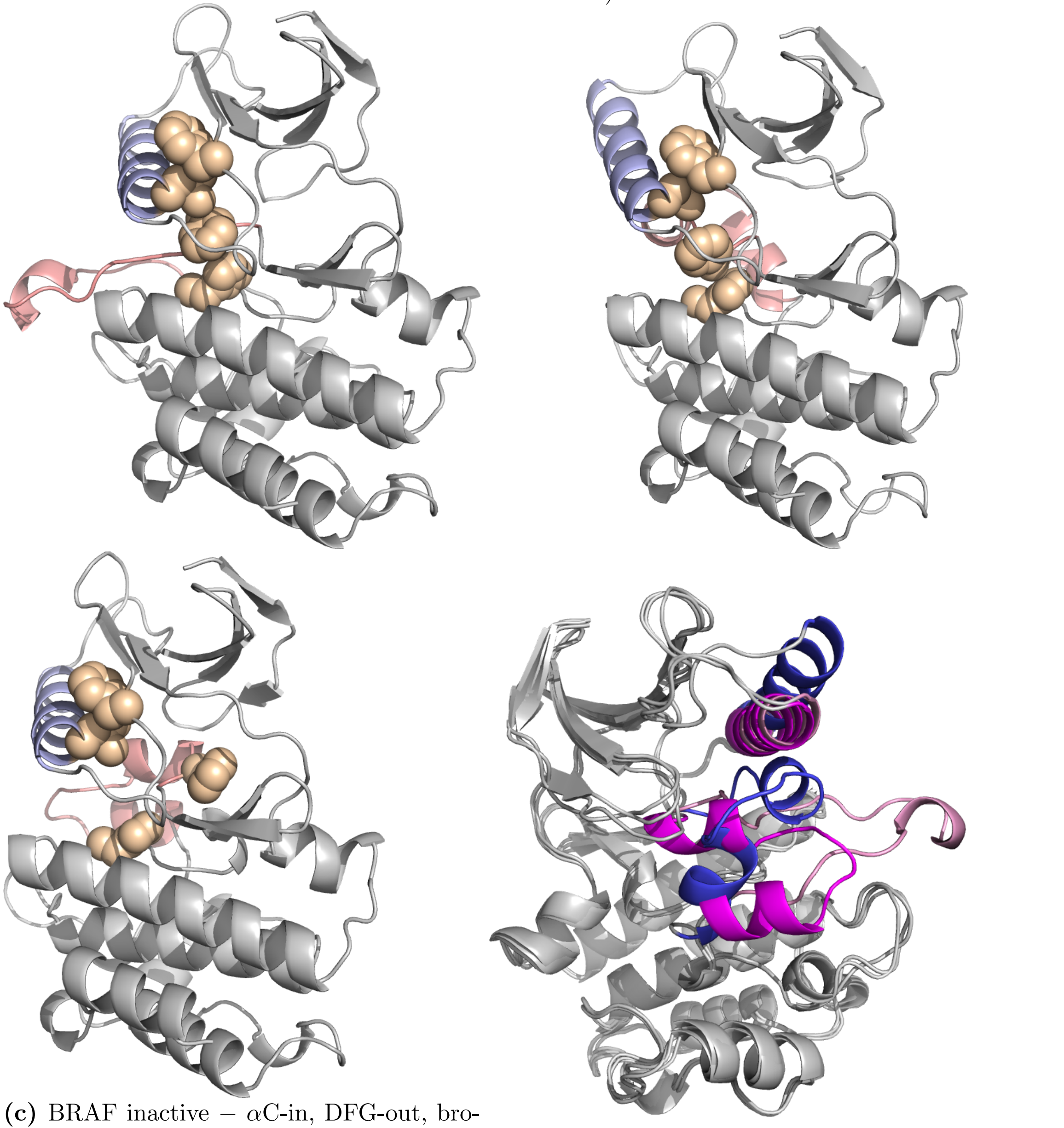
1.3.4 DFG-in vs DFG-out

At the start of the activation loop there is a conserved Asp-Phe-Gly (DFG) motif which plays an important role in kinase activation. Extensive studies utilizing both crystallization and MD experiments have shown the importance of the orientation of the DFG motif [109, 146, 250, 148, 157, 251]. In the active conformation, the Asp points towards the active site of the kinase, which is known as the DFG-in conformation for this reason. In the DFG-in conformation the Asp can coordinate Mg^{++} which are important for ATP catalysis and binding [71]. In the DFG-out inactive conformation the Asp points away from the active site, preventing coordination of Mg^{++}. There is also a DFG-in inactive conformation [251] which is observed in crystal structures of kinases much more frequently than DFG-out inactive. Another feature of the DFG-out conformation is that the Phe occupies the active site, abrogating ATP binding. Several computational studies have also shown that Asp protonation promotes the DFG-out conformation [225, 152, 164]. A DFG-in conformation is shown in Figure 1.2 (a) and (b) which are active and inactive, while a DFG-out inactive conformation is shown in Figure 1.2 (c).

1.3.5 R-spine

Another motif often discussed in the literature on the regulation of kinase domain activation is the regulatory spine (R-spine) [135, 239, 103]. The partisans of the R-spine have gone so far as to claim that αC-in vs αC-out and DFG-in vs DFG-out “is irrelevant; it is the

Figure 1.2: Nuances among BRAF structures



(a) BRAF active – αC-in, DFG-in, formed R-spine, KE salt bridge not formed (4.7 Å)

(b) BRAF inactive – αC-out, DFG-in, formed R-spine, KE salt bridge formed (2.8 Å)

(c) BRAF inactive – αC-in, DFG-out, broken R-spine, KE salt bridge formed (3.1 Å)

(d) Alignment of all three structures

BRAF structures were completed (any missing residues were added) with BioPhysCode utilizing Modeller as outlined in chapter 4. A was modeled off pdb 4MNE, B was based on 3TV4, and C was modeled from 1UWH.

assembly of the R-spine in its entirety that must be considered" [239]. Skepticism of conceptual apparatuses such as αC helix and DFG orientation is generally warranted, but such unequivocal claims are themselves guilty of the sort of reification they set out to critique. At any rate, the formation and dissolution R-spine seems to be a common theme in kinase activation. The R-spine is composed of 4 residues, generally hydrophobic, which come into alignment in the active conformation but are not seen to form a spine in some inactive conformations. These 4 residues are as follows: the His residue from the catalytic HRD motif (the Asp in HRD is the catalytic residue), the Phe from the DFG motif, an aliphatic residue in the αC helix, and an aliphatic residue from β strand 4 [103]. Workers have demonstrated that mutations that promote R-spine formation in BRAF can cause constitutive kinase activation [103]. When a kinase is in the DFG-out inactive conformation, the Phe points towards the active site, thus disrupting R-spine formation; this is shown in Figure 1.2 (c) while an inactive conformation with the R-spine formed is shown in Figure 1.2 (b). Concomitant with the transition from DFG-out to DFG-in, and thus R-spine formation, is the formation of the KE salt bridge [135], though an active conformation with R-spine but not KE salt bridge formation is shown in Figure 1.2 (a). Thus it does appear that a concerted movement of a number of components is important in the activation pathway of kinases, but this does not require us to set up hierarchies of importance surrounding which reified components *must be considered*.

1.4 Philosophical Introduction

The question of the existence and importance and existence of the R-spine gives us a focal point for analyzing some questions about reification and the scientific method. When attempting to model any process a scientist may be confronted with a number of technical challenges, but these are logically secondary to the question of *what to model* and *how to model it*. The answer to these questions is generally provided by convention in the field a scientist works in [139].

For example, an economist may be concerned with the workings of a single firm [47], an industry or nation, or even the total economy of the planet [23], but generally not with the entire balance of energy on the planet [16]. Our hypothetical economist must not choose a model that is overly complex for the problem at hand; quantum mechanics won't give any useful insight into the functioning of the economy. Further, our economist may be shocked to learn that their object of study (if it is the economy as quantified by GDP) is a social construct, which was not even conceived as an object for analysis before the late 1930's [167]. Finally, the economist, like any social scientist, must confront the troubling situation that any explanations that they provide for how an economy functions may themselves impact the functioning of the economy [153].

Fortunately, we are in a slightly less precarious situation. The motions of proteins, and the equilibria that govern their conformations, are governed by the rigorous laws of statistical mechanics [56], which tells us how to estimate the behavior of large numbers of atoms. Unlike our unlucky friend the economist, we do not find ourselves in a situation where some of the atoms in the system we hope to apply statistical mechanics to *are* statistical mechanics; our understanding of how we think the system works should not affect the workings of the system.

Even this situation can be seen as rather precarious though, for we have a number of choices to make. First, we must choose what part of the problem we are interested in investigating and which we want to discard in the hope that they are not relevant or important. For example, see chapter 2 for discussion of the ambiguity around studying different isoforms of ABL that are observed in the clinic. Second, we must chose how to analyze our results. For example, we may think that the R-spine plays a critical role in kinase activation and prioritize methods of analysis that can highlight this importance. There is nothing wrong with conducting scientific experiments in such a way as to make them easier to perform and interpret, but we must always recognize that this reductionism may exact the heavy price that we are not looking for our keys where we lost them, but instead only looking for our keys underneath the nearest street lamp. Even leaving aside the question of how much of the system of interest to model or how to interpret results, there are a large number of experimental conditions that must be carefully selected such as ion concentrations or incubation/simulation time (important *in vitro* and *in silico*), which cell line to use for *in vitro* experiments, and how to model electrostatic interactions in *in silico* experiments.

In fact, no *a priori* information can ever tell us which experiments to perform, how to set up these experiments, what analyses to perform, or even whether to perform experiments at all. In a landmark work in the history and philosophy of science, Feyerabend [68] outlines the role Galileo played in transforming science from its previous, scholastic, method of argument from authority, to a new method of argument from experiment. Remarkably, only by repeated appeals to authority, his own as well as the authority of the newly invented telescope (which in some cases performed demonstrably worse than the naked eye), was Galileo able to begin this transformation. Not long after this, the chemist Robert Boyle, eponymous of the gas law, engaged in a heated debate with Thomas Hobbes, author of "The Leviathan" and *"De Cive"* but also an aspiring chemist, on the role of the witness in an experiment [226]. Boyle thought the best way to get people to believe in scientific results was to have large numbers of witnesses assent to the truth of experimental results. This proved to be a challenge since the experimental apparatuses he used precluded all but the richest scientists from replicating his experiments on their own, and also since the difficulty of performing the experiments made it unlikely that more than a few people would ever witness the results of an experiment. Hobbes, raised a number of important questions about Boyle's empiricism. Put into modern terms these questions can be formulated as follows: (1) If new instruments such as the air pump show that the senses do not reveal everything we could learn about nature, what reason do we have to trust the instruments over the senses? (2) If some groups of people, like Oxford professors, are more suited to act as witnesses than others, like Oxfordshire peasants, then doesn't this say that there is a subjective component to witnessing? and (3) How do we know the boundary between facts and theories (or explanations of facts) if we wish to define a boundary between them at all? It would be a mistake to assume that these problems have been solved today, or that they could in principle be solved. All three of these questions can be seen revolve around a central axis, that of the relations of power in society to determine what questions can be asked and how they can be investigated. For an interesting appraisal of how power affects question (1), see Bertold Brecht's play "Life of Galileo" where the power of the Catholic church not only prevents assent to Galileo's astronomical theories, but even to learned cardinals looking through a telescope. For insight into the role of power in question (2), see "Merchants of Doubt" by Naomi Oreskes and Erik Conway for a detailed description of how well funded and well connected scientists worked for fossil fuel and tobacco companies to sow public doubt about climate change and the dangers of smoking. The role of power in question (3) is ably considered in "A History of the Modern Fact" where Mary Poovey outlines how the modern discipline of statistics has its roots in attempts by various people to be able to make a *unique* claim about how resources should be used based on supposedly values-free measurement and calculation. Poovey's exploration actually only covers the period from the late 16th century to the middle of the 19th century. If she had gone another century further, she could have also documented the role of power in question (3) by noting that both that the early statisticians Irving Fisher [70] and Karl Pearson [189] were interested in statistics and biology to further (what they viewed as) the science of eugenics. Both also sought to outline purportedly scientific justifications for colonialism by attempting to demonstrate that their racist understanding of the world was backed up by a 'neutral' body of facts.

Thus, we must always ask ourselves, as scientists, as citizens, and as human beings, whether we are in fact being 'objective' since what is deemed objective will always be the result of both societal and scientific convention. In the final analysis, we must take heed not only of the potentially distorting impacts of power on scientific discourse, but also on which interests we chose to pursue. As Walter Benjamin (quoted in Agamben [5]) wrote,

> The mastery of nature (so the imperialists teach) is the sense of all technology. But who would trust a cane wielder who proclaimed the mastery of children by adults to be the sense of education? Is not education, above all, the indispensable ordering of the relationship between generations and therefore mastery (if we are to use this term) of that relationship and not of children? And likewise technology is the mastery not of nature but mastery of the relation between nature and humanity.

We must be cognizant of what precisely it is we would like to order our understanding of. In this thesis, we will be attempting to order our understanding of the relationship between mutations which have primarily been observed in cancer patients and the dynamics or chemistry of the kinase domains which harbor these mutations. The related questions of whether this understanding can be translated into clinical practice or used to develop new treatments is left to future researchers, not only because the author has reservations about the distribution of resources and agency this implies [114], but more proximately because life is short.

Chapter 2

Mutations

> Then David said to God, "I have sinned greatly by taking this census. Please forgive my guilt for doing this foolish thing."
>
> I Chronicles 21

2.1 Introduction

2.1.1 Analysis of COSMIC

The ongoing decrease in price of genome and exome sequencing has lead to the creation of online databases for cancer genome sequence information such as the catalog of somatic mutations in cancer, COSMIC [72] and the cancer genome atlas, TCGA[1]. This in turn has driven a longstanding effort to catalog which proteins are frequently mutated in various cancer types [83, 73], as well as efforts to determine the driver status of mutations [84, 100] and to understand tumor heterogeneity through clonal evolution [178, 257]. A more thorough overview of the field of computational classification of cancer mutations will be given in chapter 5. In this chapter I will give some insight into what can be learned about kinase mutation by comparing the distribution of kinase mutations in COSMIC with what is known about the activation status of frequently mutated residues. I will provide a brief analysis of the most frequently mutated genes, residues, and pfam domains, concluding with a look at the distribution of mutations within the kinase domain. This will lead onto a detailed discussion of what is known about the activation status of a number of kinase domain mutations and conclude with a look at commonalities of mutant kinase activity.

2.2 Results

2.2.1 Mutations in all proteins

We proceed first by analyzing the total distribution of non-silent mutations in the COSMIC v81 whole genome and exome mutation database [72], which contains 29,805 whole genomes and millions of non-silent mutations. In this dataset there are 4 genes that are mutated

[1] cancergenome.nih.gov/

more than 10,000 times, shown in Table 2.1 and 441 genes that are mutated more than 1,000 times. There are over 4.1 million amino acid substitution mutations which result in 20,498 unique substitutions, including truncation mutations. The most frequently mutated protein is titin, which is over 38 thousand amino acids long, and thus probabilistically likely to be frequently mutated. It is unlikely that titin plays a role in cancer given that it is involved in muscle function and not cell growth or proliferation [39]. The next most frequently mutated protein, p53, regulates progression through the cell cycle and apoptosis and many mutations are known to be important in cancer progression for a number of tumor types [192]. Mucin 16 overexpression is used as a biomarker for ovarian cancer but mutations are not known to have effects on mucin 16 function [91]. Workers have shown that loss or downregulation of nesprin 1 can lead to misregulation of the DNA damage response, but to date no biochemical evidence has shown that individual mutations have a role in cancer [236]. Given the fact that three of the four most frequently mutated proteins are not known to drive cancer progression, and that these three also are very large proteins, there is good evidence that the frequency of mutations in a protein as a whole is not necessarily informative and is not pursued further.

Table 2.1: COSMIC frequently mutated proteins

Protein	Mutations	# amino acids	Role in cell
Titin	44,941	38,148	muscle contraction
p53	29,839	393	cell cycle checkpoint
Mucin 16	16,188	14,507	acts as mucosal barrier in epithelial cells
Nesprin 1	12,859	8,749	connects cytoskeleton to nuclear lamina

2.2.2 Frequently mutated residues

Shown in Table 2.2 are all proteins that have either more than one residue mutated more than 100 times or at least one residue with more than 500 mutations. Of 78 mutations seen over 100 times, 56 are found in the gene TP53 (the protein is called p53), which is also the second most mutated protein after titin. One striking thing about the list of frequently mutated residues is that, with the exception of p53 mutations which is mostly impacted by loss of function (LOF) mutations, most of the rest of the frequent mutations are gain of function mutations (GOF). Even in p53, which has had all 2,314 possible single nucleotide polymorphism possible tested, around 25% of mutations are GOF, though most of the frequently observed cancer mutations are LOF mutations [192]. Mutations in the GTPases KRAS and NRAS are frequently observed and these frequent mutations have been shown to be GOF, primarily through altering GTP catalysis and exchange rates [8, 218, 231]. The phosphotidylinositol kinase PIK3CA is important for regulating the phospholipid makeup of the cell membrane and several of the frequently occurring PIK3CA mutations, including two kinase domain mutations at residue H1047, have been shown to lead to increased kinase activity [113]. Isocitrate dehydrogenase 1 (IDH1) is involved in the citric acid cycle and thus a key node in cellular metabolism. The frequently observed mutation R132H alters the catalytic site of this enzyme to change the major product of IDH1 from α-ketoglutarate to 2-hydroxyglutarate, leading in turn to an upregulation of the protein hypoxia inducible factor 1α, promoting tumor cell growth [270]. The frequent mutation BRAF V600E causes constitutive kinase activity and is a well studied cancer driver mutation [254]. The most curious case is that of the transmembrane serine protease TMPRSS13, which does not have a known role in cancer but is involved in proteolytic activation of hepatocyte growth factor [92], one of the hallmarks of cancer [89]. Thus there is a rather large body of evidence that correlates frequently observed mutations to GOF alterations in protein activity.

Table 2.2: COSMIC frequently mutated residues

Protein	Residues mutated > 100 times	Residues mutated > 250 times
BRAF	1	1
IDH1	1	1
KRAS	7	5
NRAS	2	0
PIK3CA	6	3
TMPRSS13	2	0
p53	52	15
Total	**78**	**25**

2.2.3 Frequently mutated PFAM domains

We proceed next by analyzing the total distribution of non-silent mutations in the COSMIC v81 whole genome and exome mutation database [72] which alter a PFAM domain [69]. Briefly, PFAM domains are constructed by performing sequence alignment of a set of nonredundant representative proteomes to determine conserved protein motifs across individual proteomes and across proteomes in many species. PFAM domains, or architectures, have a conserved structure and have even been used to find conserved interfaces [263], thus demonstrating the relation between PFAM domains and protein structure and function. Data in Table 2.3 shows the percentage of mutation in a PFAM domain relative to the total mutations observed in all PFAM domains (Domain mutation %) as well as the fraction of the exome composed of a given PFAM domain (Domain exome %). When taken together, kinase domains (shown in red in Table 2.3) account for a higher percent of mutated domains than any other domain. When combining this knowledge with the fact that 2 of the 10 non p53 mutations in the Table 2.2 list of residues mutated over 250 times are kinases we have a strong case that a more detailed study of the pattern of kinase domain mutations could yield important insights. For this reason, we embarked on an analysis of the pattern of mutations in kinase domains in this dataset.

2.2.4 Kinase subdomain mutation distribution

In order to obtain information on subdomain clustering of kinase domain mutations, the COSMIC whole genome/exome screen (WGES) database (version 81) [72] was used as the source for mutational data. A multiple sequence alignment was performed using ClustalW2 [145], and the residues comprising functionally important subdomains were extracted. These subdomains are as outlined in chapter 1 and include the nucleotide binding loop (p-loop), the catalytic loop (c-loop), the αC helix, and the activation loop (a-loop). By binning the clinically observed cancer mutations in kinase domains in this manner, we can observe

Table 2.3: COSMIC mutations mapped to PFAM domains

Protein domain	Domain mutation %†	Domain exome %‡
Cadherin domain	1.77	1.14
Zinc-finger double domain	1.67	2.24
Class I Histocompatibility antigen	1.63	0.28
Immunoglobulin I-set domain	1.46	1.21
Protein kinase domain	1.24	1.92
Zinc-binding dehydrogenase	1.18	0.05
7 transmembrane receptor (rhodopsin family)	1.18	1.59
Seven additional domains	**5.85**	**-**
Protein tyrosine kinase	0.64	0.96

$\dagger \frac{\text{Total \# mutations observed in all residues in this PFAM domain}}{\text{Total \# residues in this PFAM domain}}$

$\ddagger \frac{\text{Total \# residues in this PFAM domain}}{\text{Total \# residues in exome}}$

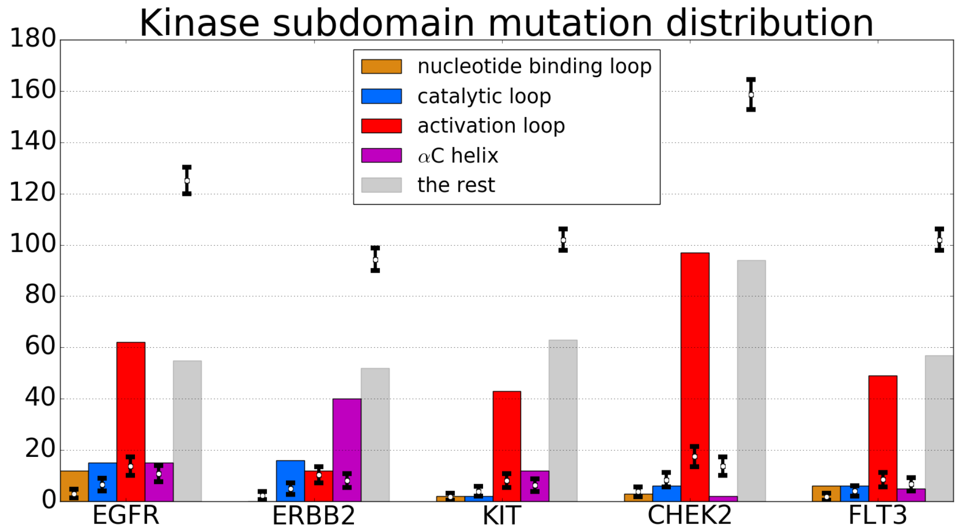
whether or not mutations preferentially segregate to any of these subdomains or whether there is a more uniform distribution of mutations across the kinase domain. This is shown in Figure 2.1. There are 22 proteins that have a kinase domain with more than 100 mutations in COSMIC WGES v81 and only one of these shows a mutational distribution where each subdomain is within one standard deviation of an expected uniform distribution. For this study, an expected uniform mutation distribution was calculated using a multinomial distribution, which can be thought of as akin to making n draws, with replacement, from an urn containing balls of k different colors, with the probability of drawing a ball of color k, p_k, proportional to the number of residues in a subdomain relative to the total number of residues. The number of draws n is in this case the total number of observed mutations. These distributions are shown in Figure 2.1 as black errorbars with a white circle as the mean and the bar length proportional to one standard deviation. Only proteins with particularly sharp deviation from uniform mutation distribution are shown in Figure 2.1, but almost all frequently mutated kinase domains show a non-uniform mutational distribution. Also notable in Figure 2.1 is the fact that the activation loop shows the largest deviation from a uniform mutation distribution. This is even more in evidence in the case of BRAF (not shown), where almost 1,000 of the roughly 1100 observed mutations are in the activation loop. This is largely due to the frequently observed BRAF V600E mutation discussed in frequently mutated residues.

2.3 Discussion

2.3.1 Description of Kinase Mutations

As noted in the sections frequently mutated residues and kinase subdomain mutation distribution one of the most frequently observed single mutation in COSMIC is the BRAF V600E mutation, which is located in the activation loop of the BRAF kinase domain and has been shown to lead to constitutive activation [254, 112]. In this kinase, several orders of magnitude more mutations are observed in the activation loop than the entire rest of the kinase domain. Almost all of these mutations are the V600E mutation. This mutation is

Figure 2.1: Several kinases with large number of mutations



Errorbars represent expected mutation distribution based on a uniform (multinomial) distribution.

frequently observed in several cancer types, including thyroid, colon, skin, leukemia, and lung cancers. While V600E is by far the most common BRAF mutation, there are a number of BRAF mutations that have been studied to determine their activation status. Most of these mutations occur in the activation loop or the nucleotide binding loop [254, 112, 111], though there are a few outside of these subdomains. These mutations are instructive of the complexity of kinase based signaling networks. A number of mutations to the BRAF active site result in kinase dead BRAF that nonetheless drives activation of downstream partners: D594A/V, the aspartate of the DFG motif; G596R, the glycine in DFG; as well as K483M, the catalytic lysine [254, 94]. It has been demonstrated that this effect is the result of increased dimerization and can also be seen to result from drug treatment of cells with WT BRAF, but not not constitutively active BRAF [94]. A detailed listing of BRAF, as well as other kinase, mutations that have been characterized is in Table A.1. Examining the large number of BRAF mutations that have been characterized provides an object lesson in the difficulties of model selection outlined in the chapter 1. Namely, that how we perform an experiment will overdetermine the result (not determine; it is a dialectic and not strictly a causal process). In [112, 111] many of the same BRAF mutations as in [254] are measured. However, in [112, 111] a subphysiological [ATP] is used, whereas in [254] a putatively more biologically relevant concentration is used. This casts some doubt on the measurements in [112, 111] since BRAF WT is autoinhbited at physiological [ATP], but not at subphysiological [ATP], by phosphorylation of S465 and S467 in the nucleotide binding loop [102]. Also of note is the fact that several BRAF studies compare the kinase activity to BRAF incubated in cells with mutant HRAS and find that under these conditions BRAF activity is several fold higher in many mutations [269, 112, 111], with one study citing an 11.5 fold increase [112] and another a 6 fold increase [269] in BRAF activation under these conditions. This is a proxy for measuring the activity of unphosphorylated versus phosphorylated protein. Studies in ALK have shown that the activity of unphosphorylated protein correlates well with transformation capacity in cells while phosphorylated protein activity does not [24]. For this reason, activity in mutant HRAS cells are not reported here.

Two proteins in the ErbB receptor tyrosine kinase (RTK) family are seen on the list of frequently mutated kinase proteins, and many mutations in both have been experimentally characterized. In the epidermal growth factor receptor (EGFR), the most common mutation is L858R, sometimes called L834R due to the presence of a cleaved signal peptide. This mutation is in the activation loop and is known to cause constitutive activation [267, 268, 266]. Even though this mutation accounts for the plurality of the mutations that are listed in COSMIC WGES v82 for EGFR, there are many other mutated residues. Most of these mutations are of unknown consequence, but several have been shown to lead to constitutive activation Table A.1. Other frequently observed mutations are the activation loop L861Q mutation, the constitutively active nucleotide binding loop G719S mutation [267], and the so-called gate-keeper mutation T790M, which is not located in any of the kinase subdomains under consideration but does lead to constitutive activation and decreases sensitivity to EGFR inhibitors [268]. Curiously, although some studies have shown that L861Q increase kinase activity, transforms cell lines, and promote drug resistance [36, 41, 129], no study to date has made a direct measurement of L861Q kinase activity. The EGFR mutations E709G and S768I, which are often seen mutated in targeted EGFR mutation screens have also been shown to lead to increased activity and resistance to targeted inhibition [41] but have not undergone direct kinase activity measurements.

The ErbB family member HER2, also known as ErbB2, is frequently mutated in breast cancer and also often mutated in colorectal cancer as well. The most frequent mutations are the catalytic loop V842I and the αC helix I767M, D769Y, and V777L mutations. All of these mutations have been studied *in vitro* and *in vivo* to determine their activity [21]. The L755S mutation, which is outside of any subdomain and is N-terminal to the αC helix, has been shown to have minimal cell transforming abilities but to confer drug resistance, and has as yet resisted attempts to characterize its catalytic activity [21, 131, 273].

Some insights can be gained by comparing the effects of mutations in these two ErbB family members. The most striking difference in the mutation patterns of these two proteins can be easily seen in Figure 2.1, which shows that EGFR is primarily mutated in the activation loop while HER2 is mostly mutated in the αC helix. A few studies have also sought to directly compare mutations in EGFR and HER2. One such study looked at the HER2 L866M mutation, which is equivalent to the EGFR L858R mutation. While both lead to the ability to transform cell lines, HER2 L866M only leads to moderate increase in activity [66]. It is also of note that the BRAF activating mutation L597V [254] also aligns to the same location as EGFR L858R and HER2 L866M, indicating that this site, which follows immediately after the DFG motif, is conserved to maintain a precise hydrophobic character and that minor steric alterations can lead to large changes in activity. As discussed in chapter 1, there is an important salt bridge (the **KE** salt bridge) that is characteristic of the active kinase conformation and generally requisite for binding ATP. The Lys of the **KE** is L745 in EGFR and K753 in HER2 and in both cases mutation to Met leads to loss of kinase activity, but in the HER2 this mutation is still transforming due to increased EGFR phosphorylation. HER2 K753M also confers resistance to some targeted therapies [273].

Another kinase that is frequently mutated in many cancers is the stem cell growth factor receptor known as c-kit. Again the majority of the mutations in this protein are observed in the activation loop, primarily consisting of the D816V or D816Y mutation, both of which are known to lead to constitutive activation [75, 235, 6]. The case of D816V (and of the less frequently observed D816H) highlights the need for detailed mechanistic analysis, as D816V does not differ significantly from WT in terms of K_{cat} but is instead activated by a relatively more rapid bimolecular autoactivation [75]. Another common activation loop mutation is N822K, which has been shown to have weak transforming potential in cells [179] but has not been investigated for changes in catalytic activity. Mutations are also observed in the αC helix of c-kit or just KIT, with K642E being the most common. This mutation has been demonstrated to have the potential to transform cells [116, 169] but again has not been investigated for changes in catalytic activity. There is also one relatively frequent mutation in kit that is not in a subdomain, namely the V654A mutation that falls in between the αC helix and the catalytic loop and leads to increased kinase activity [75].

An interesting case is the protein CHEK2, which positively regulates the DNA damage response by phosphorylating numerous proteins in response to DNA damage. This means that CHEK2 cancer driver mutations would actually be inactivating mutations instead of mutations that lead to constitutive activity, and this principle is demonstrated by the fact that CHEK2 knockout mice are more prone to develop cancer from ionizing radiation than CHEK2 positive mice [15]. This makes that fact that mutations in CHEK2 cluster in the activation loop, as seen in Figure 2.1, interesting as it demonstrates that the same principle that can lead to increased activity can lead to loss of activity. The only frequently observed mutation that has been catalytically characterized is the K373E mutation, which leads to a decrease in kinase activity comparable to the D368N mutation, which is the Asp of the DFG motif and thus catalytically important [99]. The frequently observed Y390C mutation has been shown in cells to be unable to activate p53 in response to DNA damage [258] but has not been catalytically characterized.

Almost all of the activation loop mutations observed in FLT3, frequently mutated in leukemia, occur at D835, with the rest of the observed mutations scattered throughout the rest of the kinase domain. While many of these mutations have been shown to be transforming in cells [265], only the most common, D835Y and D835H, have been shown to result in increased kinase catalytic activity [42].

Perhaps the most remarkable case of constitutive activation of a kinase by a mutation is that of JAK2. JAK2 has two kinase domains called JH1 and JH2, but the second kinase domain, JH2, has only weak kinase activity and is thought to act to regulate the activity of the JH1 kinase domain [247]. The V617F mutation is on the list of residues mutated more than 100 times discussed in section on frequently mutated residues. This mutation occurs in a loop in the N-lobe of the JH2 kinase domain, which lacks secondary structure and leads to constitutive JAK2 activation [14]. The mechanism that causes this mutation to be activating is still a subject of debate but results from microsecond molecular dynamics (MD) simulations show that the V617F mutation may stabilize the αC helix of JAK2 JH1 [14]. This would be an interesting corollary to the mechanism of the EGFR L858R mutation which MD and hydrogen/deuterium exchange (H/DX) have indicated may promote constitutive activation by stabilizing the helical character of the EGFR αC helix, allowing access to an active conformation [255, 256, 223, 237] (see also chapter 1).

There are a number of clinically observed kinase domain mutations that have not been frequently observed in WGES but have been frequently seen in targeted cancer mutation screens. Particularly well studied in this category are the members of the fibroblast growth factor receptor (FGFR) family. As seen in Table A.1 there are a number of mutations of FGFR2 and FGFR3 that have been demonstrated to lead to constitutive kinase activity [272, 40, 62, 186] The most common FGFR2 mutations are various mutations to N549 which is just C-terminal to the αC helix and K659E in the activation loop. In FGFR3 the most common mutations are to K650 in the activation loop (in the same location as FGFR2 K659) and G697C which is located in the C-lobe of the kinase domain. While the K650E mutation in FGFR3 has been demonstrated to activate FGFR3 even in the absence of activation loop phosphorylation, the mechanism of G697C activation is uncertain as this residue is located on an unstructured loop distal from the active site [106] (also of note is that different studies have measured widely varying catalytic activities for the G679C mutation [106, 186]). Another protein that has an activating mutation outside of a kinase subdomain is RET. The protein RET is frequently mutated in thyroid cancer at a M918T which is a few residues N-terminal to the activation loop. This mutation has been shown to lead to constitutive activation by decreasing the stability of closed activation loop conformation, leading to rapid phosphorylation of activation loop tyrosines relative to WT RET [197]. This mutation as again a call to exercise care in experimental design. One study found that RET M918T was not activating using a generic phosphorylation substrate [163] and another study found little difference between a construct composed of kinase domain or of kinase domain plus juxtamembrane (JM) [133]. However, a number of other studies have shown increased catalytic activity using more physiologically relevant substrates for kinase activity assays [198, 197, 196]. These studies also showed that the kinase activity of the isolated kinase domain was lower than that of a construct containing the kinase domain and the juxtamembrane segment [197, 196], in line with evidence from EGFR [206].

The protein ABL has been the subject of intense study owing to it being the first kinase for which a targeted inhibitor, imatinib, was developed [33]. The first identified genetic abnormality linked to cancer (leukemia), named the Philadelphia Chromosome, was eventually shown to the result a a chromosomal rearrangement involving the coding sequence of ABL and and a genomic loci call the breakpoint chromosome region (BCR), leading to the BCR-ABL fusion protein [177, 33]. An emerging consensus has it that by the time of diagnosis and treatment, clonal expansion of BCR-ABL WT containing cells has already resulted in a small population of BCR-ABL mutated cells, that these cells are selected for by targeted inhibition [232], and that further treatment with second and third line inhibitors also select for mutations present in the population at the outset of treatment [183, 184]. Most sequencing of patients who harbor BCR-ABL is done in a targeted manner rather than via WGES due to the need for sensitivity in detecting low frequency mutations [183, 184]. The most frequent clinical mutation in BCR-ABL is T315I, which is not in any kinase subdomain and is known as the gate-keeper mutation since it has been observed to inhibit drug binding in ABL and in many other kinases [13] The T315I mutation reduces the catalytic activity while increasing affinity for ATP in ABL and BCR-ABL [85], and the substitution of the larger and hydrophobic Ile for the smaller and hydrophilic Thr sterically hinders the binding of ATP competitive inhibitors, and also biases the system toward the active conformation by promoting hydrophobic spine formation [13, 207]. H/DX experiments have also shown that the T315I mutation leads to increased flexibility in the

αC helix relative to WT [110]. Two further ABL mutants are frequently observed in cancer and known to lead to resistance to targeted therapy while maintaining or increasing kinase activity, namely Y253H and E255K [166, 110]. Y253 is in the nucleotide binding loop and E255 is one residue C-terminal to the nucleotide binding loop, and mutations at these residues seem to have similar effect as mutations to the gate-keeper residue [85, 166, 110]. It is again worth sounding a note of caution related to how mutations are assessed as kinase catalytic activity has been studied under many conditions: including only the kinase domain (KD) [230, 85, 60], including the SH2 domain and the kinase domain (SH2-KD) [60], including the SH3 domain which preceded the SH2 domain [22], using full length ABL [85], for the effect of myristoylation [13, 110], and the fusion proteins p210 BCR-ABL [85, 166], and p185 BCR-ABL [230, 210]. It can be difficult to compare the activity of mutations from separate studies where catalytic activity is normalized to WT activity, and this problem is illustrated forcefully by one study [60] which seeks to compare KD and SH2-KD activity which for each plot normalizes KD activity to 1, resulting in 3 separate relative activities between KD and SH2-KD (see Table A.1), precluding any easy comparison between these two constructs. Finally, some studies [230, 22] only report K_M and V_{MAX} but not K_{cat} or [protein] so that catalytic efficiency cannot be assessed.

In PDGFRα the D842V activation loop mutation is frequently observed in targeted screens of patients with gastro-intestinal stromal tumors (GIST). This mutation, along with a few other PDGFRα activation loop mutations seen in GIST have been shown to lead to increase PDGFRα in cells, to lead to transformation in cell lines, and to confer resistance to targeted inhibitors [95, 96], but have not to date had their catalytic activity reported in the literature.

2.3.2 Similarities of kinase mutations

Interesting phenomena emerge when we look at the most prevalent mutations as a group rather than individually. Already noted was the effect of the gate-keeper mutation, which has been shown to confer resistance to targeted inhibition in ABL T315 [13], ALK L1196 [44, 24], BRAF T529 [94], CRAF T421 [94, 201], EGFR T790 [13, 268], ErbB2/HER2 T798 [129], KIT T670 [75], PDGFRα T674 [13], PDGFRβ T618 [13]. Also previously mentioned was the fact that EGFR L858R, HER2 L866M, and BRAF L597V are all activating mutations [66, 254]. This points to the interesting fact that several prevalent mutations are actually at the same residue in different proteins, as determined by sequence alignment. As previously discussed, EGFR L858R, ERBB2 L861M, and BRAF L597V are at the same location, just after the DFG motif which begins the activation loop, and are all constitutively active [66, 254]. Another activation loop mutation site that leads to constitutive activation maps to 4 residues past the DFG motif. The mutations BRAF V600D/E/F/K/R, EGFR L861Q, FLT3 D835H/Y, KIT D816V/Y, and PDGFRα D842V all lead to constitutive activation. It is of note that BRAF (see chapter 1 for a figure which shows this), EGFR [271], and KIT [144] all have a short activation loop helix that is formed in the inactive conformation and MD simulations in KIT have shown that KIT D816V may function by destabilizing this helix [144]. It has also been proposed that this activation loop residue position can alter the formation of the R-spine in BRAF [103]. In contrast to the above mutations which lead to constitutive activation, there are a number of residues that have been studied for their ability to prevent activation. One major thrust in this area has been to characterize the effects of mutations to phosphorylatable residues [143]. A large number of serine, threonine, and tyrosine residues have been mutated to alanine to prevent phosphorylation or glutamate or aspartate to mimic phosphorylation. A recent review of the literature found 145 such kinase domain mutations that have been experimentally characterized [143]. Another class of kinase mutations that have seen thorough investigation is mutations to the DFG motif aspartate. ALK D1270G [24], BRAF D594V [112, 254], CHEK2 D368N [99], DAPK3 D161A [82], and LIMK1 D460A/N [168, 63] have all been shown to abolish or reduce kinase activity. Finally, many studies include mutation of the lysine of the KE salt bridge as a negative control. AKT1 K179A [9], EGFR K745M [66], ERK1 K71A [143], HER2 K753M [66], NEK2 K37R [209], RET K758M [197], and RPS6KB1 K167N [93] all fall under this category and all lead to loss of kinase activity.

Chapter 3

Kinase MD

To dissimulate is to feign not to have what one has. To simulate is to feign to have what one hasn't.

Jean Baudrillard *Simulacra and Simulation*

3.1 Introduction

3.1.1 Kinase Activity

Kinase proteins play important roles in diverse cellular processes including signaling, differentiation, proliferation, and metabolism [147], are frequently mutated in cancer [89] and are the targets of a large and growing number of specific inhibitors [76]. These proteins have large differences between the conformation of the active and inactive conformations [109] and there is mounting evidence that drugs can specifically target one conformation over the other [182] and that, paradoxically, kinase inhibitors can activate wild type (WT) kinases [94]. While a large number of kinase domain mutations have been observed in patient tumor samples [72], more detailed analyses have shown that only some of these mutations can be classified as cancer driver mutations, which have a demonstrated impact on cancer progression, while many others will be passenger mutations that have no known effect on cancer progression as demonstrated by catalytic and colony formation assays (see chapter 2 for a detailed discussion on this topic).

Studies in anaplastic lymphoma kinase (ALK) have demonstrated the relationship between increases in isolated kinase domain catalytic efficiency (K_{cat}) and increased ability of transfected cells to form foci in colony growth assays [24]. Changes in substrate binding affinity (K_M) are less relevant here since the cellular ATP concentration is in the millimolar range, and many mutations actually increase affinity for ATP, leading to decreased efficacy of ATP competitive inhibitors [85, 66, 268, 75]. Tests of 23 ALK mutations observed in neuroblastoma patients showed a correlation coefficient of 0.95 between *in vitro* K_{cat} for non-phosphorylated ALK and colony transformation while analysis of phospho-ALK showed no significant correlation. It is also noteworthy that every mutation that increased K_{cat} by larger than 4.6 fold lead to colony transformation and variants with K_{cat} increase of 3 fold could show activation [24]. Analysis of a series of 22 mutations in BRAF, which is frequently mutated in melanoma and colorectal cancer showed that all tested mutants with a K_{cat} increase of more than 5 fold could transform cells in focus formation assays, though not all variants were tested in focus formation assays [254]. A group of studies on BRAF mutant catalytic activities with sub-physiological [ATP] showed that increases in K_{cat} of more than 3 fold lead to colony formation in focus formation assays [112, 111], though without K_M measurements it is difficult to compare the different measurements of BRAF kinase mutants. Studies in HER2, also known as ErbB2 and frequently mutated in breast cancer, have again shown that increases in K_{cat} of greater than 4 fold in monomeric assays lead to transformation in colony formation assays [21, 273], with mutations showing increased catalysis of as little as 2.2x in monomeric assays also demonstrating transformation potential [21]. Enforced dimerization of HER2 leads to increased activation relative to soluble monomers [21] and it has been demonstrated that HER2 mutation activation is not a result of altered interactions with HSP90 [66]. Studies in BCR-ABL mutants have also shown a good correlation between increased V$_{MAX}$ relative to WT p210 BCR-ABL and increased transformation, though this comparison is complicated by the fact that WT p210 BCR-ABL is itself moderately transforming [230].

3.1.2 Kinase Structure

For a detailed elucidation of the intricacies of kinase structure, please refer to the chapter 1.

3.1.3 Computational studies

3.1.3.1 Kinase MD Methods

Given the large number of kinase domain mutations observed in cancer and the laborious nature of performing kinase catalytic and colony assays, many workers have sought to develop computational approaches to understand the effects of mutations on kinase dynamics. One computational approach that gives unique insights into short time-scale dynamics is a molecular dynamics (MD) simulation. These simulations probe motions on the order of nanoseconds to microseconds while in kinases catalysis takes place on the scale of milliseconds to seconds [24, 254, 21], so careful analysis of a simulation trajectory is needed to gain insight into how mutations affect dynamics. These analyses can generally be fit into 3 broad categories: (1) analysis of alteration in chemical or physical quantities, (2) analysis of collective motions, and (3) computation of free energy landscapes. Under the first class are methods such as analysis of hydrogen bonds and salt bridges, changes in solvent accessible surface area, or analysis of hydration dynamics. In the second class are measurements such as root mean squared deviation (RMSD) or fluctuation (RMSF), and calculations based on interatomic covariance matrices such as protein structure networks (PSNs), or principal component analysis (PCA). The final class contains a large and growing number of methods for understanding the energetic relationship between different conformational states of a protein. These methods generally rely on some prior knowledge of different conformational states of a protein and apply some sort of energetic potential to help the system explore desired states.

3.1.3.2 MD Studies

Nanosecond MD Early kinase mutation MD studies were limited by computational power and were generally only tens of nanoseconds (ns) long but were still able to provide insight into the impact of kinase mutations on kinase dynamics. Advances in computer power quickly lead to simulations with hundreds of nanoseconds of simulation time. In [58], 10 ns simulations in both active and inactive conformation of EGFR WT, T790M and L858R, as well as ABL WT, T315I, and L387M were performed. On this timescale no difference in RMSD was detected but in inactive conformation simulations increased fluctuations were seen in the activation loop. Targeted MD and MM-GBSA calculations showed a decrease in stability of the inactive and increase in stability of the active conformation. In [74], 15 ns simulations of BRAF WT, D594V, V600E, and K601E did not show any differences in RMSD and RMSF on this timescale but hydrogen bond networks were altered. An early study to investigate the effects of mutation on kinase dimers was performed by Shih *et al.* [229]. Here 10 ns simulations on WT monomers of EGFR, Her2, and HER4 in active and inactive conformation as well as inactive dimers of these three systems plus EGFR L858R, and the common EGFR deletion mutant del_L771_P777_ins_S (Δ). Both SASA and water density analysis, a measure of hydrophobicity, show that inactive conformations of the ErbB kinases are stabilized in dimers relative to monomers. EGFR L858R and Δ are shown to counteract this hydrophobic stabilization. Further, the network of hydrogen bonds that stabilize inactive ErbB conformations is shown to be disrupted by mutations. Using four 200 ns replicates each in both active and inactive conformation of EGFR WT and L858R [255] show that only one simulation of 16 displays substantial change in RMSD, and this localized to the activation loop which moves from an active towards and inactive conformation. Analysis of salt bridges shows a weakening of a salt bridge that stabilizes the inactive conformation in the L858R relative to wild type but little difference between active conformation salt bridge patterns. A combination of PCA and MM/PBSA analysis of shows that in L858R the active conformation is favored whereas in WT inactive is favored. The existence of an intermediate conformation between active and inactive in L858R was also observed. A study in phosphotidylinositol 3-kinase alpha (PI3Kα) using five 150 ns replicates of active WT and H1047R show changes in polar contacts lead to greater activation loop flexibility and more positive charge on the protein surface known to be involved in membrane binding. Surface plasmon resonance studies then confirmed that H1047R has greater affinity for negatively charged membrane [78]. Another MD analysis of EGFR mutants, T790M, L858R, and the double mutant T790M/L858R using 50 ns simulations showed that mutations alter collective motions as assessed by both PCA and RMSF [118]. An analysis of protein structure using force constant analysis (FC), a scaled form of Cα-Cα carbon distance correlation matrix, show results similar to PCA analysis. This is as would be expected by the fact that they both rely on transformations of a correlation matrix, in the case of FC using only Cα and PCA using all backbone heavy atoms. In another analysis by the same group, a number of kinases from diverse kinase families were simulated in WT active and inactive conformation for 20 ns and analyzed via PCA [59]. Mutations were then mapped onto conformational flexibility profiles to suggest that activating mutations cluster in more flexible regions while inactivating mutations cluster in more rigid regions. One limitation of this study is that only some of the cancer mutations considered have been functionally validated via studies of altered catalytic efficiency or colony transformation ability, calling in to question the generality of the identified trend. A pair of articles [141, 140] investigates the impact of mutations on the active conformation dynamics of Aurora kinase A (AurkA) with S155R simulated for 100 ns [141] and G325W simulated for 200 ns [140]. A number of methods show differences between mutant and WT dynamics including PCA, RMSD/RMSF, SASA, and radius of gyration. One limitation of all of these studies is that they only simulate mutations that are known to be activating. A pioneering paper to address this issue is the work of [24] which uses 40 ns simulations of a series of 23 ALK mutants which range from inactivating to strongly activating. A scoring function based on hydrogen bonding patterns, SASA analysis combined with free energy perturbation, and PCA was shown to reliably distinguish between activating and non-activating mutations as assessed by changes in catalytic efficiency.

Microsecond Conventional MD Supercomputers of various types have lead to the ability to simulate microseconds worth of kinase motions. These timescales are not generally long enough to observe major conformational changes in proteins as large as kinases. Although fluctuations as large as folding/unfolding events can be observed in peptides and small, fast folding proteins on this timescale at physiological temperature [149], even in 100 amino acid proteins it is necessary to simulate near the melting temperature of a protein to observe folding events on the millisecond timescale [194, 195], allowing a delineation of a full conformational landscape. Nonetheless, longer simulation timescales have been important both as a validation of insights from shorter timescales and also because they provide a test of stability of various conformational (sub)states. An early use of long conventional MD to understand the effects of mutations on kinase domain dynamics was carried out to understand how mutations in the second kinase domain (denoted JH2) of JAK2, which is kinase dead, leads to constitutive activation of the first kinase domain (denoted JH1) [14]. Analysis of secondary structure of JH2 variants at this timescale showed that the helical character of the αC helix of WT JH2 is low but that the V617F mutation common in myeloproliferative neoplasms stabilized the helicity of the αC helix. This insight was in agreement with the crystallographic evidence in the same report. Finally, the JAK2 JH2 double mutant F595A/V617F was shown in phosphorylation assays to bring JAK2 activity down to levels similar to WT and shown in MD to destabilize the JH2 αC helix. One of the most extensive uses of long conventional MD to understand dynamic effects of kinase mutation was performed in [224]. Performing four replicates of 10 μs each on the active conformation of EGFR, it was seen that in all four simulations the αC helix goes from an 'in' conformation indicative of the active state to an 'out' conformation generally associated with the inactive conformation within 100 ns. At longer times, the WT αC helix was seen to adopt a disordered conformation, which was confirmed by hydrogen/deuterium exchange (H/DX) experiments. Even more intriguingly, in all four simulations the salt bridge between Lys-745 and Glu-752 (KE salt bridge) which is characteristic of the active conformation is broken within 40 ns. This provides good evidence that shorter MD simulations can still provide insight into the effects of kinase mutations. Long simulations of active conformation L858R show that the KE salt bridge is stable for between 2 and 30 μs, leading to a more stable αC helix in the L858R system. Simulations of G719S, S768I, L861Q, and Δ also show αC helix stabilization. Interestingly, HER2/EGFR and HER4/EGFR heterodimers were both shown to stabilize the αC in conformation in microsecond MD simulations, providing insight into the activation mechanism in non-mutant systems. Another study to investigate microsecond dynamics in EGFR is [256] where 2 replicates each of WT and L858R in the active conformation were run for 10 μs each. As in a previous study by the same group [255] and the work of *Shan et al.* [224], this work showed that WT EGFR has a disordered αC helix but that EGFR suppresses this disorder. Interestingly, the 2nd principal component corresponds to unwinding of the αC helix. An interesting application of microsecond MD is an investigation of the effects of protonation state on the active state dynamics of SRC kinase [71]. When the Asp of the DFG motif in SRC is unprotonated, no change from the αC-in active conformation or breaking of the salt bridge Lys290-Glu310 (similar to the EGFR KE salt bridge; conserved in active kinases) is seen in simulation of between 167 and 375 ns. In the protonated state, this salt bridge is broken within 100 ns in 4 of 8 simulations, with concomitant transition to αC-out and breaking of the R-spine, as measured by contact area of R-spine residues. Simulations of protonated D404 with ADP but no Mg^{++} show ADP leaving the binding site after 225 and 385 ns while with the addition of Mg^{++} lead to ADP remaining bound to SRC for 3.5 μs of simulation time. While not simulating any mutations, this work shows that changes in charge can greatly alter kinase and ligand binding dynamics, which should apply as well to mutations. An application of microsecond MD to understanding kinase mutations in the context of EGFR/HER3 heterodimers is [150] which investigates HER3 WT, Q790R, and S827I with 9 μs simulations in the active conformation. These simulations showed that in EGFR/HER3 WT a stable interaction between HER3 and the EGFR juxtamembrane (JM) region, known to be important for activation, is not formed while in both mutants there is a stable interaction between HER3 and EGFR JM. Since HER3 has a low intrinsic catalytic activity [228] and since these mutations activate EGFR moderately over WT HER3 in vesicle assays but strongly activate in solution assays [150], these simulations are in good agreement with experimental data.

Metadynamics Although there are a number of methods for determining either the free energy or conformational landscape of a biomolecular system, the one that has seen the most use in computational studies of kinase mutations is metadynamics. Without belaboring the details of this method, it allows an efficient exploration of conformational and energetic landscapes by adding energy to the system along a predefined reaction coordinate, termed a collective variable, providing a description of both the conformation and energy at each point along the reaction coordinate. These simulations generally require significant simulation time, hence the distinction long conventional MD as compared to free energy calculations. One of the first reports of the use of metadynamics on kinase mutations was an investigation of monomeric EGFR WT, T790M, L858R, and the double mutant T790M/L858R [237]. The conformational free energy landscapes reported here confirm the insight from short [255] and long [256, 224] conventional MD that WT EGFR has disordered αC helix as a distinct state in the inactive conformation, and also showed that a fully active conformation is not accessible to monomeric WT EGFR. Again in line with previous work, these workers showed that the L858R mutant suppresses the disordered αC helix while also allowing the system to visit a fully active conformation. Interestingly, the T790M mutant is shown to be able to access the disordered αC helix but to strongly prefer a fully active conformation while the Y790M/L858R double mutant is shown to preferentially visit semi- and fully active conformations. In none of the mutants is there a minimum corresponding to the WT inactive conformation. In work with important lessons for minimal model selection, work on ABL using 1 μs simulations showed that inclusion of the SH2 domain of ABL results in decreased nucleotide binding and αC helix flexibility, and increased activation loop flexibility, as shown by both RMSD and PCA analysis. This result was confirmed with metadynamics. One microsecond simulations of ABL kinase domain mutants M297G and E294P/V299P were shown to have similar flexibility with and without SH2 domain as part of the model. This was confirmed by kinase activity assays that show these mutations have similar activity whether or not the SH2 domain is part of the construct. A similar study combined two 500 ns simulations each of BRAF WT and V600E in active, DFG out (semi-active), and αC-out (inactive) with metadynamics [157]. On the timescale of the conventional MD simulations there are not significant differences between flexibility of WT and mutant, with the primary difference being a persistent salt bridge formed between Glu600 and Arg603. Metadynamics shows that BRAF V600E does not appreciably populate the inactive conformation but is mostly found in what the authors call a semi-active conformation; the Glu600-Arg603 salt bridge is common in this conformation. A metadynamics study of FGFR1 WT and V561M showed that the WT has a much smaller energy well in the active state than the mutant, even though both have similar activation barriers [28].

3.2 Methods

3.2.1 Simulations

Simulations and analysis were carried out using the BioPhysCode software suite[1]. The initial structures of ALK are as previously reported [24]. BRAF active was modeled off pdb 4MNE while BRAF inactive was based on 3TV4. HER2 active was modeled after 3PP0 chain B while HER2 inactive was constructed from a homology model based on EGFR structures 2GS7 chain A, 4HJO chain A, 3W32 chain A and ErbB4 structures 3BBW chain A, 2R4B chain A. All homology models were constructed using MODELLER [216] and all mutations were introduced using a BioPhysCode Automacs routine based on MODELLER. Simulations were run with Automacs using GROMACS 4.6 [181] with the CHARMM27 force field [154] with TIP3P explicit solvent [124] in a periodic water box with at least 12 Å between the protein and box edge. An ionic concentration of 0.15 M NaCl was used and the final charge of the full system was zero. Minimization was carried out using steepest descent and the system was eqilibrated first at constant volume, then at constant pressure using Berendsen [18] before production MD was carried out constant pressure using Parinello-Rahman [185]. All equlibration and production MD were carried out at constant temperature using velocity rescaling [31], using particle mesh Ewald electrostatics [65] with linear center of mass motion removal. LINCS [98] was used to constrain all bonds during equlibration and hydrogen bonds constrained during production MD. Simulations were run for a total of 101 ns and two replicates were performed for each simulation.

[1] github.com/biophyscode

3.2.2 Analysis

Analysis was performed, unless otherwise noted, on the last 100 ns and the two replicates were averaged together. Structures were sampled from each trajectory at 20 ps intervals, resulting in a total of 5001 structures for analysis. Plotting was performed with Omnicalc using matplotlib [108].

3.2.2.1 Hydrogen bonding

Each amino acid is considered to have a maximum of 3 possible hydrogen bonds (H-bonds), a main chain donor, a main chain acceptor, and the side chain, meaning that some residues such as Arg or Asp can have more than one side chain H-bond in a single frame; however, bonds are counted uniquely so that this could only happen if e.g. Arg-i and Asp-j side chains make both possible H-bonds. For each structure in a trajectory and for each H-bond the hydrogen bond occupancy ($\mathbf{O}$) was calculated by dividing the number of frames in which an H-bond is observed by the total number of frames. After computing the occupancy for each residue i in the inactive WT ($\mathbf{O_{WT,i}}$) and residue i in the inactive mutant ($\mathbf{O_{MUT,i}}$) the occupancy difference in mutation MUT for residue i ($\mathbf{\Delta_{MUT,i}}$) was calculated as $\mathbf{\Delta_{MUT,i}} = \mathbf{O_{MUT,i}} - \mathbf{O_{WT,i}}$. For each residue i occupancy difference, if $|\mathbf{\Delta_{MUT,i}}| > threshold$ then $\mathbf{\Delta_{MUT,i}}$ is added to an accumulator ($\mathbf{\Delta_{MUT,Total}}$). The absolute value is checked against the threshold to allow for loss or gain of hydrogen bonds, but the signed value is added to the accumulator to see whether an individual system is gaining and/or losing H-bonds. Here *threshold* is set to 0.75 and a mutation considered to have a different occupancy than WT if $\mathbf{\Delta_{MUT,Total}}$ is nonzero. The threshold value of 0.75 was chosen by varying the threshold from 0 to 2 and plotting either the receiver operating characteristic area under the curve, a measure of how well a classifier can distinguish between positive and negative examples, or true positives minus false positives. In both cases, each system had a peak value between 0.7 and 0.8, though in some cases this peak spanned a larger region (data not shown).

3.2.2.2 Solvent accessible surface area

The DSSP [126] algorithm implemented in BIOpython [48] was used to calculate solvent accessible surface area for each residue in each structure. Residues used for analysis plots are indicated in the plot.

3.2.2.3 Kinase activity assays

Kinase activity assays were performed by Jin Park as outlined previously [24]. Briefly, mutations were introduced using the QuikChange method (Stratagene) on Sf9 cell lines. Recombinant protein was then isolated and assayed using an ALK activation loop peptide mimic.

3.2.2.4 Transformation assays

Colony transformation assays were performed by Jin Park as outlined previously [24]. Briefly, NIH 3T3 cells were transfected with ALK variants. Cells were then grown in serum before being fixed in formaldehyde and stained to assess transformation.

3.3 Results

3.3.1 Analysis of kinase activity

In a previous study we showed that analysis of hydrogen bonding and solvent accessible surface area from molecular dynamics simulations correlates well with ALK activation [24]. To test the robustness of this method, which performed very well on activating activation loop and αC helix mutations, we envisioned a series of activation loop and αC helix mutants which we hypothesized would be non-activating. We also wanted to test the method out on a further set of ALK mutations observed in patients. We also ran simulations on a series of mutations in BRAF and HER2, since there are several mutations with known effect in these kinases. The mutations analyzed here are given in Table 3.1, sorted by whether they are non-activating, mildly activating, or activating and which kinase subdomain the mutation occurs in. See the section on kinase structure (chapter 1) or kinase mutations (chapter 2) for further information on kinase subdomains or the distribution of kinase domain mutations.

In order to understand the effects of the series of ALK mutations observed in patients and designed as decoys for MD, Jin Park performed kinase activity and colony transformation assays. The results of these experiments are shown in Figure 3.1.

3.3.2 H-bonds analysis

As discussed in MD studies, analysis of hydrogen bonding patterns on relatively short timescales can be used to differentiate between structures with different energy landscapes and that hydrogen bonds that persist for many 10s of nanoseconds are often stable for much longer times. Previous work has generally only compared a WT protein and one or a few mutants of a single kinase. We sought to undertake a more comprehensive study to elucidate commonalities between a series of mutants in one kinase and between mutations in different kinases. In order to come to an understanding of differences between hydrogen bonding patterns of mutant kinases in the inactive conformation and the WT in the inactive conformation, we sought to understand the general features of hydrogen bond occupancy for a series of inactive conformation mutant simulations in ALK, BRAF, and HER2, as well as active conformation simulations.

3.3.2.1 H-bonds occupancy maps

An example H-bonds occupancy for inactive WT of ALK, BRAF, and HER2 is given in Figure 3.2 as well as a second longer (400 ns) ALK WT inactive contact map. One striking feature of these plots is that they point out the dynamic nature of some H-bonds while showing that most are relatively static. While most H-bonds are formed at the start of a simulation and persist for the duration, some H-bonds flicker in and out of existence, especially in the case of polar and basic residue types. Indeed, there are generally not more than a few H-bonds that show a larger than 30% occupancy difference between the first and second 50 ns of a simulation, as discussed further below. Mutations that can stabilize or

Table 3.1: **Experimental characterizations of mutant kinases**

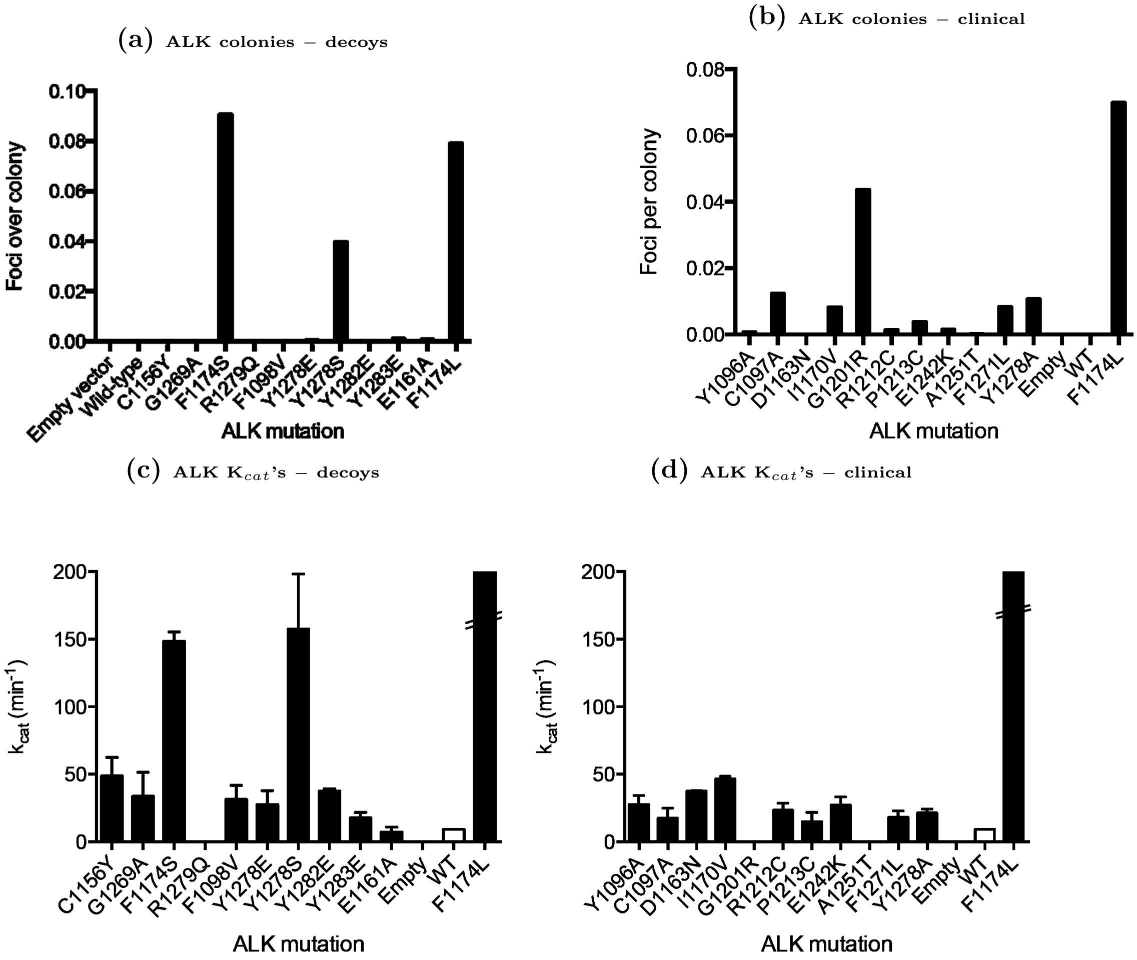
Activation status	Kinase subdomain	ALK Mutations	BRAF mutations	HER2 mutations	Total
Non-activating	P-loop	-	G466V G466E	-	2
	αC helix	E1161A	-	S760A I767M	3
	C-loop	I1250T A1251T	-	-	2
	A-loop	F1271L G1286R Y1283E R1279Q	D5994V G596R	-	6
	rest	C1079A A1200V G1201R P1213C R1231Q T1343I D1349H	G469E	Y835F	9
	total	14	5	3	22
Mildly activating	P-loop	-	-	-	0
	αC helix	D1163N	-	-	1
	C-loop	-	-	V842I	1
	A-loop	G1269A Y1278A Y1278E Y1282E	F595L T599I	-	6
	rest	Y1096A F1098V I1183T L1204F R1212C E1242K	-	R896C L755S	8
	total	11	2	3	16
Activating	P-loop	-	G466A G464E G464V	-	3
	αC helix	M1166R I1170N I1170S I1170V I1171N F1174L F1174S	-	V777L D769H D769Y V773L L768S	12
	C-loop	F1245C F1245V	-	-	2
	A-loop	R1275Q Y1278S	L597V K601E V600R V600K V600E V600D	-	8
	rest	G1128A T1151M C1156Y R1192P L1196M	G469A N581S I463S R462I	-	9
	total	11	9	5	25
Total		41	20	11	72

destabilize these interactions may have impacts on the overall dynamics and conformational landscape of a protein.

3.3.2.2 H-bonds occupancy patterns

To investigate the impact of mutations on H-bonds occupancy patterns, we computed the total Hbonds occupancy difference, $\boldsymbol{\Delta}_{\mathbf{MUT},\mathbf{Total}}$, between a series of inactive conformation mutant simulations in ALK, BRAF, and HER2, as well as active conformation simulations, as outlined in H-bonds methods. This measure only takes into account H-bonds which have rather large differences in occupancies between the mutant and WT simulation, as small occupancy differences could be the result of fluctuations around a well defined minima. Any mutant where the computed $\boldsymbol{\Delta}_{\mathbf{MUT},\mathbf{Total}}$ is nonzero is taken to have an altered hydrogen

Figure 3.1: ALK mutation catalytic and transformation assays



(a) ALK colonies – decoys

(b) ALK colonies – clinical

(c) ALK K$_{cat}$'s – decoys

(d) ALK K$_{cat}$'s – clinical

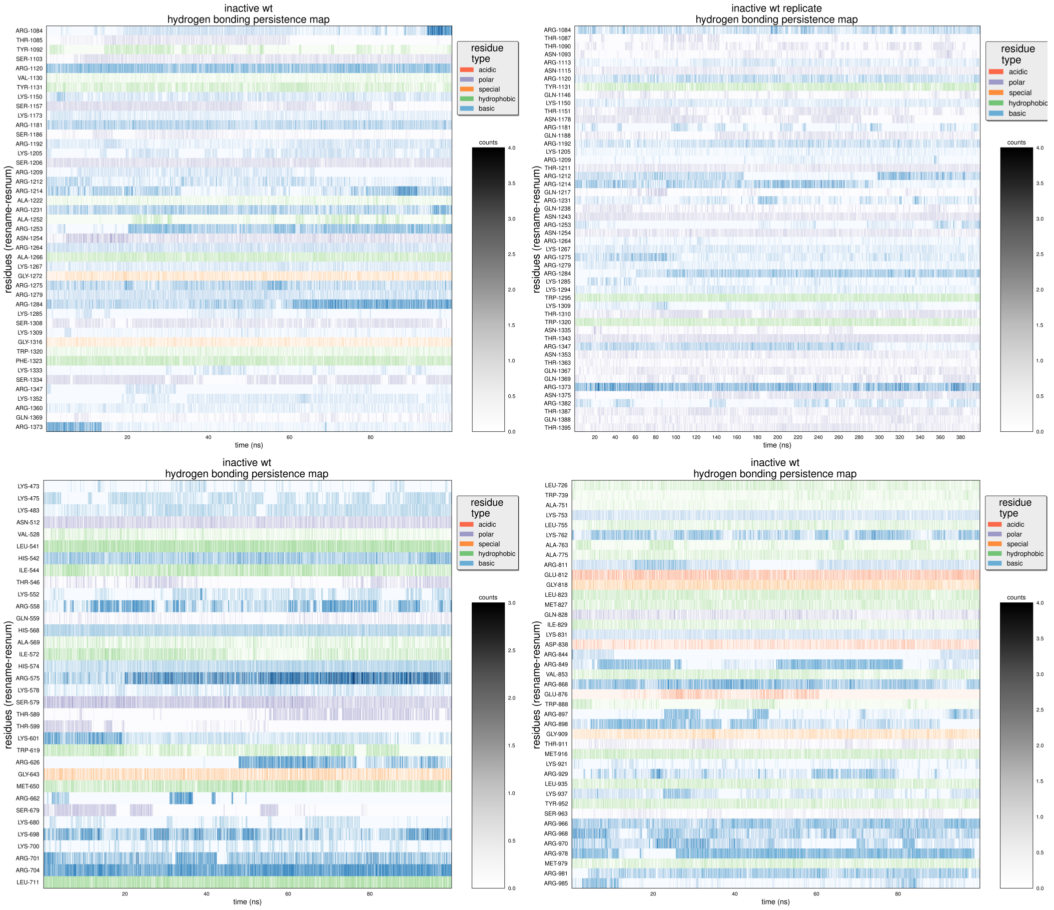
A and C show designed mutations while B and D show patient derived mutations

bonding pattern. The reason to consider mutations with $\mathbf{\Delta_{MUT,Total}}$ close to but not zero is that any mutant with $\mathbf{\Delta_{MUT,Total}} < 0.75$ must actually have at least two altered H-bonds, one gained and one lost, since in our scheme values less then 0.75 are not counted. The maximum difference in occupancy for one H-bond between two simulations is 2 since side chains are considered as a whole and some residues have two H-bond donors or acceptors.

It is important to take into account the extensive work that has gone into characterizing the conformational dynamics of the kinase domain and the differences between the active and inactive conformation as outlined in chapter 1. Confirming previous insights that motions in a few functionally crucial subdomains are important in the kinase activation process is the fact that looking at hydrogen bonding patterns in these subdomains are better able to differentiate between activating and non-activating mutants than the kinase domain as a whole.

In the entire series of mutations studied here, only one, HER2 I767M, shows no difference between WT and mutant $\mathbf{\Delta_{MUT,Total}}$ when looking at all residues in the kinase domain. Considering only H-bonds within αC helix only finds an altered $\mathbf{\Delta_{MUT,Total}}$ in

Figure 3.2: Inactive Hbonds contact maps



(a) ALK WT inactive 100 ns

(b) ALK WT inactive replicate 400 ns

(c) BRAF WT inactive 100 ns

(d) HER2 WT inactive 100 ns

Residues are colored by type of side chain. Darkness determined by number of H-bonds a residue participates in during a single frame. Backbone and side chain contributions are taken together. Only residues that participate in at least one H-bond for at least 10% of frames are shown.

the three ALK mutants E1161A, I1171N, and I1183T while considering H-bonds within the nucleotide binding loop never finds a difference between WT and mutant simulations. In contrast, analysis of H-bonds pattern in several kinase subdomains is assessed for ability to differentiate between activating and non-activating mutation in Table 3.2. Only H-bonds within the top 3 subdomains, or H–bonds within and/or between combinations of subdomains, as assessed by balanced accuracy, (see Table 3.4 for descriptions of measures of binary classifiers) are given in Table 3.2. For each kinase, the balanced accuracy is computed for two different fold changes in catalytic activity, since a large body of evidence has accumulated that mutations which increase activity by greater than 4.5 fold are transforming, but changes of 2 fold could still reasonably be classified as 'altered' (as discussed in the section on kinase activity).

Table 3.2: H-bonds occupancy classificatory power by subdomain(s)

Protein	**activating** K_{cat} **cutoff**	**H-bonds region**	**Rank Score**	**TP**	**FP**	**TN**	**FN**	**BACC** (%)	**TPR** (%)	**TNR** (%)
ALK	2	(P-loop) αC helix & A-loop	1	18	7	7	9	58.3	66.7	50.0
ALK	2	C-loop & A-loop	2	20	8	6	7	58.5	74.1	42.9
ALK	2	(P-loop) A-loop	3	15	4	10	12	63.5	55.6	71.4
ALK	4.5	C-loop & A-loop	1	13	15	10	3	60.6	81.3	40.0
ALK	4.5	(P-loop) αC helix & A-loop	2	12	13	12	4	61.5	75.0	48.0
ALK	4.5	(P-loop) A-loop	3	10	9	16	6	63.3	62.5	64.0
BRAF	2	αC helix & C-loop	1	10	3	2	5	53.3	66.7	40.0
BRAF	2	(P-loop) αC helix & A-loop	2	7	2	3	8	53.3	46.7	60.0
BRAF	2	(P-loop) αC helix & C-loop & A-loop	3	13	3	2	2	63.3	86.7	40.0
BRAF	4.5	(P-loop) αC helix & A-loop	1	6	3	4	7	51.7	46.2	57.1
BRAF	4.5	C-loop & A-loop	2	3	1	6	10	54.4	23.1	85.7
BRAF	4.5	(P-loop) αC helix & C-loop & A-loop	3	11	5	2	2	56.6	84.6	28.6
HER2	2	(P-loop) αC helix & C-loop	1	4	1	2	4	58.3	50.0	66.7
HER2	2	(P-loop) A-loop	2	3	0	3	5	68.8	37.5	100.0
HER2	2	(P-loop) αC helix & A-loop	3	5	0	3	3	81.3	62.5	100.0
HER2	4.5	(P-loop) C-loop	1	3	1	5	2	71.7	60.0	83.3
HER2	4.5	(P-loop) αC helix & C-loop & A-loop	2	5	3	3	0	75.0	100.0	50.0
HER2	4.5	(P-loop) αC helix & A-loop	3	4	1	5	1	81.7	80.0	83.3

See Table 3.4 for definitions of column headers. If two or more sets of subdomains have the same classificatory power they are listed together. Only the top 5 subdomain sets in terms of BACC are shown and the BACC column is highlighted for ease of readability.

Table 3.4: Measures of binary classifiers

Short name	Full name	Alternate name	description or formula
P	Positive	-	activating mutation
N	Negative	-	non-activating mutation
TP	True Positive	-	positive predicted positive
FP	False Positive	type I error	negative predicted positive
TN	True Negative	-	negative predicted negative
FN	False Negative	type II error	positive predicted negative
TPR	True Positive Rate	Sensitivity	$TPR = \frac{TP}{P}$
TNR	True Negative Rate	Specificity	$TNR = \frac{TN}{N}$
BACC	Balanced ACCuracy	-	$BACC = \frac{TPR+TNR}{2}$

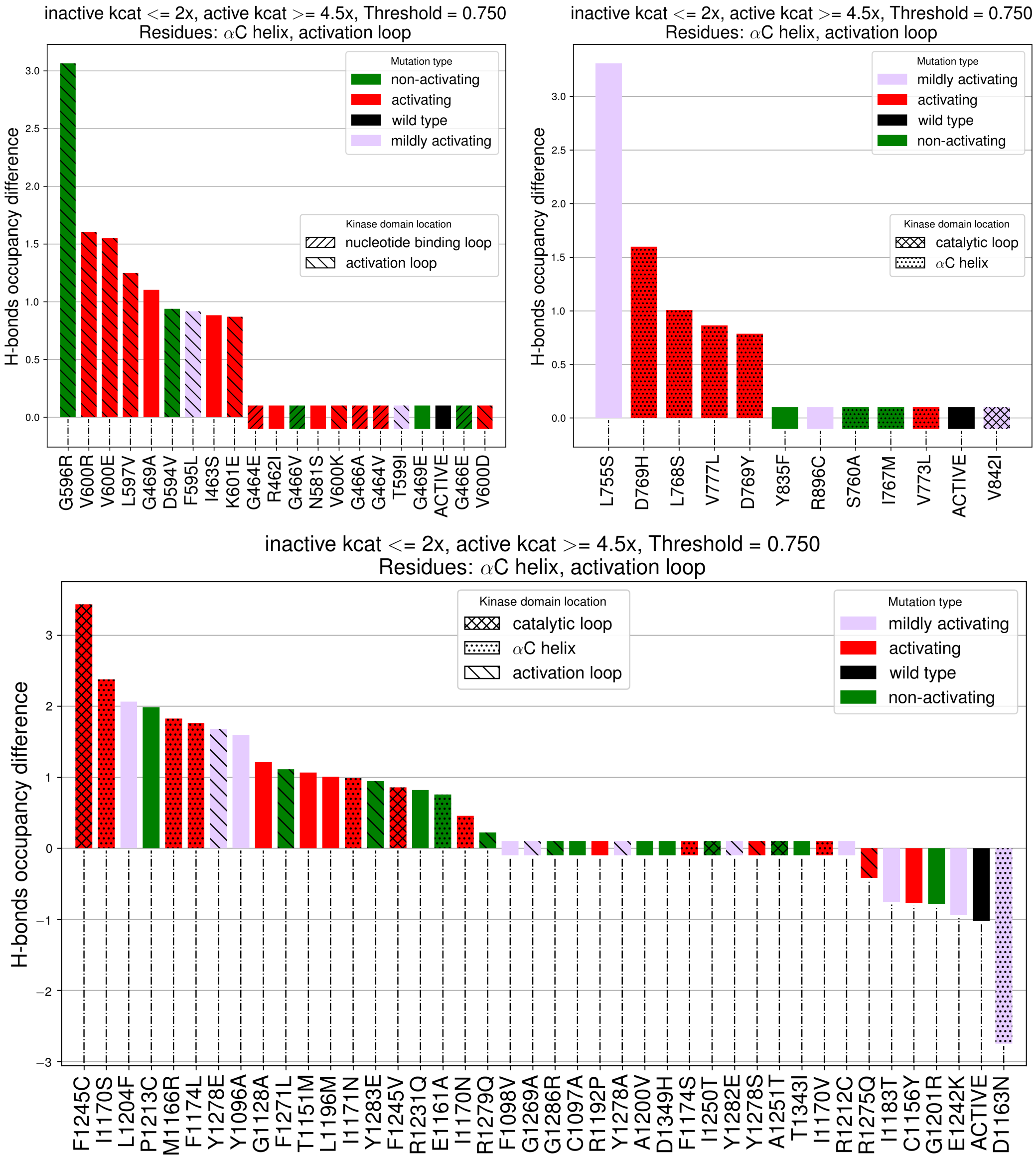
A number of interesting features of hydrogen bonding patterns are discernible from Table 3.2. In each of the proteins studied, not only does the nucleotide binding loop on its own not differentiate between activating and non-activating mutations, there is no case where adding the nucleotide binding loop increases balanced accuracy over leaving it out. This is true even in the case of BRAF where many of the mutations are actually in the nucleotide binding loop (see Table 3.1). This is highlighted in Table 3.2 where (P-loop) indicates that addition of the nucleotide binding loop does not alter the scoring. Only in 4 out of the 18 cases listed does adding the nucleotide binding loop alter the balanced accuracy; in all four cases the balanced accuracy decreases. Another interesting feature is that the cutoff used for catalytic threshold of an 'activating' mutation does affect the orderings of which set of subdomains changes. In order to find an objective measure of which set of subdomains best discriminates between activating and non-activating mutations, we score each of the top 3 sets of subdomains according to their rank. Using the lower threshold of 2x increase in catalysis, the best set of kinase subdomains to use is αC helix and activation loop with a score of 5, followed closely by only the activation loop with a score of 4. For the higher threshold of 4.5x increase, again the best set of kinase subdomains to use is αC helix and activation loop with a score of 5, but here looking at $\mathbf{\Delta_{MUT,Total}}$ in the αC helix, catalytic loop, and activation loop also gives a score of 5. The fact that using the αC helix and activation loop resulted in the best balanced accuracy score in both cases provides verification of our previous work which used this as a scoring function in a series of ALK mutations [24], which comprise about half the ALK mutations here. A plot of $\mathbf{\Delta_{MUT,Total}}$ within and between the αC helix and activation loop for all mutations studied here is given in Figure 3.3.

3.3.2.3 H-bonds occupancy histograms

Several salient features of Figure 3.3 can be noted. Since inclusion of the nucleotide binding loop in H-bonds scoring has almost no effect it is not surprising that the only activating mutations in this dataset to fall within the nucleotide binding loop, all in BRAF, are not scored properly. This points to the need for alternate scoring functions since there may be different activation mechanisms involved in different mutants. Indeed, there is evidence that BRAF P-loop mutations function by reducing binding affinity between BRAF and MEK, leading to increased BRAF-CRAF dimerization and increased downstream signaling [87]. BRAF has two false positives, D594V and G596R, which are both in the DFG motif that

Figure 3.3: H-bonds occupancy plots

(a) BRAF (b) HER2



(c) ALK

starts the activation loop in most kinases [109]. Both of these mutations have been shown to be capable of activating their downstream partners even though they have very low intrinsic kinase activity relative to WT, due to a propensity to promote dimerization with CRAF [254, 94]. In the case of HER2, L755S has proven difficult to express at levels sufficient for detailed catalytic measurements, but also to result in increased activation of EGFR in some cell lines [21]. HER2 L755S is the only kinase mutation in this study that has not been characterized in terms of it catalytic activity, but has been demonstrated to lead to increased phosphorylation of downstream partners, to be able to transform cell lines, and to confer lapatinib resistance [246, 129, 21, 131, 273, 264] and is given a moderately activating score, though it could actually lead to a decrease in activity and still be transforming. In the ALK case, L1204F (from [24]), G1201R, and R1279Q (see Figure 3.1) have also proven difficult to express in sufficient quantity for *in vitro* measurement. This may suggest that some of the false positives in the $\mathbf{\Delta_{MUT,Total}}$ analysis presented in Figure 3.3 may adopt a conformation distinct from the WT inactive conformation but still not be biased towards an active conformation, as has been shown for some mutations studied via MD as discussed in introduction on metadynamics simulations of kinase mutants (also see the discussion on H-bonds false positives).

3.3.2.4 H-bonds occupancy is strongly influenced by labile hydrogen

In an effort to understand the variability in individual H-bonds, as illustrated in Figure 3.2, we undertook an analysis of how labile individual H-bonds are. As a metric for labile H-bonds, we investigated which H-bonds show a change in occupancy of greater than 30% between the first and second half of any mutant simulation, again limiting our analysis to H-bonds that show a change in occupancy of at least 0.75 between WT and mutant. We found that only a small number of H-bonds are labile in this manner, across all systems. The H-bonds most frequently found to be labile is given in Table 3.5.

We ran a total of 86 ALK simulations, 44 BRAF simulations, and 26 HER2 simulations, in each case simulating each mutant twice and also two simulations each of WT in active and inactive conformations. The total number of labile H-bonds across all simulations is 390 in ALK, 147 in BRAF, and 123 in HER2, an average of about 4 per simulation. This confirms the intuition derived from Figure 3.2 that most H-bonds do not fluctuate much over the course of a simulation.

The data in Table 3.5 shows that a small number of labile H-bonds recur across simulations of different mutants and that this pattern is true for all three kinases studied here. Remarkably, these labile H-bonds also account for most of the $\mathbf{\Delta_{MUT,Total}}$ value found for the subdomains that contain them. This is shown in Table 3.6 which for each kinase give BACC, TPR, and TNR for (1) the top 3 subdomains for the 4.5x K_{cat} cutoff as listed in Table 3.2 but here using only the few labile bonds found in those subdomains, and (2) the top three single residues as scored by BACC for differentiating activating from non-activating mutations, also using the 4.5x K_{cat} cutoff. Not only does only considering labile bonds effectively recapitulate considering all residues in a subdomain, in each kinase the single labile residue with the highest BACC has a similar BACC to considering all residues in a (set of) subdomain(s). Also of note is the fact that the H-bonds most frequently found to be labile in Table 3.5 are not all found within a subdomain, but only one of the high BACC residues in Table 3.6 is not in a subdomain.

Table 3.5: Labile H-bonds

Kinase	H-bond donor	H-bond acceptor	# simulation bond is labile
ALK	Arg-1181 side	Glu-1197 side	47
	Arg-1231 side	Glu-1384 side	40
	Arg-1253 side	**Asp-1249 side**	33
	Arg-1284 side	**Asp-1163 side**	30
	Arg-1279 side	**Asp-1163 side**	23
	Arg-1275 side	**Asp-1276 side**	22
	Arg-1284 side	**Asp-1276 side**	21
	Listed bonds/total labile bonds		216/390
BRAF	Arg-662 side	**Glu-611 side**	23
	Arg-701 side	Asp-702 side	19
	Arg-701 side	Asp-629 side	19
	Arg-626 side	Asp-629 side	16
	Arg-558 side	Glu-715 side	9
	Arg-603 side	**Glu-501 side**	9
	Arg-558 side	Asp-555 side	8
	Arg-575 side	**Glu-501 side**	7
	Listed bonds/total labile bonds		95/147
HER2	Arg-981 side	Asp-982 side	12
	Arg-868 side	**Glu-770 side**	10
	Arg-849 side	**Asp-845 side**	10
	Arg-844 side	**Asp-873 side**	9
	Arg-968 side	Glu-837 side	8
	Arg-970 side	Glu-837 side	7
	Listed bonds/total labile bonds		53/123

Residues are colored by location within the kinase domain as **C-loop**, αC helix, and **A-loop**. All listed H-bonds are between residue side chains (side) but main chains were also considered.

Finally, we point out that labile H-bonds are at least partially conserved, as demonstrated by Table 3.7 which shows labile H-bonds that are in the same position in the kinase domain. Only 2 of the 8 conserved labile H-bonds in Table 3.7 are not found in a subdomain. Three of the six labile H-bonds in Table 3.7 are in the catalytic loop, and two of these are found in the catalytically important HRD motif. The highly conserved catalytic residue is the D of the HRD motif [135] and the pair ALK-D1249:HER2-845 are at this position. The R of the HRD is thought to help coordinate the active site, though it is sometimes absent in kinases that do not undergo activation loop phosphorylation [135] and the pair HER2-R844:BRAF-R575 are at this position. A cross-referencing between residues in Table 3.6 and Table 3.7 shows that many of the conserved labile H-bonds are also good at differentiating activating from non-activating mutations based on BACC.

Table 3.7: H-bonds at the same position in kinase domain

αC helix	BRAF Glu-501	HER2 Glu-770
C-loop	ALK Asp-1249	HER2 Asp-845
C-loop	BRAF Arg-575	HER2 Arg-844
C-loop	ALK Arg-1253	HER2 Arg-849
A-loop	ALK Arg-1279	BRAF Arg-603
A-loop	ALK Arg-1275	HER2 Arg-868
-	ALK Arg-1231	BRAF Arg-558
-	ALK Glu-1384	BRAF Glu-715

Position within kinase domain is given in the first column.

Table 3.6: H-bonds occupancy classificatory power by residue(s)

Protein	**Donors**	**Acceptors**	**BACC** (%)	**TPR** (%)	**TNR** (%)
ALK	**Arg-1275 Arg-1279 Arg-1284**	**Asp-1163 Asp-1276**	60.1	56.3	64.0
ALK	**Arg-1253**, **Arg-1275 Arg-1279 Arg-1284**	**Asp-1249 Asp-1276**	59.5	75.0	44.0
ALK	**Arg-1275**, **Arg-1279**, **Arg-1284**	**Asp1276**	57.0	50.0	64.0
ALK	-	**Asp-1163**	55.6	31.3	80.0
ALK	**Arg-1275**	-	56.5	25.0	88.0
ALK	-	**Asp-1276**	63.3	62.5	64.0
BRAF	**Arg-603**	**Glu-611**, **Glu-501**	40.1	23.1	57.1
BRAF	**Arg-575**, **Arg-603**	**Glu-611**	50.0	0	100
BRAF	**Arg-575**, **Arg-603**	**Glu-501**, **Glu-611**	48.9	69.2	28.6
BRAF	-	**Glu-501**	48.9	69.2	28.6
BRAF	-	**Glu-611**	54.4	23.1	85.7
BRAF	Arg-662	-	54.4	23.1	85.7
HER2	**Arg-844 Arg-849**	**Asp-845**	71.7	60.0	83.3
HER2	**Arg-844 Arg-849 Arg-868**	**Glu-770 Asp-845 Asp-873**	55.0	60.0	50.0
HER2	**Arg-868**	**Glu-770 Asp-873**	61.7	40.0	83.3
HER2	-	**Glu-770**	71.7	60.0	83.3
HER2	-	**Asp-845**	71.7	60.0	83.3
HER2	**Arg-868**	-	81.7	80.0	83.3

Residues are colored by location within the kinase domain as **C-loop**, **αC helix**, and **A-loop**.

3.3.3 Solvent accessible surface area analysis

3.3.3.1 R-spine

As outlined in the introduction to kinase structure in chapter 1, much recent work has gone into elucidating the role of a regulatory spine in kinase activation. In the active conformation the R-spine is assembled, with all four residues in close proximity. In the inactive DFG-out conformation, the F of the DFG motif points towards the active site, breaking the R-spine [135]. In the case of BRAF workers have shown that replacing L505 residue in the R-spine with a Phe residue leads to constitutive kinase activation, potentially by promoting R-spine formation [104]. In order to see if a similar activation mechanism may be at play in any of the mutants considered here, we used changes in R-spine solvent accessible surface area (SASA) as a proxy for R-spine assembly, as outlined in R-spine methods. The reason SASA was chosen as a metric as opposed to a distance or dihedral angle based metric is that the method originally used to discover the role of the R-spine was based on changes in surface exposure as measured by SASA using DSSP [135], as we have done here. We note that in [135] a probe radius of 1 Å is used but that we use here a 1.4 Å radius, the default in DSSP since it is close to the radius of a water molecule [126]

None of the simulations considered here were prepared in the 'DFG-out' inactive conformation but instead have the catalytic Asp pointing towards the active site. This was not a conscious modeling decision but the result of the crystal structures used for homology modeling. Indeed, a majority of structures in the PDB adopt a DFG-in conformation, with around 3 times more DFG-in structures than DFG-out [251]. Recent studies using free energy calculations have shown that the DFG-in conformation is lower in energy in both ABL and SRC kinases [164], though protonation of the DFG Asp can alter this [152], while free energy studies of BRAF have shown that both the WT and V600E systems preferentially adopt a DFG-in conformation [157].

Since all systems have a starting configuration with the R-spine formed, we should not expect a large difference between the active and inactive WT R-spine SASA, and indeed this is confirmed in Figure 3.4. Only in the case of BRAF do the means of the active and inactive WT differ by more than 1 standard deviation. Curiously, BRAF is also the only case where the active WT mean is higher than inactive WT. Mutational analyses have demonstrated that R-spine substitutions of larger hydrophobic residues can lead to constitutive kinase activation [104, 103], while free energy calculations have shown that ABL has a larger active site SASA [152], so we hypothesized that mutations that lead to larger R-spine SASA might correlate with activating mutation. Unfortunately, no such patter is detectable in Figure 3.4. We also investigated whether there was any correlation between activation status and the standard deviation of SASA, but again there was no discernible relationship (data not shown).

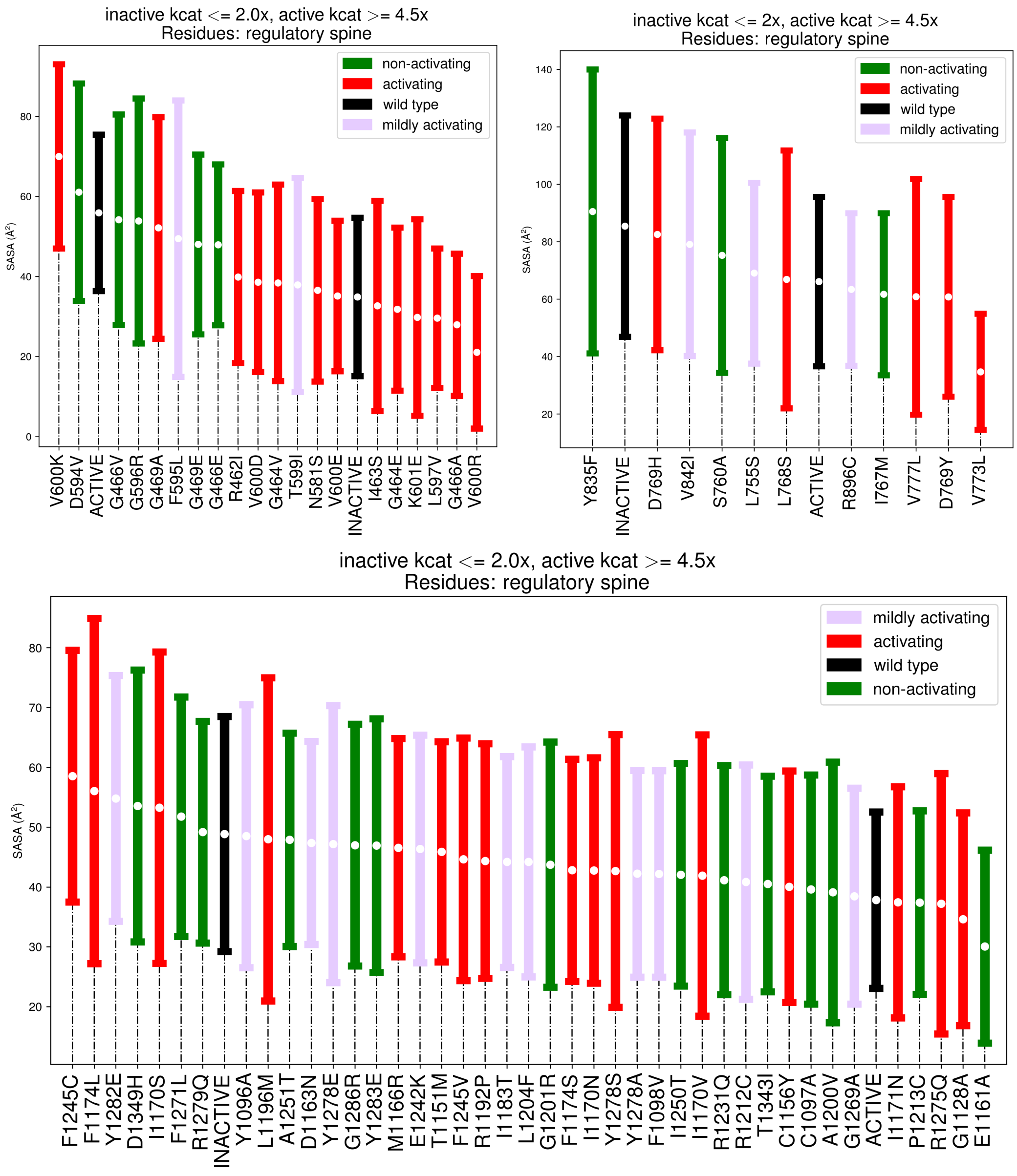
3.3.3.2 Hydrophobic core

We also sought to measure changes in SASA of a hydrophobic core of residues as we had previously for some of the ALK mutations under study here [24]. While in that earlier study the hydrophobic core correlated well with activation status, the new set of ALK mutations did not show a similar correlation, nor did BRAF or HER2 (data not shown). By analyzing the contributions of each residue included in the hydrophobic core analysis, we determined that the two ALK JM residues Y1096 and F1098, both located rather peripherally in the hydrophobic core, are subject to larger fluctuations in SASA relative to other core residues. When these residues were left of the hydrophobic core analysis the correlation to activation status in the ALK mutants studied in [24] disappeared and SASA patterns of ALK mutants more closely resembled those of BRAF and HER2, which only have the kinase domain in the structures used here.

3.3.3.3 Genetic algorithm

Finally, in an attempt to determine whether there might be some previously unrecognized pattern of hydrophobic residues that could separate activating from non-activating mutations in the systems under study here. To this end, we implemented a genetic algorithm (GA) that used as its fitness function (1) difference in SASA of the two classes, activating and non-activating mutations, (2) number of residues included in the analysis, and (3) receiver operating characteristic area under the curve of classifying mutations activating above or below a SASA threshold. In the case of both (1) and (2), both minimization and maximization were tried. In all cases attempted, it was found that there is a set of residues for which SASA can distinguish without error on the training set but do very poorly on a test set. Combinations of residues as small as 11 and as large as 39 were able to achieve an AUC of 1 on training sets (data not shown).

Figure 3.4: H-bonds occupancy plots



(a) BRAF

(b) HER2

(c) ALK

3.4 Discussion

3.4.1 H-bonds false positives

As discussed in H-bonds occupancy histograms, the BRAF D594V and G596R mutations diminish catalytic activity but lead to transformed cell lines. The same is also true of the kinase dead BRAF mutation K483M and it is thought that the underlying mechanism in these cases increased ability to act as a scaffold for CRAF [254, 94]. The K753E mutation of HER2 is at the same position in the αC helix, which is conserved among kinases and important for forming a salt bridge important for ATP binding and catalysis. This mutation has been shown to lead to decreased kinase activity of HER2 but to increase signaling through EGFR [66, 273]. In EGFR the K745M mutation, often discussed in the literature as K721M due to a numbering scheme based on the removal of a targeting sequence from the mature peptide, is also known to reduce kinase activity [66]. This kinase deficient EGFR is still known to be able to stimulate activation of downstream targets upon EGF stimulation and this stimulation is dependent on HER2 kinase activity, suggesting that EGFR is acting as a scaffold for HER2 in this case [55]. It has also been shown that EGFR K745 may be a target for monomethylation and that this methylation leads to EGF independent EGFR C-terminal tail phosphorylation in cell lines [217]. This evidence, when taken together, points to a mechanism where mutations that reduce kinase catalytic activity in several proteins nonetheless result in increased activation of downstream signaling pathways. Also of note is that HER2 L755S has been demonstrated to lead to cell line transformation and lapatinib resistance [246, 129, 131, 273, 264] and is difficult to express for kinase activity assays [21]. The ALK mutations G1201R and L1204F both express poorly and lead to transformation in cell lines [24] (also see Figure 3.1). It is noteworthy that though ALK G1201R and L1204F, BRAF D594V and G596R, and HER2 L755S are all known to lead to transformation of cell lines despite lessened (or difficult to measure) kinase activity relative to WT, all of these mutations show strong changes in H-bonds occupancy within and between the αC helix and activation loop, as shown in Figure 3.3. This raises the question of whether these mutations are actually 'false positives.' These mutations with diminished activity but paradoxical activation of downstream targets could plausibly be attributed to structural changes that bias the system away from the inactive conformation. Future studies of the three dimensional structure or conformational free energy landscape of these mutants could lead to interesting insights into how these mutants function and potentially validate the use of $\boldsymbol{\Delta}_{\mathbf{MUT,Total}}$ as a scoring function for kinase domain mutations.

3.4.2 Utility of SASA

As seen in Figure 3.4, SASA of the 4 R-spine residues undergo rather large fluctuations relative to their means. This also holds when a larger number of residues like the hydrophobic core is considered. We were also able to find sets of residues using a GA which perfectly separated activating and non-activating mutations in a training set, but which performed poorly on cross-validation. This suggests that SASA may be inherently too noisy to be useful for distinguishing between activating and non-activating mutations. However, we cannot rule out that a measure that accounted for changes in solvent density as opposed to solvent area [187] would yield better results.

3.4.3 Importance of R-spines and H-bonds

The inability of SASA analysis of R-spine residues to distinguish activating from non-activating mutations does not rule out their importance in kinase activation. Even if SASA is a plausible metric for monitoring the DFG flip which would coincide with breaking of the R-spine, the time scales of the simulations performed here may preclude actually observing such an event. An NMR investigation of p38 MAP kinase found that the DFG flip occurs in the microsecond to millisecond time scale, though only by inference as the actual flip was not directly measured, only shown not to happen at sub-microsecond timescales and to have already occurred at millisecond time scales [253]. For protein kinase A, NMR has shown that DFG flip is not observed in the apo state but that upon nucleotide binding DFG flip, and other rearrangements concomitant with activation, occur on millisecond time scales with similar rates to measured K_{cat} time scales [159]. As such, the DFG flip and thus R-spine (dis)assembly is unlikely to be observed on the time scales accessible to conventional MD. One study did report R-spine breaking in SRC on the 100s of nanoseconds, but did not observe DFG flip, only a change in orientation of the αC helix [71]. We also frequently observe motion of the αC helix in our simulations. Many enhanced sampling studies have computed free energy landscapes for DFG flip [152, 250, 148, 164, 171], showing that in most cases the DFG-in conformation is energetically favorable. Taking into account both the long time scales of DFG flip in R-spine (dis)assembly and the fact that an analysis of H-bonds on much shorter time scales does seem to correlate with functional alterations, the question of the priority of spines vs H-bond networks can be asked.

A small number of residues changing their hydrogen bond occupancy does seem to correlate with mutational activation status and these residues also happen to be among the only ones that are regularly seen to undergo large changes in occupancy during MD simulations. This points towards a sensitive electrostatic switch that can be disrupted by changes in charge such as BRAF V600E and ALK R1275Q or electrostatic screening effects that result from changes in size or configuration of hydrophobic residues. This could explain how HER2 V773L and ALK I1170V, which are at the same position, could both be activating though in both cases the WT and mutant residues only differ by one carbon. A similar argument could be made for HER2 V777L and ALK F1174L, or to explain why, in addition to BRAF V600E/D/K/R all being activating, BRAF V600F also leads to constitutive activation [103]. This electrostatic switch seems to be active on a much shorter time scale than (dis)assembly of the R-spine or DFG flip, and thus it seems likely that R-spine formation is a result of the changes in H-bonds patterns and not a driver of these changes.

Chapter 4

Biophyscode

> Everything rests here on the mode in which the passage from potentiality to act comes about.
>
> Giorgio Agamben *The Coming Community*

4.1 Introduction

The use of molecular dynamics (MD) simulations to study and understand biomolecular systems such as proteins and membranes allows insights into biophysical mechanisms that can be difficult or impossible to gain with other techniques. This has combined with ongoing decreases in cost and increases in availability of computational resources to make MD an increasingly popular technique for investigating biomolecular systems [213]. A number of popular software packages exist for setting up, running, and analyzing MD simulations, but these operations are not always integrated into a complete pipeline and often require extensive use of a command line interface (CLI). A recent report [213] showed that GROMACS [203, 181, 2] was the most popular MD engine, followed by VMD [107], LAMMPS [199], NAMD [193], CHARMM [26], and AMBER [37]. In this paper we will outline the need for, and workings of, an integrated GROMACS simulation and analysis pipeline with a web-based graphical user interface (GUI), but first we will give a brief overview of existing tools and software used for biomolecular MD.

4.1.1 Existing MD tools

NAMD Not Another Molecular Dynamics program (NAMD [193]), and Visual Molecular Dynamics (VMD [107]) are popular for running and analyzing MD simulations [213]. VMD has a graphical user interface with plugins that allow for the generation of NAMD input files for solvated and ionized protein or protein membrane system. Mutations can also be made to proteins using VMD. These NAMD files can then be run using conventional MD or enhanced sampling techniques such as replica exchange, accelerated MD, free energy perturbation, and thermodynamic integration. VMD has a number of analysis tools such as root mean squared deviation or fluctuation (RMSD, RMSF), hydrogen bonding and salt bridges, solvent accessible surface area (SASA), normal modes, and tools for analyzing collective variables. It is in principle possible to write scripts to chain together simulation setup, run, and analysis, but this requires knowledge of the tcl programming language since that is what the VMD interface is written in. Recently, the QwikMD [213] package was released to allow NAMD/VMD users to have an integrated pipeline for preparing, running, and analyzing MD simulations. In addition to the features of NAMD and VMD, QwikMD allows users to select a structure from the protein databank, select which chains or residues from the structure to use, and to model in any missing residues. QwikMD also allows users to run MD simulations on Amazon Webservice.

CHARMM Chemistry at HARvard Molecular Mechanics (CHARMM [26]) is another popular MD engine. Although there is a free academic version of CHARMM available, there is a fee to obtain a version compatible with high performance compute cluster usage. CHARMM can be run at the command line using a CHARMM specific scripting language but also has a graphical user interface called CHARMM-GUI [120]. CHARRM-GUI allows users to download coordinates from the PDB, select chains, add any missing residues, and solvate a protein and can also be used to create protein-membrane systems. Once the system is set up CHARMM-GUI allows users to run conventional, targeted, steered, or replica exchange MD, and is compatible with a number of all atom and coarse grain force fields. The majority of the tools available in CHARMM-GUI are for setting up and running protein, protein-membrane, and protein-ligand systems. The main CHARMM package has a large number of analysis tools however, including measuring quantities such as RMSD and RMSF, radius of gyration, principal component analysis (PCA) and autocorrelation, SASA, hydrogen bonding, bond angle distributions, and density fluctuations.

AMBER Assisted Model Building with Energy Refinement (AMBER [37]) is another non-free MD engine; the associated analysis suite, AmberTools is available free. AMBER has command line tools for preparing PDB structures but users must generally already have found or generated a complete protein structure. After obtaining a structure, AMBER users can then solvate, minimize and equilibrate the system, and carry out conventional, replica exchange, accelerated, and nudged elastic band MD; umbrella sampling and TI are also available. Various analysis tools allow for the calculation of various quantities such as RSMD and RMSF, bond angle distributions, secondary structure, density, radius of gyration, diffusion, and velocity autocorrelation. Recently a Kepler workflow has been introduced which allows for GUI based set-up, local or cluster based execution, and analysis of AMBER MD simulations [204].

GROMACS The GROningen MAchine for Chemical Simulation (GROMACS) is the most widely used biomolecular MD tool and supports a wide variety of use cases. GROMACS has CLI tools that allow a protein structure to be solvated in a water box with ions, minimized, equilibrated, and run as production MD with a number of different all atom or coarse grained force fields and can be run either on a local machine or on a compute cluster. Tools to create biomembranes and to place proteins on or in a membrane also exist. One drawback of GROMACS is that users must already have a complete protein structure as no mutation or homology modeling tools are provided. In addition to the core functionality of setting up and running MD simulations, there are a large number of analysis tools that are included in the GROMACS package such as PCA, RMSD and RMSF, SASA and protein secondary structure, clustering structures, radius of gyration, density maps, hydrogen bonds and salt bridges, bond angle distributions, velocity auto-correlation, as well as tools for selecting subsets of atoms and timesteps from a trajectory. Additionally, GROMACS works with a number of free energy calculation methods such as free energy perturbation, thermodynamic integration, and replica exchange with built-in tools to analyze the result of such calculations. One reason that GROMACS has such a large selection of features is the admirable commitment of the developers to maintaining an open source codebase with many contributors. One potential downside here is that all of these codes are run from the command line with separate calls and runtime options, making it necessary to either run each program manually from the command line in turn or to write scripts that chain together large numbers of commands. Another drawback is that since many of the analysis tools were written by independent contributors, there may not be ongoing support for some analysis tools.

One tool that has seen extensive use in the MD community, and which is used extensively in the work discussed here, is MDAnalysis [165, 80]. MDAnalysis is a python based interface for reading in MD topology and trajectory files and performing analysis using built-in codes such as RMSD, hydrogen bonding, water density, and PCA as well as any operation that can be performed on the coordinates of the trajectory via NumPy or SciPy [248]. MDAnalysis can read in topology and trajectory information from GROMACS, NAMD, CHARMM, LAMMPS, or Amber and give programmatic access to the underlying coordinates as NumPy arrays, making it a very important addition to the MD ecosystem. Another similar tool is MDtraj [161], which offers trajectory readers for a number of MD engine trajectory formats and also several NumPy and scikit-learn [190] based analysis tools. Although there have been efforts to develop GUI based tools for GROMACS [214, 221, 155], as yet there has not been a single tool which integrates simulation setup, execution, and analysis.

4.1.2 Need for a GROMACS framework

The fact that GROMACS is only available on the command line and requires several commands to be run sequentially necessitates the creation of scripts to chain together GROMACS calls. Discussions that the authors have had with other groups that run MD using GROMACS points to each lab having its own in-house codebase to set up and run simulations. One major drawback of this approach, and one of the major driving forces for the development of BioPhysCodes, is the fact that taking an *ad hoc* approach whereby scripts or inputs are modified as needed can make it very difficult to keep track of which simulations were run under which conditions. The other principle diving force behind the creation of BioPhysCodes is that having a relatively (or completely) automated system to set up and run simulations allows workers unfamiliar with running MD simulations, or even with use of the command line, to rapidly start running simulations and analysis. This has proven especially useful for people who are only working on a project for a period of a few months such as rotation, undergraduate, and even high school students. The need for an integrated framework for setting up, running, and analyzing simulation data is widely appreciated, as demonstrated by the recent development of such frameworks for NAMD [213] and AMBER [204], as well as the previous development of CHARMM-GUI [120]. The authors hope that the BioPhysCodes can be a useful and welcome addition to the MD field that brings

GROMACS ease of use into line with other MD engines.

4.1.3 Need for “big green button”

Reproducibility has been shown to be a major problem in many fields of science [115], with efforts to reproduce preclinical results proving successful only 11% [17] to 25% [202] of cases in two widely cited studies. Leaving aside epistemological considerations [138], only issues related to sampling should prevent easy reproducibility in a field like MD which has a rigorous underpinning in statistical mechanics. Nonetheless, one roadblock to reproducibility is the large number of parameters that need to be set appropriately in order for an MD simulation to run correctly. While small changes in temperature or pressure coupling, and even large changes like simulating in a different ensemble, should theoretically eventually lead to the same result [188], in practice and at shorter time scales choices of parameters and minimization and equilibration protocols can lead to different results in systems with the same atoms [180].

To ensure both ease of use and reproducibility, the BioPhysCodes, which consist of the Automacs module for setting up and running simulations as well as the Omnicalc module for analysis, is set up in a modular fashion. The underlying Python framework is minimal and the actual analysis and simulation protocols are retrieved (automatically) from `git` repositories. BioPhysCode also has a Factory module which is a web based interface built on a Django framework that can be used both as a GUI and as a way to archive simulation and analysis data. The combination of the Factory, Automacs, and Omnicalc codes which are contained in BioPhysCodes allows for maximum reproducibility of MD experiments.

4.2 Methods

4.2.1 Inefficiency in standard methods

A typical simulation workflow for both experts and trainees starts when they conjure a target simulation system from an interesting scientific question. This choice is often heavily constrained by the molecules available in state-of-the-art force fields, or their purported reliability for simulating the right physics. After selecting a target, researchers must build a coherent model from a combination of useful experiments (e.g. X-ray, NMR, or cryo-EM for protein structure), protein homology modeling (for incomplete proteins), or the careful relaxation of an approximate structure for soft matter systems like polymers and lipid bilayers. Rarely can complicated new systems or materials be simulated from scratch in atomic detail without knowledge of some sort of constraints on possible configurations.

Because of their high computational cost, most simulation tools are written in fast, low-level languages like C and Fortran. The scientist typically calls them from a linux command line in a “bucket brigade” fashion in which the outputs for one simulation or model-building step are fed into the next. The precise computational “experiment” they create takes effort to generalize and automate. Moreover, there are often multiple, equally valid ways to accurately construct a good model. For example, you can make a bilayer by randomly scattering lipids in a box or by carefully arranging them in a grid. After the data are generated, there are a number of different ways to package and archive the data for analysis and future study. As a result, there are major barriers to auditing it, replicating the procedure precisely, or sharing the underlying methodology without personalized training. Measuring physical properties from a simulation either generates a massive amount of new data specific to that calculation (e.g. a pressure tensor calculation) or a highly-reduced dataset that still needs to be paired with the source data to ensure that its conclusions are robust. The necessary "manual" archival imposes a maintenance cost on many large sets of data.

The process described above contains both community-developed standards and personal taste in seemingly equal measure. Each researcher has a unique, presumably rational approach to building, simulating, and storing molecular dynamics trajectories. Nevertheless, constructing these systems can sometimes resemble a fine art. This is common to most basic scientific research, and the commensurate training and development costs also ensures that most molecular dynamics practitioners are experts in their trade. In the next section we will describe some modernization methods which seek to automate and streamline this task without replacing subject matter experts and instead making their (our) work more transparent.

4.2.2 Modernizing methods

Despite the sometimes individual character of molecular simulations conducted in research groups small and large, there are a number of community standards and best practices that software developers have used to standardize the research in this field. Many of these practices have informed our approach, and hence they are worth reviewing here.

(1) All popular integrators have extensive documentation, tutorials, and examples so that new users can get started (c.f. AMBER[1], CHARMM[2], GROMACS[3], VMD[4]). Most codes are distributed in an open- or quasi-open-source format so that scientists can carefully audit them.

(2) Users can build on existing codebases when they are combined with a scripting language. One prominent example of this is VMD, which is built on top of the TCL scripting language [107]. This provides two major advantages. First, users can easily interact directly with VMD features, calculation functions, and even automate tasks that typically occur at the GUI. Second, having access to the raw data makes it possible for users to save their results in a more durable format. Other programs like Modeller [216] and pymol [220] use python as a backend, and hence benefit from its well-known syntax. Interoperability between simulation formats is also provided by tools that lack a graphical interface, like MDAnalysis [165] or MDtraj [161], both of which are capable of reading data generated by several different integrators.

(3) The integrators themselves provide a number of features that standardize their outputs. For example, GROMACS and LAMMPS both use standard force field inputs so that they can use force fields native to *other* integrators. GROMACS also outputs simulation data in multiple common trajectory formats.

Embedding simulation tools in preexisting languages, writing standard trajectory formats, and integrating analysis tools are all necessary components of a standard pipeline.

[1] ambermd.org/doc12/Amber14.pdf

[2] charmm.org/charmm/documentation/

[3] manual.gromacs.org/documentation/2016.4/manual-2016.4.pdf

[4] ks.uiuc.edu/Research/vmd/current/ug/

However, there are many more modern software engineering tools that can help to improve on these tools. First, many codes must be compiled to match underlying hardware. For new users, this can be a major bottleneck. Virtual machines like Docker make it possible to run these calculations in an isolated software environment which can be endlessly replicated across different hardware (and also provides a reproducible compilation procedure). Other language-specific tools, like Python Anaconda allow users to install a large set of software dependencies (including external linux-packages). Modern software engineering makes heavy use of so-called "test sets" which test developing codes on previous use cases to ensure that backwards compatibility is not broken. Syntax formats like YAML (yet another markup language) make it possible to describe arbitrary data structures and control flow without clumsy programming syntax. Scientific python packages, namely SciPy [248], are now competitive with native C and Fortran code when running expensive calculations (typically because they call on these codes directly). Cross-platform binary output formats provided by tools like HDF5 make it possible to store data in durable, machine and hardware independent, format. Finally, multiple simulations or calculations can easily be run at once by taking advantage of the embarrassingly parallel nature of running simulations of or calculations on separate systems. Basic multiprocessing tools available in python also make it possible to run many calculations in parallel without switching to a native-parallel language or writing lower-level code manually.

The rich ecosystem of different software tools described above lends itself to many correct solution methods, however many of the tools described above have peculiar limitations, specific use-cases, and a lack of full interoperability. In the remainder of this methods section we will outline a framework which uses all of these tools while making as few arbitrary design choices as possible in order to close the loop and eliminate the high costs of extending the functionality of a particular code. In the following results section, we outline how different modular elements of BioPhysCode can be combined to set up and analyze diverse protein, lipid, and carbohydrate systems.

4.2.3 Our objective

Our present task is to solve the "last mile" problem for molecular simulation tools by building a single application that automates and supervises the construction, simulation, and analyses of common models. Just as software design efforts have lead structure solution methods like X-ray crystallography [3] and cryo-EM [219] to become common methods used in support of answering biological questions, we hope that eventually MD can also become a standard(ized) tool in the toolbox of anyone hoping to investigate biological questions. The guiding design principle for this code is: "don't be arbitrary". Since there are many often redundant molecular dynamics tools, our code is designed to resemble a software framework in which codes interact in such a standardized way that one can easily swap in different components. To this end, we have designed a modular framework with a few, strict rules for making connections with other codes and datasets.

4.2.4 The BioPhysCode method

Given the design constraints outlined above, we have developed the following software framework for organizing simulations from start to finish. Software frameworks are less a strict set of syntax rules that a coherent design philosophy. For that reason, we will describe the framework by outlining some key features of the BioPhysCode workflow and how this differs from a typical workflow described earlier.

4.2.4.1 Software dependencies

The pipeline begins with with either a basic linux workstation that has some pre-compiled scientific software (GROMACS) or a Docker container. Dockerfiles (currently available upon request from the authors) can be used to reliably generate the software environment from a minimal linux installation, and they can serve as instructions for users who are setting up new computer systems. We minimize the number of system-level packages users must install; most users need only to install (1) GROMACS, which has detailed compilation instructions[5] and benefits from being compiled from source because this optimizes the code to your hardware, and (2) mod_wsgi, which is a server interface that allows python-django to function. Both packages are found in many linux repositories and on homebrew for mac. The remainder of the pipeline, specifically Python, required modules, and linux packages necessary for serving the web interface, are all managed by the factory, which programatically installs Anaconda with dependencies. An important design principle is that with the two exceptions listed above, we choose never to formally "install" software, but instead to build a local software environment. This allows users to run a factory instance with no worry of conflicting versions of or dependencies on installed system packages.

Users clone the factory from github, download a copy of Anaconda, and run GNU `make`. Users interact with a customized `makefile` to run the factory, automacs, and omnicalc codes. Each instance is entirely local, and the use of Anaconda ensures that users do not need superuser permissions to install software (with the rare exception of opening ports for serving public websites). The factory maintains a list of installed software. Users that add new calculations to the calculator can share customized lists with other users so there is no ambiguity about which software is required to complete an analysis.

4.2.4.2 Managing the datasets

Once the factory installs the required Anaconda environment, users can modify a standard connection file written in YAML. These connection files centralize all of the typically hard-coded paths for a single computational project, including: where to store new simulations, how to import preexisting simulation data, which analysis packages available on github to use, and even which subsets of these analyses to activate (so that one analysis package can be used on multiple projects), and even how to serve the interface (i.e. which ports). Users then "connect" a project with a single command (`make connect`) and have the option to non-destructively reconnect whenever they update these settings. This is particularly useful when datasets move to other disks, when simulations must be merged into the current dataset (each project can contain multiple distinct data sets). The default connection is designed to parse simulations generated with automacs (described below), but the connection file specifies a set of regular expressions which allow users to parse existing datasets without modifying them on disk.

[5] manual.gromacs.org/documentation/2016/install-guide/index.html

4.2.4.3 Running simulations

Once a project is successfully connected, new simulations can be generated programatically via recipes that are packaged as `git` repositories that function with automacs. These recipes are easy to customize in a simple text editor. The factory also exposes their parameters to the graphical interface so users can rapidly set system size, composition, starting structures, and many other parameters. Most automacs functionality is written in plain Python, which means that users who wish to make novel simulation construction procedures only need a basic understanding of GROMACS (or the underlying integrator) and python. Automacs is nothing more than a program designed to run a series of GROMACS commands automatically, but it also includes explicit log files for each step. These logs can be used to reproduce an automacs simulation without using the automacs codes. Developers can make new simulation routines on-the-fly, without restarting, by continuing a stalled simulation after modifying the code. Most importantly, automacs is highly modular, reusing many common functions and parameters. For example, there is a single equilibration function which allows users to define an arbitrary sequence of integrator parameter changes before running the production simulation. There are also built-in tools for building bilayers that work equally well for atomistic and coarse-grained systems. Using more generic, modular functions enables users to adopt previous methods more easily. Automacs scripts are meant to be human-readable and only call on python functions like “minimize” or “solvate”. The interface to GROMACS is also rather independent from the data structures that govern simulation protocols. This means that any future changes in the structure of command line calls in GROMACS can easily be harmonized with automacs, and also that it is possible to extend simulation protocols to other integrators or molecular simulation software like NAMD or LAMMPS.

Automacs simulations are saved in a simple data structure, subdivided by modular simulation steps, that the factory can automatically parse. It is important to note here that the factory serves three primary purposes: it manages the connection files for multiple projects, installs and updates the software environment provided by Anaconda, and it runs the graphical user interface to automacs and omnicalc. It does so with almost no customized message passing. Instead, it simply executes terminal commands via make that more experienced users might do on their own. In that sense, the Django-based factory interface can easily be replaced with a more advanced option in the future, particularly since many graphical interface and web-based tools evolve quickly. Factory clones a fresh copy of automacs for each new simulation, so it is always current and can be easily expanded. It also reads arbitrary “experiment” files written in the automacs style, meaning that any new simulation procedures will automatically have their settings exposed to the graphical interface.

4.2.4.4 Performing analysis

Data generated from automacs are automatically available in the calculator package, called omnicalc. Users can also merge and import previous datasets as long as they can formulate a regular expression to identify their directories and files. Omnicalc is designed to read and sub-sample large GROMACS datasets automatically and also accepts data from NAMD. Calculations proceed by applying python-based analysis functions to groups of simulations.

All user defined parameters, cutoffs, simulation names, and even plot labels can and should be set in a single YAML file (the metafile) for easy readability and reproducibility. Each calculation function must package data in a manner suitable for long-term storage via the binary HDF5 format [240], which saves disk space and load times. This is the only constraint on how calculations can be performed, but ensures that results are always accessible to omnicalc. Calculations are never repeated once they are successful, and previous results can be loaded into a "downstream" calculations in order to perform a chain of calculations. This allows analyses to be built incrementally, and easily divided into small parts if they are processor- or memory-intensive.

Analysis codes can use a number of trajectory readers which are not fixed by omnicalc. The authors tend to use MDAnalysis [165] however it is easy to read simulations natively in python, or use another package such as MDtraj [161]. Completed calculations can be summarized using plotting software like matplotlib [108] in a customized interactive python terminal which allows users to edit the code without reloading the data. All plotting scripts are convertible to interactive Jupyter notebooks [191], which can be served within the factory GUI and then customized for publication. This feature also allows easy sharing and manipulation of plotting among collaborators. Omnicalc includes a number of utility functions for saving large libraries of similar images with additional metadata so they are easy to sort and filter. The factory GUI also has access to this metadata, allowing users to easily select figures based on parameters set at the time of plot creation via the YAML metafile. Each omnicalc instance is in bijection with a single `git` repo. The reason for this is that while one omnicalc `git` repo can contain multiple calculation and plot functions which manipulate data in specific ways, there is no reason to expect all developers to adhere to a strict data structure (aside from the calculation data being HDF5 compatible). Omnicalc only coordinates the execution of these functions and the organization of the data on disk. If a user wants to analyze data with multiple omnicalc `git` repos, the factory connection file can be set up so that different omnicalc instances have access to the same underlying data.

By abstracting the analysis codes into simple functions, it is also easy for users to understand how the authors have analyzed their data so they can implement these methods outside of the BioPhysCode framework. This framework also allows for the development of complex ecosystems of analysis codes that have been published in the literature, similar to e.g. Kepler [262] or Taverna [130]. The authors hope that this can help to both make molecular simulation and analysis procedures more reproducible, but also a more tightly integrated part of answering biological questions.

4.3 Results

4.3.1 System setup

4.3.1.1 The starting protein structure

In order to learn about the biophysical properties of a protein system the user must first select a 3-dimensional protein structure. This structure can itself be the result of a biophysical measurement such as X-ray crystallography, NMR, cryo-EM, or can be the result of a previous modeling effort. Depending on the structure determination method used to obtain the initial structure, there may be missing residues that need to be filled in, multiple chains from which a subset needs to be selected, heteroatoms that need to be removed, or mutations that were made in the course of structure determination that need to be reverted back to the wild type residue. Automacs, at the command line, and Factory at a GUI, have a built in set of protein homology tools to address these problems. The Protein Data Bank [19] (PDB[6]) has a standardized format to report missing residues, as well as sequence conflicts between the PDB chain sequence(s) and the sequence found in the Uniprot database [51], as well as differences in numbering of the sequence. Automacs homology tools can parse this data, renumber residues to match the sequence database (or any numbering the user desires), and report sequence conflicts. The user may also select which chains, which set of residues in each chain, and any point mutations to include in the final structure. The homology tools will also determine if any residues are missing from the structure and use Modeller [216] to ensure that the final structure is complete. Homology modeling can also be used to make mutations in existing complete protein structures. It is also possible to convert an all atom structure into a coarse-grained (CG) structure based on the Martini force field [158, 170] Once a user is satisfied with their prepared starting protein structure, they can proceed to further setup steps or to running simulations.

4.3.1.2 Building a bilayer

There are a number of ways to construct a membrane bilayer system, and efforts to automate these procedures have resulted in the Membrane Builder [119] package which works with CHARMM-GUI [120].

A common method is to programatically place lipids on either side of a plane normal to the Z-axis of the simulation box that represents the midpoint of a bilayer and then perform careful equilibration while gradually releasing constraints on Z-axis motion of the lipid, or lipid head groups, while leaving XY motion unconstrained. As this procedure can be computationally demanding for even systems of a few hundred lipids, another possibility is to equilibrate a relatively small lipid solvent system and then to replicate this system periodically in the X and Y directions until the desired system size is attained. It is important to reinitialize particle velocities so that the system does not display anomalous long-range order. This larger system composed of pre-equilibrated bilayer can then be eqilibrated somewhat more rapidly than a similarly sized system that has not undergone any equilibration.

While the above methods can in principle be used for all atom (AA) or CG systems, there is another method available that is primarily computationally feasible for CG bilayer systems. This method involves randomly placing the desired number of lipids for a bilayer in random positions in a water box and running the simulation long enough for the lipids to spontaneously assemble into a bilayer. One disadvantage of this method is that it does not allow for explicit selection of bilayer asymmetry or geometry.

These methods of bilayer construction, for both AA and CG systems are available in automacs. Users may select the number of each type of lipid to use, grid spacings for lipid placement, initial distances between leaflets, membrane geometries such as flat, curved, or wavy, and even different numbers of lipids in each leaflet, possibly resulting in membrane

[6] rcsb.org

curvature. The modular nature of automacs allows for complex bilayer setup protocols to be rapidly prototyped and deployed. For instance, in the case of equilibrating a small bilayer patch and then replicating this in the XY plane, the steps can be broken down in such a way that if any step in the process fails, the procedure can be restarted at that step without having to start at the beginning.

4.3.1.3 Adding a protein to a bilayer

Adding a protein or proteins to a bilayer can be an especially challenging procedure. This is because not only must the protein be placed in the correct orientation on or in the correct leaflet(s) of the membrane bilayer, but also because this placement often necessitates removal of some lipid molecules. Automacs has a number of procedures to allow users to specify the orientation of proteins relative to the bilayer. Users can specify the orientations of proteins relative to principle protein and bilayer axes, as well as specifying distances of individual residues from the bilayer. As even this rather detailed level of control can still lead to equilibrated systems in undesired configurations, the modular and multistep nature of such complex procedures means that only a single equilibration step needs to be rerun. This allows for, for instance, a detailed investigation of how rapidly to release restraints on protein orientation, with all trials starting with the same equilibrated bilayer and leaving behind a record of both the exact procedure used and the resulting biomolecular system configuration.

4.3.1.4 Polymer systems

Automacs also has procedures for setting up polymer melt and gel systems. This allows users to simulate systems of sugar n-mers constructed using either a lattice or non-lattice based initial configuration. This is of course dependent on the existence of parameters for sugar monomers and dihedral angles for their connectivity. Users can also collect statistics on angle distributions of shorter n-mers to parametrize longer ones.

4.3.2 Simulation

Automacs is a command line interface for GROMACS, and thus the simulation of systems constructed with automacs relies primarily on calls to the GROMACS (or potentially some other) integrator. Nonetheless, there are a number of ways in which automacs eases this process. The primary function of automacs in running simulations is to allow users to globally set molecular simulation run parameters, or to locally modify these parameters, while also keeping a concise log file of all such parameters. One of the design philosophies of the BioPhysCodes is that users should have the *ability* to minutely control the execution of the codes, but they should not have a *necessity* for such minute control most of the time. A number of MD execution parameters can be set, including which version of GROMACS to use, via Linux `module` commands, how many processors and cores to use, and how long in system or simulation time to run the simulation for. These parameters, and their values for a number of XSEDE compute resources [245] that the authors have found useful can be stored in a single, globally accessible file, or in a local version. Automacs also includes headers needed for execution on these XSEDE compute resources alongside functionality to upload and download relevant files to a cluster. It is easy to add execution parameters for novel compute clusters, allowing utilization of local compute resources or extension to national or international compute resources beyond XSEDE. Also, the fact that BioPhysCodes can be utilized in a docker environment allows for the use of commercial compute resources such as Amazon Webservices. Taken together, this means that a user can upload files, start a simulation, and fetch the results with only a few commands. This not only allows experienced users to focus on more demanding tasks like system construction and analysis, it allows newer users to begin simulations even if they are relatively unfamiliar with the use of compute clusters.

4.3.3 Analysis

4.3.3.1 Access to the data

Once the desired number of simulations has been run for the desired amount of simulation time, and all data has been consolidated such that it is on a storage volume accessible to omnicalc, analysis procedures can be run. We would like to stress that omnicalc is designed to be very flexible in terms of how the underlying molecular simulation data is generated. Just as is the case with automacs, the code that loads data for omnicalc analysis is relatively independent of the bulk of the omnicalc codes. The reason for this is that omnicalc crucially serves to organize data and expose it to the user for ease of analysis, and this should be independent of how the data is stored on disk or the file format of the data as long as an appropriate trajectory reader in MDAnalysis [165], MDtraj [161], or any other package is available. This means that data does not have to be generated with automacs for it to be used by omnicalc or the factory GUI; heterogeneous datasets, even comprising results in different binary formats, can be incorporated into a single dataset. While any data generated by a factory instance via automacs will automatically be available to omnicalc, users can point to simulations in arbitrary directory structures by writing regular expressions to help automacs parse the data.

4.3.3.2 Selection of data to process

Once simulation data has been exposed to automacs, the user may specify how to process this data. These specifications are outlined in a YAML file, known as the metafile since it contains all the relevant metadata, which passes on the user desired parameters to omnicalc. Before any calculations can be carried out, the underlying MD data must first be subsampled to obtain ‘slices’ of each trajectory on which to perform calculations. Users can specify: start and end simulation times, how many frames to skip in between recorded frames, which molecules or molecule types from the simulation to keep, and how to handle periodic boundary conditions for the molecules of interest. Once all slices have been made, calculations can be performed on these slices, and all metadata relevant to the construction of each slice is stored so that there is no ambiguity surrounding data generation. If a factory instance is used to generate data, a template metafile can automatically be generated for slice creation.

4.3.3.3 Performing calculations

Calculations are performed on individual slices, allowing for an easy parallelization since the results of a calculation involving one slice can never affect another calculation. If a single calculation with different parameters needs to be performed (e.g. with varying distance cutoffs), this can also be specified in the YAML-based metafile. All calculation parameters are also specified in the metafile, and these parameters are again propagated to calculation results to eliminate ambiguity. Calculations can either be performed on the simulation data contained in the slices or can be performed on the results of previous calculations, allowing for modular and memory efficient codes to be written. All calculation results must ultimately be made compatible with the HDF5 data format, to ensure easy sharing of data across different machines but also to ensure that omnicalc will always be able to read calculation data. It is often not necessary to write calculation scripts from scratch as packages such as MDAnalysis [165] and MDtraj [161] have previously implemented many simulation analysis tools, and BioPython has many tools for analyzing PDB structures, which can easily be generated from frames in a slice. Finally, since omnicalc can automatically pull from a `git` repo, it is easy for authors to share analysis (and plotting) codes with the larger molecular simulation community.

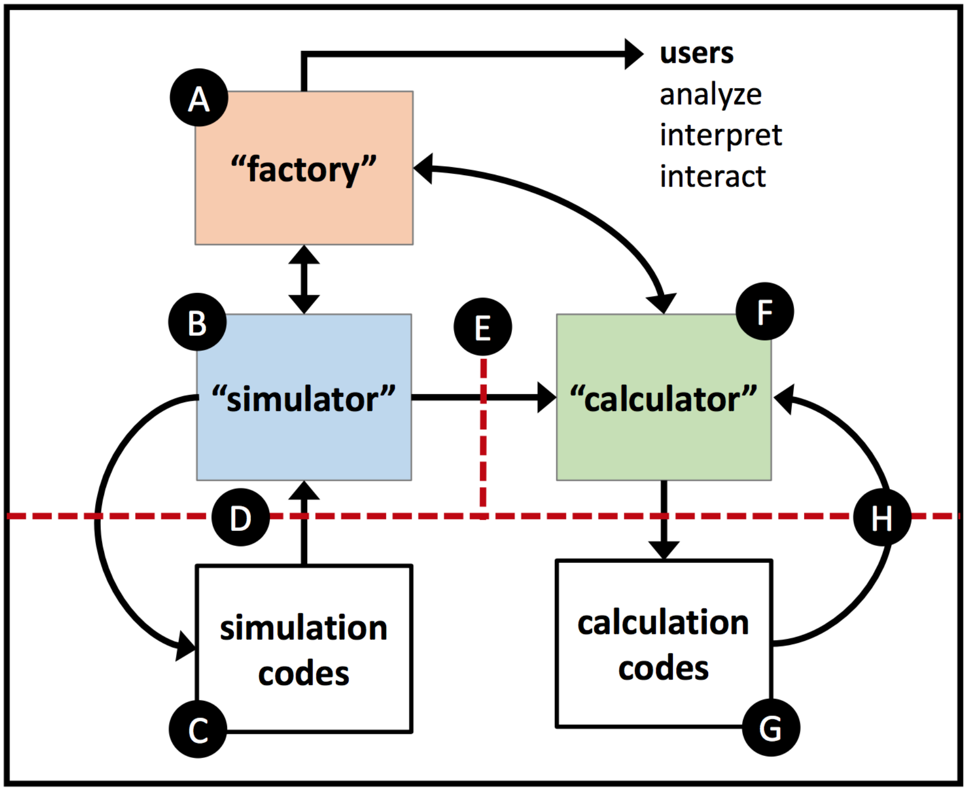
4.3.3.4 Visualization

Once simulation data has been adequately processed by calculations, users can plot the data in a number of ways. Since calculations can only act on a single slice, plots are the only way to aggregate data from multiple simulations. In practice this means that some calculations may be necessary as part of a 'plot' function. However, plotting data is inherently a form of data compression (n.b. a plot can be a few kilobytes in size yet display information from 100s of gigabytes of simulation data), so the philosophy of omnicalc is that data for use by a plotting function should already be small enough to fit in memory. Just as in the case of selecting data for processing and performing calculations, all data needed for plotting should be stored in the YAML metafile to ensure that plots have metadata which details how they were created. If the user is performing analysis with the aid of factory, any plot scripts can be exported to a JuPyter Ipython notebook [191] that allows for interactive plot development. This can be useful for collaborative efforts, as it is then easy for collaborators to see the results of plotting using, for instance, different distance-based cutoff values. Omnicalc also allows for a detailed inspection of the underlying calculation data, which is exposed to the user after the plot script is executed; prototyping can occur in a JuPyter notebook or as part of an Ipython [191] session. Another feature of the factory that can be used to aid collaboration but which also allows for detailed comparison of different plotting methods and styles is that the factory can generate lower resolution thumbnail images which can then be displayed all on a single page. Since the factory is based on a Django interface, this can also be served over the web.

4.4 Conclusion

The history of all hitherto existing MD interfaces is the history of the class (or object) struggle. Expert and novice, undergrad and post-doc, stood to have to write a new interface for each question to which they wished to apply molecular dynamics. Here we have outlined a flexible, modular, extensible interface to allow for easy setup, execution and analysis of molecular dynamics simulations. We hope that other members of the MD community will find it as useful for their own research as we have for ours and that the BioPhysCodes can help make MD experiments easy to set up, analyze, and share.

Figure 4.1: Schematic of the framework.



The factory (A) acts as a graphical user interface for the simulator and calculator and also manages the software dependencies, locations of the data, and configuration of sub-modules. The factory manages two separate codes which perform the simulation and calculation steps of a typical experiment. The simulator (B) is designed to call the GROMACS molecular dynamics integrator (C), however the interface to these codes (D) is concise and depends only on a robust (Python-BASH-binary) call, hence it can easily be extended to other popular integrators and even interfaced with other simulation methods to create a hybrid simulation. Trajectories are organized and stored in their native formats in a structure that makes it easy for the calculator to parse (E). The calculator (F) contains multiple methods for reading simulation data which are appropriate for the integrator and also calls external libraries (G) for performing the computations. At the end of this pipeline, calculations are stored in a standardized, durable binary format so that they are computed once and retrieved for more detailed analysis, plotting, and visualization.

Chapter 5

Svm

> The trick is to follow the line that links the experience of concrete situations in everyday life to the spectacular falsification of totality.
>
> McKenzie Wark *The Spectacle of Disintegration*

5.1 Introduction

5.1.1 Application of machine learning to protein sequences

The technological developments that lead to the sequencing of the human genome at the end of the 20th century [49] lead to increasingly focused efforts to understand human genetic diversity via efforts such as the 1000 genomes projects [50] and eventually to a regime where targeted and even whole genome sequencing (WGS) are routinely used in the clinic as a routine part of cancer treatment [128] This in turn has lead to a significant accumulation of cancer mutation data and to the curation of cancer mutation databases such as the Catalog of Somatic Mutations in Cancer (COSMIC) [72] and The Cancer Genome Atlas (TCGA)[1]. Investigation of these mutations has had a number of effects.

A significant role has been ascribed to the role of mutations in cancer (see chapter 1). The increasing understanding of the function of individual kinase domain mutations in particular (see chapter 2) has lead to the development of targeted therapies for treatment of specific mutations and a sustained effort to develop computational tools that can predict the effect of point mutations in cancer. One widely used computational method, which has only rarely been used to *predict* the effects of point mutations on protein function [24], is molecular dynamics (MD) simulation of protein dynamics. The literature around the use of MD to understand biophysical repercussions of kinase domain mutations is reviewed in chapter 3. One major limitation to using MD to study large numbers of mutations in cancer has been the difficulty of setting up, running, and analyzing large numbers of simulations. The transcendence of this limitation by the development of an integrated MD workflow is reported in chapter 4. The other principle limitation of using MD to study the large number of clinically observed mutations (over 21,000 unique protein coding substitutions in COSMIC, see chapter 2) is that the computer processing power required to study even

[1] cancergenome.nih.gov/

a single mutation is on the order of thousands of processor hours. For this reason, and since computer processing power was even more limited than today at the beginning of the 21st century when genome sequencing became widespread, much work has gone into finding computationally inexpensive ways to classify large numbers of mutations.

The earliest attempts to understand how well sequence changes would be tolerated was not in the context of cancer, but instead was used for understanding evolutionary distances between sequences. These methods give probabilities of mutation frequencies based on phylogenetic trees [54] or sequence alignments [97]. These methods, while innovative when developed, were not designed to predict the effect of mutations on protein function. One of the first methods to predict whether a mutation would be deleterious, and still a benchmark in the field of mutation classification, is called sorts intolerant from tolerant (SIFT) and uses sequence conservation to determine deleteriousness [175, 174]. Since this pioneering method, several other algorithms that use sequence conservation or homology to predict the effects of SNPs [46, 234, 25, 211, 4]. While these methods should in principle work on any observed mutations, they have largely been developed and validated for use on SNP data and not on cancer mutations specifically. As cancer genome sequences have become more available, the desire to separate driver from passenger mutations has only increased. An early attempt to solve this problem was that of [83], which used the mutation rate of noncoding genomic regions as a baseline and then tried to determine genes in which there was a statistically significant deviation from this baseline. More recently, several groups have developed machine learning techniques to separate driver from passenger mutations. Methods used include random forest [127, 200], entropic methods [212], and support vector machines (SVM) [117, 244].

The SVM technique is a machine learning method that falls under the broad category of supervised machine learning (see below for a detailed description of how SVM works). This method has enjoyed some success in attempting to classify kinase somatic cancer mutations, but has generally been applied to the whole protein as opposed to only the kinase domain, causing one of the leading predictors of driver status to be location within the kinase domain [117] in one predictor. While this is useful for determining which proteins are likely to be involved in cancer, it is not necessarily accurate at the residue level, for e.g., to accurately predict the effects of different mutations in the same protein. In particular, the SVM methods listed here have focused on kinase proteins specifically since they play such an outsized role in cancer progression. Machine learning methods can be quite powerful and the have been somewhat successful when applied to cancer mutations but these methods are generally only as good as their training sets, with a balanced training set giving better results [260, 259]. Furthermore, the methods that are used to construct training sets may be biasing the results of the methods. For instance, [117] took all mutations that were found in a cancer sample to be driver mutations. In a comparison of various methods for cancer mutant classification, [79] took every mutation that is observed to be mutated at least twice in COSMIC to be a driver mutation. The method outlined in [34] reports to have the ability to differentiate driver mutations from passengers 98% of the time but uses a dataset where driver mutations are taken to be any mutations that are observed in COSMIC and passenger mutations are taken from a synthetic dataset of computationally generated mutations of unknown function. The reliability of a method should be suspect if it is making a priori decisions about what is a driver or passenger mutation in a bid to predict driver or passenger mutations.

5.1.2 Support vector machines

A general problem in (supervised) machine learning can be formulated as being given a labeled set of examples, each associated with a number of features, and trying to learn a function which can distinguish between the different labels based solely on the features [249, 29, 10]. In mathematical notation, the examples have labels $y_i \in \{-1, +1\}$, $i = 1, ..., n$ with associated feature vectors $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, ..., n$ where the number of features is equal to the dimension, d, of the vector $\mathbf{x}_i$ and the total number of examples is n. The goal then is to find a classification function $f : f(\mathbf{x}_i, \alpha) = y_i,\ \forall\, \mathbf{x}_i, y_i$ where α are a set of parameters, that minimizes what is known as a loss function,

$$loss = \sum_{i=1}^{n} |y_i - f(\mathbf{x}_i, \alpha)|. \tag{5.1}$$

An SVM solves the problem posed in Equation 5.1 by relying on a simple insight from geometry, namely that any set of labeled points $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d,\ y_i \in \{-1, +1\},\ i = 1, ..., n\}$ that are linearly independent (i.e. are not found along a line) can be separated on the basis of the labels by a line in $\mathbb{R}^{d-1}$. In the case where $n <= d$ and the $\mathbf{x}_i$'s are linearly independent, this solution can be found by solving the system of linear equations $\mathbf{X}\mathbf{w} = \mathbf{y}$ where $\mathbf{X}$ is the data matrix composed of the $\mathbf{x}_i$'s, $\mathbf{y}$ is the vector composed of the $\mathbf{y}_i$'s, and $\mathbf{w}$ is a line (or hyperplane; we don't call a line or a point a *hypo*plane though) that separates the data based on the labels. This is equivalent to linear regression.

In real world problems it is often the case that there are more examples in the data than there are features, $n > d$, and in this case it may not be possible to find an exact solution. There are two principle solutions to this problem that are taken advantage of by SVM. One is to impose a cost on points which are miscorrectly classified. In this case the function f will take the form $f(\mathbf{x}_i, \alpha) = \langle \mathbf{w}, \mathbf{x}_i \rangle + C\xi$ where $\xi_i = |\langle \mathbf{w}, \mathbf{x}_i \rangle|$ is the distance from an incorrectly classified point to the line $\mathbf{w}$ and C is a penalty selected by the user. The loss function will then be

$$loss = \sum_{i=1}^{n} |y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle + C\xi_i|. \tag{5.2}$$

A second method to overcome the problem of $n > d$ is to use a simple trick: mapping the data into an arbitrarily high dimension so that once again $d > n$. These mappings are based on the fact that $\mathbf{w}$ is the solution to an eigenvalue problem for the matrix of inner products between the data points. The matrix of inner products between the data points is known as the Gram matrix, $\mathbf{G}$, and has elements $\mathbf{G}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. This means that any mapping that preserves distance relationships can also be used to solve for $\mathbf{w}$. Note also that the Gram matrix is invariant under rotations. While a number of potential mapping functions exist, the one that has seen the most widespread use in machine learning is a Gaussian radial basis function (RBF) of width γ:

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = e^{-\gamma ||\mathbf{x}_i - \mathbf{x}_j||} \tag{5.3}$$

Since the values of Equation 5.3 are based on the difference between points, this means that RBF kernels are translation invariant.

To conclude, SVM classifiers have many excellent qualities and are relatively easy to interpret when compared to other methods such as artificial neural networks [227]. These excellent qualities include that the solution of an SVM problem is invariant to translation and rotation and only dependent on the distance between training examples. This allows for preprocessing of data without affecting the final result. An SVM can in principle classify an arbitrary number of points perfectly based on a limited number of features. There are only two parameters in an SVM using RBF's. The error penalty C in Equation 5.2 controls how smooth the decision surface is, with larger values of C leading to an increasingly jagged boundary that attempts to classify every example correctly. The Gaussian width γ in Equation 5.3 controls how large of a region in feature space (or any mapping of feature space) that the training examples take up, with larger values meaning training examples are 'felt' in a smaller region. Both C and γ can be tuned in cross-validation.

5.2 Methods

5.2.1 Construction of training set

The dataset was constructed via text mining of the Uniprot database [52] using a perl script. Regular expressions were used to parse the MUTAGEN and VARIANT fields in Uniprot (unfortunately the MUTAGEN field is no longer included in the database). Mutated residue entries in Uniprot were classified as non-activating if they contained any of the following stings: 'impairs', 'strongly impairs', 'reduce','strongly reduce', 'abolishes', 'diminished', 'loss.+normal.+order' (where .+ denotes at least one other character of any type), and 'abolishes down-regulation'. Mutated residue entries in Uniprot were classified as activating: 'increase', 'strongly increase', 'constitutive', and 'does not.+constitutive'. The resulting training set was validated by searching the literature for a subset of the entire dataset to ensure that class assignments were correct. This not only showed the utility of the underlying method, but led to many papers that had mutations not in the original set in addition to those in the original set. Final set used in this work contains 756 total mutations with 192 positive, activating, and 564 negative, non-activating mutations (compare to the Kinase mutation experimental measurements).

5.2.2 Creation of feature vectors

For each mutation, a feature vector of the following values was constructed. This leads to a feature vector for each mutation with 68 elements. Each element of the resulting vectors is normalized so that all values are in [-1,1]. A large number of the elements will be zero for each mutation.

1. Wild type residue (one feature element for each of the 20 amino acids)
2. Mutant residue (one feature element for each of the 20 amino acids)
3. Wild type residue type (from aliphatic, acidic, basic, aromatic, and polar)
4. Mutant residue type (from aliphatic, acidic, basic, aromatic, and polar)
5. Difference between wild type and mutant residue for the following

   1. Kyte-Doolitle hydropathy [142]
   2. Free energy of solvation [38]
   3. Normalized van der Waals radius [67]
   4. Polarity difference [81]
   5. Charge difference

6. Difference between mutant residue and the average of all wild type residues at the same position for the following

   1. Kyte-Doolitle hydropathy
   2. Free energy of solvation
   3. Normalized van der Waals radius
   4. Polarity difference
   5. Charge difference

7. Whether the mutation falls in one of the following kinase subdomains

   1. nucleotide binding loop
   2. αC helix
   3. catalytic loop
   4. activation loop

8. PHD [215] prediction for the following

   1. α helix
   2. β sheet
   3. turn
   4. solvent accessibility

5.2.3 SVM parameter search

For this study, SVM^{perf} [121, 122, 123] was used for model construction (training) and testing. SVM^{perf}, like any ML algorithm, has a number of parameters which can be optimized. We elected to us a radial basis kernel function (RBF, see SVM introduction), and focused our efforts on finding the best values of loss function, *loss* (see Equation 5.1 and Equation 5.2), margin error, C (see Equation 5.2), and kernel width, γ (see Equation 5.3). To this end, we performed a grid search over all combinations of values of $\gamma \in [1 \times 10^{-5}, 1 \times 10^{4}]$ increasing by a factor of 10 in each iteration, $C \in \{0.01, 0.1, 1, 2, 3, 4, 5\}$ for loss functions which maximize one of {error rate, F1, AUC}. The grid search was conducted by performing 10 fold cross validation whereby for each of the 10 trials, 75% of examples are used for training and the other 25% are used for testing. During each trial, a balanced dataset was used to prevent model bias [259] by considering all examples from the smaller class and randomly selecting the same number of examples from the larger class, meaning that not every example is used in a single trial. We then average the test set results over all

10 trials for F1 and AUC. For the dataset used here, The F1 value was maximized for the parameters $\{(\gamma = 0.01, C = 4, loss = \texttt{F1})\}$. AUC was maximized for the parameters $\{(\gamma = 0.1, C = \{2,3,4,5\}, loss = \texttt{AUC})\}$. Further attempts at refining these parameters yielded only small increases in either AUC or F1, but nonetheless allowed the selection of $\{(\gamma = 0.1, C = 2, loss = \texttt{AUC})\}$ as the set of parameters which optimize AUC. Other $\texttt{SVM}^{perf}$ parameters that are not default values are {`-w 3;-# 50;-t 2;--b 0`}. In words, we used a 1-slack dual structural learning algorithm, the QP suboptimization problem is terminated after 50 iterations if no progress is made, a radial basis function was used, and no L2-bias feature was used.

5.2.4 SVM model evaluation

Once a set of SVM parameters is chosen, the resulting model is again evaluated by making predictions on a validation set. This procedure is similar to the cross-validation described above, but instead of randomly assigning examples to training and test sets, the user supplies a predetermined validation set on which predictions are made after a model is trained on the training set. For this procedure, we again elected to use a balanced dataset. Since the data we are using here is heavily skewed in favor of negative examples, we take all examples from the smaller, positive class and an equivalent number of randomly selected examples from the larger, negative class. Solving the problem of imbalance in the class sizes in this way leads to a new problem, that not all examples in the larger class are considered. We solve this problem by repeatedly training a model with different subsets of the training set and making predictions on the same validation set. We keep a running average of the prediction for each element in the validation set and call the difference between this average in subsequent iterations the residual. Once the sum of the residuals for all elements in the validation set is less than 0.01 and at least 10 iterations have been performed, the prediction is considered to have converged as subsequent iterations have negligible impact on the output validation set predictions. This procedure is mathematically valid since the data only enters the picture in the form of a Gram matrix, which is invariant to translations and rotations; since the model is trained on distances, taking the average distance should not skew the result.

5.3 Results

5.3.1 Selection of model parameters

As outlined in SVM parameter search, we attempted to find an optimal set of parameters by performing a grid search using 10-fold cross-validation. This resulted in two optimal parameter sets. The parameters which optimized the value of F1 gave values of F1=80.79 and AUC=87.43 while the AUC optimal parameters had values of AUC=88.67 and F1=79.65. Further attempts to refine the parameter values led to only small increases in F1 or AUC values, on the order (a few percent), of the differences in F1 and AUC values that occur between separate cross-validation attempts. To decide on which set of parameters to use, we performed model evaluation as outlined in SVM model evaluation by using the same dataset as training and validation set. In order to prevent biasing the resulting model towards performance on the actual validation set (the mutations studied with MD in chapter 3), we performed model evaluation at this stage using a dataset with the validation set removed. Thus on each iteration a model was trained using 152 each of positive, activating and negative, non-activating examples while predictions were performed on 152 positive and 539 negative examples. The results are given in Table 5.1.

Table 5.1: Results for separate iterations with different SVM model parameters.

Parameters	TP	FP	TN	FN	Iterations	BACC	TPR	TNR
F1	152	205	334	0	24	80.98	100	61.97
	152	205	334	0	24	80.98	100	61.97
	152	206	333	0	23	80.89	100	61.78
AUC	152	21	518	0	17	98.05	100	96.10
	152	19	520	0	20	98.24	100	96.47
	152	20	519	0	18	98.14	100	96.29

See Table 3.4 for definitions.

Of the false positives in the F1 validation, 184 occurred in all three iterations while 20 occurred twice and 13 occurred once. While it is possible that some of the false positives from the F1 validation are incorrect entries in the dataset, many of the mutations are well characterized enough to appear in the mutation table in chapter 2. This demonstrates that the method used here for ensuring a balanced dataset leads to reproducible results. For the false positives in AUC validation, 13 occurred in all three iterations while 5 occurred twice and 11 occurred once. The majority of these false positives have been experimentally characterized, with many showing up in the table of mutant kinase activities in Appendix A. Given the much smaller number of false positives using the AUC validation, we elected to use AUC further model validation.

5.3.2 SVM test set performance

In order to compare the results of SVM and MD, we made predictions on the same mutation set used in chapter 3, training the algorithm only on mutations not in this data set. The results are shown in Table 5.2. Results for SIFT [174] and Polyphen2 [4] on the same dataset are also given as a comparison for algorithms that are frequently used to predict effects of mutations.

Table 5.2: Prediction results for a number of different methods.

Method	TP	FP	TN	FN	BACC	TPR	TNR
MD	22	17	21	12	59.98	64.71	55.26
SVM	19	8	22	20	61.03	48.72	73.33
SIFT	38	26	4	1	55.38	97.44	13.33
Polyphen-2	37	27	3	2	52.44	94.87	10.00

There are more predictions for MD due to several ALK JM mutations being considered with MD. See Table 3.4 for definitions.

The results in Table 5.2 tell a number of stories. For one, the results of both SIFT and Polyphen2 are quite poor, as they predict that almost all mutations are activating. This result may be explained in both cases by noting that SIFT and Polyphen2 are designed to predict whether mutations will alter protein function at all [174, 4]; it may be the case that only a few mutations have little to no effect on protein function. Further, in the case of Polyphen2, any mutation listed as cancer related or of unknown function in Uniprot is scored as 'damaging' for the model training data. Finally, the SVM classifier performs better than the other machine learning based methods or MD based classifier.

5.3.3 SVM performance on synthetic datasets

In order to better evaluate the SVM classifier developed here, we sought to determine the predictions of the classifier on datasets comprising all protein coding alterations in COSMIC and all possible single nucleotide polymorphisms (SNP) in ALK, BRAF, and HER2. While the actual effects of most of these mutations are not known, we hope that they can serve as a proxy for a dataset biased towards (COSMIC) and against (SNP) activating mutations. For some residues in each protein no prediction was able to be made based on the SVM model, likely because of lack of training examples at these residues. In these cases, mutations were scored as negative. The results of this analysis are given in Table 5.3.

Table 5.3: Prediction results on synthetic datasets.

Protein	Dataset	Positive	Negative	Total
ALK	COSMIC	23	33	56
	all SNP	602	1026	1628
BRAF	COSMIC	45	36	81
	all SNP	601	926	1527
HER2	COSMIC	26	23	49
	all SNP	617	964	1581

As Table 5.3 shows, in all cases the number of predicted non-activating mutations is higher in the SNP set than the number of predicted activating mutations. In the case of BRAF and HER2, the number of predicted activating mutations is higher in the COSMIC mutation set than the predicted non-activating mutations. While the actual activation status of most of these variants is not known, this trend is as we would expect based on the sources of the data, providing a further validation of the SVM classifier outlined here.

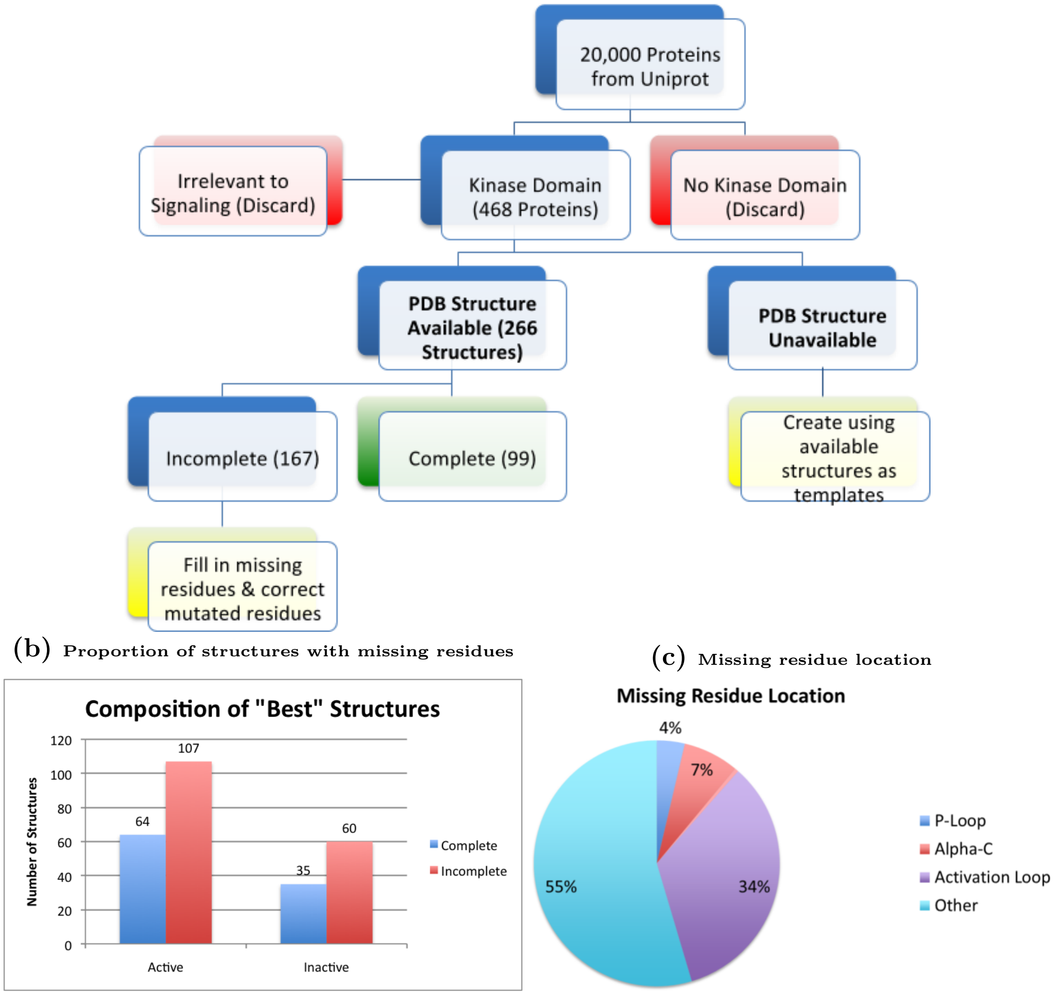
5.4 Future Directions

Recent studies have used structural information both to make predictions about the effects of mutations [35] and about the conformation of kinase domain PDB structures [162]. The work described in this chapter does not include any explicit three dimensional structural information, though it does contain information about where in the kinase sequence a mutation is found. Previous work (performed by Arjuna Keshevan under my supervision) has lead to the generation of a set of active and inactive structures for a large number of kinases. This work combined mining of Uniprot and the PDB for kinase structures and the use of Modeller [216] to model missing residues for incomplete kinase structures. This process is depicted in Figure 5.1 (a).

With this kinase structure dataset, it should be possible to improve upon the biochemical and sequence based classifier of the current SVM model. This can also allow for testing

Figure 5.1: Structural Bioinformatics

(a) Structure curation



(b) Proportion of structures with missing residues

(c) Missing residue location

of application of the results from kinase MD studies outlined in chapter 3. There, it was shown that there are a small number of conserved charged residues that largely account for the overall changes in hydrogen bonding occupancy that is well correlated with constitutively active mutations. An example of a set of features that could be used based on the combination of these insights is to have one set of features based on sequence distance of a mutated residue to an 'important' charged residue and another set of features based on distance from a mutated residue to an important residue in both active and inactive conformation of that kinase.

Chapter 6

Conclusion

Tout est possible et rien n'a d'important

Albert Camus *L'homme Revolté*

6.1 Overview of Thesis

In this thesis we have pursued a variety of avenues in investigating the effects of mutations on the activity of kinases. In chapter 1 we examined the role of kinase mutations in cancer, the structure of the kinase domain, and the problem that reification poses to modeling specifically and science in general. We then moved on to evaluate the prevalence and frequency of kinase domain mutations as well as to attempt an analysis of the extant literature around mutant kinase catalytic activity in chapter 2. With this understanding of the landscape of kinase domain mutations, we then sought to develop a molecular dynamics based methodology to evaluate the effects of kinase domain mutations for a series of mutants in ALK, BRAF, and HER2 in chapter 3. In this study, we found that there are only a few conserved hydrogen bonds that change occupancy significantly in the course of simulations of mutant and wild type protein models. Further, it was demonstrated that these residues, which we denote as 'labile' due to their propensity to alternate between hydrogen bonding partners, are useful in distinguishing between activating and non-activating mutations. In chapter 4 we demonstrated how to set up, run, and analyze larger numbers of mutations using a graphical user interface that allows for easy replication of simulation and analysis protocols. Finally, in chapter 5 we demonstrated the use of a support vector machine based kinase domain mutation classifier.

Appendix A

Kinase mutation experimental measurements

Table A.1: Experimental characterizations of mutant kinases

Kinase	K_{cat} (min^{-1}) or relative activity	$K_{m,ATP}$ (μM)	Transforming	Inhibitor(s) tested	Source
ABL WT	2.06±0.26	17.3±5.3	-	-	[85]
ABL T315I	0.28±0.25	5.64±1.96	-	-	[85]
ABL-KD WT	0.56±0.05	18.8±3.9	no	-	[85]
ABL-KD Y253F	1.06±0.1	23.1±5.0	yes	-	[85]
ABL-KD E255K	0.56±0.05	19.4±4.2	yes	-	[85]
ABL-KD T315I	0.17±0.005	12.1±0.8	no	-	[85]
ABL-KD M351T	0.1±0.01	12.6±4.7	no	-	[85]
ABL-KD H396P	0.4±0.04	18.9±4.6	no	-	[85]
BCR-ABL WT	0.71±0.04	39.6±4.9	-	-	[85]
BCR-ABL Y253F	1.87±0.22	61±14	-	-	[85]
BCR-ABL E255K	0.24±0.04	23.9±2.3	-	-	[85]
BCR-ABL T315I	0.064±0.004	19.4±3.4	-	-	[85]
BCR-ABL M351T	0.18±0.006	43.3±3.4	-	-	[85]
BCR-ABL H396P	1.04±0.39	33.5±1.4	-	-	[85]
ABL-KD WT	1	34.59±7.42	-	-	[60]
ABL-KD T291A	0.35	-	-	-	[60]
ABL-KD T291F	0.2	-	-	-	[60]
ABL-KD T291S	0.3	-	-	-	[60]
ABL-KD T291V	0.5	-	-	-	[60]
ABL-KD E294P	1.3	-	-	-	[60]
ABL-KD E294P/V299P	3.3	-	-	-	[60]
ABL-KD M297G	1.7	-	-	-	[60]
ABL-KD M297L	0.4	-	-	-	[60]
ABL-KD V299P	1.0	-	-	-	[60]
ABL-KD Y339P	0.2	-	-	-	[60]
ABL-KD Y339G	2.2	-	-	-	[60]
ABL-SH2-KD WT	1.9, 2.9, 5.1	17.42±2.30	-	-	[60]
ABL-SH2-KD D382N	0	-	-	-	[60]

AKT1 WT	1	-	-	-	[9]
AKT1 K179A	0.3	-	-	-	[9]
AKT1 S308A	0.9	-	-	-	[9]
AKT1 S308D	5	-	-	-	[9]
AKT1 WT	1	-	-	-	[125]
AKT1 Y304F	1	-	-	-	[125]
AKT1 Y315F	0.75	-	-	-	[125]
AKT1 Y326F	0.95	-	-	-	[125]
AKT1 Y315F/Y326F	0.7	-	-	-	[125]
AKT1 WT	1	-	-	-	[93]
AKT1 K158T	2	-	-	-	[93]
AKT1 K163S	5	-	-	-	[93]
AKT1 K182S	4	-	-	-	[93]
AKT1 R222N	3.5	-	-	-	[93]
AKT1 K182S/R222N	2	-	-	-	[93]
AKT1 K158T/K163S/ K182S/R222N	0.5	-	-	-	[93]
ALK WT	9.32±0.85	134±7	no	crizotinib	[24]
ALK Y1096A	27.65±6.6	-	no	-	†
ALK C1097A	17.48±7.4	-	yes	-	†
ALK F1098V	31.35±10.05	-	no	-	†
ALK G1128A	43.4±13.8	152±8	yes	crizotinib	[24]
ALK T1151M	53.4±7.3	267±18	no	crizotinib	[24]
ALK C1156Y	48.79±13.6	-	no	-	†
ALK E1161A	7.32±3.7	-	no	-	†
ALK D1163N	37.66±0.2	-	no	-	†
ALK M1166R	127±26	149±4	yes	crizotinib	[24]
ALK I1170N	200±59	297±15	yes	crizotinib	[24]
ALK I1170S	200±14	371±13	yes	crizotinib	[24]
ALK I1170V	46.57±1.8	-	yes	-	†
ALK I1171N	188±34	250±13	yes	crizotinib	[24]
ALK F1174L	365±61	127±11	yes	crizotinib	[24]
ALK F1174S	148.4±6.87	-	yes	-	†
ALK I1183T	31.5±5.6	158±18	no	crizotinib	[24]
ALK R1192P	139±33	192±7	yes	crizotinib	[24]
ALK L1196M	45.0±9.7	387±33	yes	crizotinib	[24]
ALK A1200V	11.1±0.9	208±9	no	crizotinib	[24]
ALK G1201R	no expression	-	yes	-	†
ALK L1204F	27.7±1.1	159±2	no	crizotinib	[24]
ALK R1212C	23.36±5.2	-	no	-	†
ALK P1213C	14.96±6.7	-	no	-	†
ALK R1231Q	5.35±1.05	143±15	no	crizotinib	[24]
ALK E1242K	27.31±6.0	-	no	-	†
ALK F1245C	329±65	138±1	yes	crizotinib	[24]
ALK F1245V	341±36	152±9	yes	crizotinib	[24]
ALK I1250T	2.68±0.18	150±8	no	crizotinib	[24]

ALK A1251T	0	-	no	-	†
ALK G1269A	33.76±17.6	-	no	-	†
ALK D1270G	0.923±0.306	153±19	no	crizotinib	[24]
ALK F1271L	17.86±5.0	-	yes	-	†
ALK R1275Q	119±13	326±33	yes	crizotinib	[24]
ALK Y1278A	21.4±2.9	-	yes	-	†
ALK Y1278E	27.52±10.3	-	no	-	†
ALK Y1278S	197.1±74.6	-	yes	-	†
ALK R1279Q	no expression	-	-	-	†
ALK Y1282E	37.65±1.6	-	no	-	†
ALK Y1283E	17.73±4.0	-	no	-	†
ALK G1286R	16.4±1.4	152±6	no	crizotinib	[24]
ALK T1343I	8.57±1.27	160±7	no	crizotinib	[24]
ALK D1349H	11.2±1.8	148±14	no	crizotinib	[24]
BRAF WT	1	-	-	-	[269]
BRAF T599A	0.9	-	no	-	[269]
BRAF S602A	1.0	-	-	-	[269]
BRAF S613A	0.9	-	-	-	[269]
BRAF T599A/S602A	0.7	-	-	-	[269]
BRAF T599E/S602D	7.0	-	yes	-	[269]
BRAF WT	1	-	no	-	[112]
BRAF D594V	0.06	-	no	-	[112]
BRAF F595L	3.6	-	yes	-	[112]
BRAF G596R	0.23	-	no	-	[112]
BRAF T599I	0.84	-	no	-	[112]
BRAF V600E	11.2	-	yes	-	[112]
BRAF K601E	9.0	-	yes	-	[112]
BRAF WT	1	-	no	-	[111]
BRAF R462I	0.89	-	no	-	[112]
BRAF I463S	0.83	-	no	-	[112]
BRAF G464E	0.77	-	no	-	[112]
BRAF F468C	2.5	-	yes	-	[112]
BRAF G469A	7.2	-	yes	-	[112]
BRAF G469E	0.24	-	no	-	[112]
BRAF WT	1	-	no	-	[254]
BRAF R462I	6	-	-	-	[254]
BRAF I463S	11	-	-	-	[254]
BRAF G464E	28	-	no	-	[254]
BRAF G464V	46	-	yes	-	[254]
BRAF G466A	5	-	yes	-	[254]
BRAF G466E	0.82	-	no	-	[254]
BRAF G466V	0.65	-	no	-	[254]
BRAF G469A	266	-	yes	-	[254]
BRAF G469E	1.3	-	no	-	[254]
BRAF N581	7	-	yes	-	[254]
BRAF E586K	129	-	-	-	[254]
BRAF D594V	0.32	-	no	-	[254]

BRAF F595L	60	-	-	-	[254]
BRAF G596R	0.53	-	no	-	[254]
BRAF L597V	64	-	-	-	[254]
BRAF T599I	30	-	yes	-	[254]
BRAF V600D	706	-	-	-	[254]
BRAF V600E	478	-	yes	-	[254]
BRAF V600K	162	-	-	-	[254]
BRAF V600R	244	-	-	-	[254]
BRAF K601E	138	-	-	-	[254]
BRAF WT	1	-	-	multiple	[102]
BRAF S464E	-	33	-	-	[102]
BRAF S464V	-	38	-	-	[102]
BRAF S465A	0.7	-	-	-	[102]
BRAF S465D	0.4	-	-	-	[102]
BRAF S467A	0.8	-	-	-	[102]
BRAF S467E	0.5	-	-	-	[102]
BRAF V600E	-	44	-	multiple	[102]
BRAF S602A	1.0	-	-	-	[102]
BRAF S602D	1.3	-	-	-	[102]
BRAF S605A	0	-	-	-	[102]
BRAF S607A	0.8	-	-	-	[102]
BRAF S607D	1.0	-	-	-	[102]
CHEK2 WT	1	-	-	-	[99]
CHEK2 D368N	0.3	-	-	-	[99]
CHEK2 K373E	0.4	-	-	-	[99]
CHEK2 WT	1	-	-	-	[86]
CHEK2 S372A	0.6	-	-	-	[86]
CHEK2 T378A	1.1	-	-	-	[86]
CHEK2 T378D	0.9	-	-	-	[86]
CHEK2 S379A	0.25	-	-	-	[86]
CHEK2 S379D	0.2	-	-	-	[86]
CHEK2 T383A	0.3	-	-	-	[86]
CHEK2 T383D	0.2	-	-	-	[86]
CHEK2 T387A	0.3	-	-	-	[86]
CHEK2 T387D	1.9	-	-	-	[86]
CHEK2 T389A	5.5	-	-	-	[86]
CHEK2 T389D	1.4	-	-	-	[86]
CHEK2 Y390F	0.2	-	-	-	[86]
CHEK2 T383A/T389D	0.6	-	-	-	[86]
CRAF WT	1	-	-	multiple	[102]
CRAF S357A	0.7	-	-	-	[102]
CRAF S357D	0.6	-	-	-	[102]
CRAF S359A	0.7	2	-	-	[102]
CRAF S359D	0.1	-	-	-	[102]
CRAF S494A	0.5	-	-	-	[102]
CRAF S494E	1.1	-	-	-	[102]
CRAF S497A	1.1	-	-	-	[102]
CRAF S497D	1.4	-	-	-	[102]

CRAF S499A	2.4	-	-	-	[102]
CRAF S499D	1.3	-	-	-	[102]
DAPK3 WT	1	-	yes‡	-	[82]
DAPK3 D161A	0	-	no		[82]
DAPK3 T180A	0	-	-	no	[82]
DAPK3 T180D	0	-	-	no	[82]
DAPK3 T225A	0.1	-	-	no	[82]
DAPK3 T225D	0.15	-	-	no	[82]
DAPK3 T265A	0.2	-	-	yes	[82]
DAPK3		-	-		[82]
EGFR WT	0.780	6.9±0.9	-	gefetinib	[267]
EGFR G719S	8.580	97.4±1.8	-	gefetinib	[267]
EGFR L858R	14.040	31.5±1.7	-	gefetinib	[267]
EGFR WT	1.560	5.2±0.2	-	gefetinib	[268]
EGFR T790M	8.220	5.9±0.1	-	gefetinib	[268]
EGFR L858R	89.040	148±4	-	gefetinib	[268]
EGFR T790M/L858R	27.36	8.4±0.3	-	gefetinib	[268]
EGFR WT	6.9±0.1	2.1±0.1	-	-	[66]
EGFR K745M	0	-	-	-	[66]
EGFR WT	1	4.98±1.2	-	gefetinib	[266]
EGFR L858R	33.12	65.8±3.4	-	gefetinib	[266]
ERK1 WT	1	-	-	-	[143]
ERK1 K71A	0.5	-	-	-	[143]
ERK1 T198A	0.9	-	-	-	[143]
ERK1 T207A	0.7	-	-	-	[143]
ERK1 T207E	0.2	-	-	-	[143]
ERK1 Y210E	0.1	-	-	-	[143]
ERK1 Y210F	0.2	-	-	-	[143]
ERK1 T198F/T207F	0.1	-	-	-	[143]
ERK1 WT	1	-	-	-	[32]
ERK1 H197A	1	-	-	-	[32]
ERK1 T198A	8	-	-	-	[32]
ERK1 G199A	1	-	-	-	[32]
ERK1 F200A	1	-	-	-	[32]
ERK1 F200Y	2	-	-	-	[32]
ERK1 L201A	1	-	-	-	[32]
ERK1 T202S	1	-	-	-	[32]
ERK1 E203D	1	-	-	-	[32]
ERK1 E203Q	1	-	-	-	[32]
ERK1 Y204S	1	-	-	-	[32]
ERK1 V205A	1	-	-	-	[32]
ERK1 A206V	1	-	-	-	[32]
ERK1 T207A	1	-	-	-	[32]
ERK1 R208A	1	-	-	-	[32]
ERK1 W209A	0	-	-	-	[32]
ERK1 T198A/F200Y	1	-	-	-	[32]

FGFR1 WT	1	-	-	-	[186]
FGFR1 R675G	35	-	-	-	[186]
FGFR2 WT	1	-	no	-	[77]
FGFR2 E475K	1.15	-	no	-	[77]
FGFR2 D530N	0.65	-	no	-	[77]
FGFR2 I642V	0.25	-	no	-	[77]
FGFR2 A648T	0.1	-	no	-	[77]
FGFR2 WT	1	-	-	multiple	[62]
FGFR2 F492A	2.1	-	-	multiple	[62]
FGFR2 K526E	23.0	-	-	multiple	[62]
FGFR2 V564T	1.7	-	-	multiple	[62]
FGFR2 E565K	21.0	-	-	multiple	[62]
FGFR2 R678G	43.0	-	-	multiple	[62]
FGFR2 WT	1	840±140	-	-	[40]
FGFR2 K525E	7.5	-	-	-	[40]
FGFR2 N549T	16.0	-	-	-	[40]
FGFR2 N549H	8	-	-	-	[40]
FGFR2 E565A	32	300±60	-	-	[40]
FGFR2 E565G	7.5	-	-	-	[40]
FGFR2 K641R	8	-	-	-	[40]
FGFR2 K659N	20	1540±110	-	-	[40]
FGFR2 G663E	8	-	-	-	[40]
FGFR2 R678G	9	-	-	-	[40]
FGFR3 WT	1	-	no	-	[7, 6]
FGFR3 Y823F	1	-	no	-	[7, 6]
FGFR3 D816V	0.9	-	no	-	[6]
FGFR3 WT	1	-	no	multiple	[186]
FGFR3 A500T	2	-	-	-	[186]
FGFR3 I538F	0.7	-	-	-	[186]
FGFR3 I538V	6	-	-	multiple	[186]
FGFR3 N540K	40	-	yes	multiple	[186]
FGFR3 N540S	12	-	-	multiple	[186]
FGFR3 V555M	6	-	-	multiple	[186]
FGFR3 P572A	0.5	-	-	-	[186]
FGFR3 C582F	1	-	-	-	[186]
FGFR3 D617G	0	-	-	-	[186]
FGFR3 E627D	1	-	-	-	[186]
FGFR3 V630M	0.5	-	-	-	[186]
FGFR3 G637W	0.3	-	-	-	[186]
FGFR3 D641G	6	-	-	-	[186]
FGFR3 D641N	5	-	-	-	[186]
FGFR3 H643D	1	-	-	-	[186]
FGFR3 D646Y	2	-	-	-	[186]
FGFR3 Y647C	4	-	-	-	[186]
FGFR3 K650E	44	-	yes	multiple	[186]
FGFR3 K650N	19	-	-	-	[186]
FGFR3 N653H	1	-	-	-	[186]
FGFR3 R669G	55	-	-	multiple	[186]
FGFR3 R669Q	10	-	-	-	[186]

FGFR3 V677I	1	-	-	-	[186]
FGFR3 G697C	1	-	no	-	[186]
FGFR3 WT	1	-	-	-	[272]
FGFR3 G697C	23	-	-	-	[272]
FLT3 WT	1	-	-	multiple	[42]
FLT3 D835H	15	-	-	multiple	[42]
FLT3 D835Y	15	-	-	multiple	[42]
HER2 WT	1.4±0.2	20.8±2.3	-	-	[66]
HER2 K753M	0	-	-	-	[66]
HER2 G776S	2.9±0.2	17.4±1.8	-	-	[66]
HER2 G778D	8.5±0.3	5.4±0.4	-	-	[66]
HER2 G776S/G778D	9.7±0.4	2.3±0.2	-	-	[66]
HER2 WT	1	-	no	lapatinib neratinib trastuzimab	[21]
HER2 L755S	no expression	-	no	lapatinib neratinib	[21]
HER2 S760A	1.0	-	no	lapatinib neratinib trastuzimab	[21]
HER2 I767M	1.2	-	no	lapatinib neratinib trastuzimab	[21]
HER2 D769H	5.0	-	no	lapatinib neratinib	[21]
HER2 D769Y	4.7	-	yes	lapatinib neratinib trastuzimab	[21]
HER2 V777L	22.0	-	yes	lapatinib neratinib	[21]
HER2 Y835F	0.3	-	no	lapatinib neratinib	[21]
HER2 V842I	3.3	-	no	lapatinib neratinib	[21]
HER2 R896C	2.2	-	yes	lapatinib neratinib trastuzimab	[21]
HER2 WT	7.5	-	no	neratinib trastuzimab	[131]
HER2 L755S	-	-	yes	neratinib trastuzimab	[131]
HER2 V777L	-	-	yes	neratinib trastuzimab	[131]
HER2 V842I	-	-	no	neratinib trastuzimab	[131]
HER2 L866M	23.0	-	no	neratinib trastuzimab	[131]

HER2 WT	1	-	no	lapatinib neratinib trastuzimab	[273]
HER2 K753E	1.5	-	no	lapatinib neratinib trastuzimab	[273]
HER2 L755S	-	-	no	lapatinib neratinib trastuzimab	[273]
HER2 L768S	6	-	yes	lapatinib neratinib trastuzimab	[273]
HER2 V773L	8	-	yes	lapatinib neratinib trastuzimab	[273]
LIMK1 WT	34.8±3.2	-	-	-	[63]
LIMK1 D460N	0.0	-	-	-	[63]
LIMK1 Y507F	33.3±3.6	-	-	-	[63]
LIMK1 T508V	11.8±4.2	-	-	-	[63]
LIMK1 WT	1	-	-	-	[168]
LIMK1 D460A	0.1	-	-	-	[168]
LIMK1 T508A	0.8	-	-	-	[168]
MAP3K5 WT	1	-	-	-	[27]
MAP3K5 T813A	0.5	-	-	-	[27]
MAP3K5 T838A	0.25	-	-	-	[27]
MAP3K5 T842A	0.4	-	-	-	[27]
MELK WT	1	-	-	-	[20]
MELK T56A	0.9	-	-	-	[20]
MELK T56D	1.05	-	-	-	[20]
MELK Y163F	1	-	-	-	[20]
MELK Y163D	0.95	-	-	-	[20]
MELK T167A	0.1	-	-	-	[20]
MELK T167D	0.7	-	-	-	[20]
MELK S171A	0.1	-	-	-	[20]
MELK S171D	0.15	-	-	-	[20]
MELK S253A	1	-	-	-	[20]
MELK S253D	0.7	-	-	-	[20]
KIT WT	100.12	42.5	-	sunitinib	[75]
KIT D816H	77.40	22.0	-	sunitinib	[75]
KIT D816V	33.00	17.0	-	sunitinib	[75]
KIT V560D	88.20	35.0	-	sunitinib	[75]
KIT V654A	75.60	13.5	-	sunitinib	[75]
KIT V560D/T670I	57.60	12.8	-	sunitinib	[75]
KIT WT	1	-	no	dasatinib	[235]
KIT D816V	7	-	yes	dasatinib	[235]
KIT WT	216.0±12.0	115.9±46.1	-	sunitinib	[57]
KIT Y823F	168.0±72.0	273.5±160	-	sunitinib	[57]

NEK2 WT	1	-	-	SU11652 SU11248	[209]
NEK2 K37R	0.1	-	-	-	[209]
NEK2 T170A	1.0	-	-	-	[209]
NEK2 T170E	2.4	-	-	-	[209]
NEK2 S171A	1	-	-	-	[209]
NEK2 S171D	1.7	-	-	-	[209]
NEK2 T175A	0.5	-	-	-	[209]
NEK2 T175E	1.7	-	-	-	[209]
NEK2 T179A	0	-	-	-	[209]
NEK2 T179E	0	-	-	-	[209]
NEK2 S241A	0.1	-	-	-	[209]
NEK2 S241D	0.1	-	-	-	[209]
PAK4 WT	1	-	-	purvanalol	[261]
PAK4 E329K	1.1	-	-	purvanalol	[261]
RET WT	1	-	no	-	[163]
RET M918T	8	-	no	-	[163]
RET WT	36.57±1.62	164.1±15.6	-	-	[197]
RET E734A	31.06±1.23	159.4±14.3	-	-	[197]
RET K758M	0	-	-	-	[197]
RET V804M	38.05±1.33	184.8±20	-	-	[197]
RET M918T	48.92±1.48	106.3±7.9	-	-	[197]
RET R912A	32.17±1.44	294.1±24.2	-	-	[197]
RPS6KB1 WT	1	-	-	-	[93]
RPS6KB1 K167N	0.2	-	-	-	[93]
TTK WT	1	-	no	-	[160]
TTK D664A	0	-	yes‡	-	[160]
TTK T675A	3.2	-	no	-	[160]
TTK T676A	0.7	-	yes‡	-	[160]
TTK S677A	1.8	-	-	-	[160]
TTK S682A	2.1	-	-	-	[160]
TTK T686A	0.3	-	yes‡	-	[160]
TTK Y689F	0.5	-	-	-	[160]

Note that values without an error estimate are relative to WT in the respective study. In most of these cases only graphs and not values are presented. † These ALK mutants are characterized in chapter 3. ‡ Loss of activity is transforming.

Bibliography

[1] *Cancer Facts and Figures 2017*. American Cancer Society, 2017.

[2] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS : High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1(2):19–25, 2015.

[3] Paul D Adams, W Ralf, Randy J Read, James C Sacchettini, and Nicholas K Sauter. research papers PHENIX : building new software for automated crystallographic structure determination. *Biological Crystallography*, D58:1948–1954, 2002.

[4] Ivan a Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–9, apr 2010.

[5] G. Agamben. *The Open: Man and Animal*. Meridian (Stanford, Calif.). Stanford University Press, 2004.

[6] S Agarwal, J U Kazi, S Mohlin, S Påhlman, and L Rönnstrand. The activation loop tyrosine 823 is essential for the transforming capacity of the c-Kit oncogenic mutant D816V. *Oncogene*, (August):1–10, 2014.

[7] Shruti Agarwal, Julhash U. Kazi, and Lars Rönnstrand. Phosphorylation of the activation loop tyrosine 823 in c-Kit is crucial for cell survival and proliferation. *Journal of Biological Chemistry*, 288(31):22460–22468, 2013.

[8] Fahd Al-Mulla, E. james Milner-white, James J. Going, and George D. Birnie. STRUCTURAL DIFFERENCES BETWEEN VALINE-12 AND ASPARTATE-12 RAS PROTEINS MAY MODIFY CARCINOMA AGGRESSION. *Journal of Pathology*, 187(June 1998):433–438, 1999.

[9] D R Alessi, M Andjelkovic, B Caudwell, P Cron, N Morrice, P Cohen, and B A Hemmings. Mechanism of activation of protein kinase B by insulin and IGF-1. *The EMBO journal*, 15(23):6541–51, 1996.

[10] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2nd edition, 2010.

[11] Fabrice André, Joseph Ciccolini, Jean-Philippe Spano, Frédérique Penault-Llorca, Nicolas Mounier, Gilles Freyer, Jean-Yves Blay, and Gérard Milano. Personalized medicine in oncology: where have we come from and where are we going? *Pharmacogenomics*, 14(8):931–939, 2013.

[12] Stephen C Artim, Jeannine M Mendrola, and Mark a Lemmon. Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family. *The Biochemical journal*, 448(2):213–20, dec 2012.

[13] Mohammad Azam, Markus a Seeliger, Nathanael S Gray, John Kuriyan, and George Q Daley. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nature structural & molecular biology*, 15(10):1109–18, oct 2008.

[14] Rajintha M Bandaranayake, Daniela Ungureanu, Yibing Shan, David E Shaw, Olli Silvennoinen, and Stevan R Hubbard. Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nature structural & molecular biology*, 19(8):754–9, aug 2012.

[15] Jiri Bartek and Jiri Lukas. Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell*, 3:421–429, 2003.

[16] G. Bataille. *The Accursed Share: An Essay on General Economy.* Number v. 1 in Accursed Share. Zone Books, 1991.

[17] C Glenn Begley. Raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

[18] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[19] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[20] Monique Beullens, Sadia Vancauwenbergh, Nick Morrice, Rita Derua, Hugo Ceulemans, and Etienne Waelkens. Substrate Specificity and Activity Regulation of Protein Kinase MELK. *Journal of Biological Chemistry*, 280(48):40003–40011, 2005.

[21] Ron Bose, Shyam M Kavuri, Adam C Searleman, Wei Shen, Dong Shen, Daniel C Koboldt, John Monsey, Nicholas Goel, Adam B Aronson, Shunqiang Li, Cynthia X Ma, Li Ding, Elaine R Mardis, and Matthew J Ellis. Activating HER2 Mutations in HER2 Gene Amplification Negative Breast Cancer. *Cancer Discovery*, 3(2):224–37, 2013.

[22] Bradley B Brasher and Richard A Van Etten. c-Abl Has High Intrinsic Tyrosine Kinase Activity That Is Stimulated by Mutation of the Src Homology 3 Domain and by Autophosphorylation at Two Distinct Regulatory Tyrosines *. *Journal of Biological Chemistry*, 275(45):35631–35637, 2000.

[23] R. Brenner. *The Economics of Global Turbulence: The Advanced Capitalist Economies from Long Boom to Long Downturn, 1945-2005*. Verso, 2006.

[24] Scott C Bresler, Daniel A Weiser, Peter J Huwe, Jin H Park, Kateryna Krytska, Hannah Ryles, Marci Laudenslager, Eric F Rappaport, Andrew C Wood, Patrick W Mcgrady, and Michael D Hogarty. ALK Mutations Confer Differential Oncogenic Activation and Sensitivity to ALK Inhibition Therapy in Neuroblastoma. *Cancer cell*, pages 682–694, 2014.

[25] Yana Bromberg and Burkhard Rost. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research*, 35(11):3823–35, jan 2007.

[26] B R Brooks, C L Brooks Iii, A D Mackerell, L Nilsson, R J Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, A R Dinner, and M Feig. CHARMM : The Biomolecular Simulation Program. *Journal of computational chemistry*, 30:1545–1614, 2009.

[27] Gabor Bunkoczi, Eidarus Salah, Panagis Filippakopoulos, Oleg Fedorov, Susanne Mu, Frank Sobott, Sirlester A Parker, Haifeng Zhang, Wang Min, Benjamin E Turk, and Stefan Knapp. Structural and Functional Characterization of the Human Protein Kinase ASK1. *Cell*, 15(October):1215–1226, 2007.

[28] Tom D Bunney, Shunzhou Wan, Nethaji Thiyagarajan, Ludovico Sutto, Sarah V Williams, Paul Ashford, Hans Koss, Margaret A Knowles, Francesco L Gervasio, Peter V Coveney, and Matilda Katan. The Effect of Mutations on Drug Sensitivity and Kinase Activity of Fibroblast Growth Factor Receptors : A Combined Experimental and Theoretical Study. *EBioMedicine*, 2(3):194–204, 2015.

[29] CHRISTOPHER J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[30] Rebecca a Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–45, sep 2013.

[31] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):14101, 2007.

[32] Elizabeth R Butch and Kun-liang Guan. Characterization of ERK1 Activation Site Mutants and the Effect on Recognition by MEK1 and MEK2. *Journal of Biological Chemistry*, 271(8):4230–4235, 1996.

[33] Renaud Capdeville, Elisabeth Buchdunger, Juerg Zimmermann, and Alex Matter. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nature reviews. Drug discovery*, 1(7):493–502, jul 2002.

[34] Emidio Capriotti and Russ B Altman. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics*, 98(4):310–7, oct 2011.

[35] Emidio Capriotti and Russ B Altman. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC bioinformatics*, 12 Suppl 4(Suppl 4):S3, jan 2011.

[36] Kendall D Carey, Andrew J Garton, Maria S Romero, Jennifer Kahler, Stuart Thomson, Sarajane Ross, Frances Park, John D Haley, Neil Gibson, and Mark X Sliwkowski. Kinetic Analysis of Epidermal Growth Factor Receptor Somatic Mutant Proteins Shows Increased Sensitivity to the Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor , Erlotinib. *Cancer research*, (16):8163–8172, 2006.

[37] David A Case, Thomas E Cheatham, T O M Darden, Holger Gohlke, R A Y Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber Biomolecular Simulation Programs. *Journal of computational chemistry*, 26:1668–1688, 2005.

[38] M. Charton and B.I Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, 99:629–644, 1982.

[39] Claire Chauveau, John Rowell, and Ana Ferreiro. A rising titan: TTN review and mutation update. *Human Mutation*, 35(9):1046–1059, 2014.

[40] Huaibin Chen, Jinghong Ma, Wanqing Li, Anna V Eliseenkova, Chongfeng Xu, Thomas A Neubert, W Todd Miller, and Moosa Mohammadi. A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases. *Molecular cell*, 27:717–730, 2007.

[41] Y-R Chen, Y-N Fu, C-H Lin, S-T Yang, S-F Hu, Y-T Chen, S-F Tsai, and S-F Huang. Distinctive activation patterns in constitutively active and gefitinib-sensitive EGFR mutants. *Oncogene*, 25(8):1205–15, feb 2006.

[42] Yun Chen, Yao Guo, Jiayu Han, Wanting Tina Ho, Shibo Li, Xueqi Fu, and Zhizhuang Joe Zhao. Generation and characterization of a highly effective protein substrate for analysis of FLT3 activity. *Journal of Hematology and Oncology*, 5(1):1, 2012.

[43] Alex Chiang and Ryan P. Million. Personalized medicine in oncology: next generation. *Nature Reviews Drug Discovery*, 10(12):895–896, 2011.

[44] Young Lim Choi, Manabu Soda, Yoshihiro Yamashita, Toshihide Ueno, Junpei Takashima, Takahiro Nakajima, Yasushi Yatabe, Kengo Takeuchi, Toru Hamada, Hidenori Haruta, Yuichi Ishikawa, Hideki Kimura, Tetsuya Mitsudomi, Yoshiro Tanio, and Hiroyuki Mano. EML4-ALK Mutations in Lung Cancer That Confer Resistance to ALK Inhibitors. *New England Journal of Medicin*, 363(18):1374–1379, 2010.

[45] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, sep 2013.

[46] R. J. Clifford, M. N. Edmonson, C. Nguyen, and K. H. Buetow. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 20(7):1006–1014, jan 2004.

[47] R. H. Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

[48] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[49] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(February):860–921, 2001.

[50] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2011.

[51] The Uniprot Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[52] The Uniprot Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[53] Amy L Creekmore, William T Silkworth, Daniela Cimini, Roderick V Jensen, Paul C Roberts, and Eva M Schmelz. Changes in gene expression and cellular architecture in an ovarian cancer progression model. *PloS one*, 6(3):e17676, jan 2011.

[54] M O Dayhoff and R M Schwartz. A Model of Evolutionary Change in Proteins. In *ATLAS OF PROTEIN SEQUENCE AND STRUCTURE*, pages 345–352. 1978.

[55] T B Deb, L Su, L Wong, E Bonvini, a Wells, M David, and G R Johnson. Epidermal growth factor (EGF) receptor kinase-independent signaling by EGF. *The Journal of biological chemistry*, 276(18):15554–15560, 2001.

[56] K.A. Dill and S. Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology.* Garland Science, 2003.

[57] P. Jonathan DiNitto, Gayatri D Deshmukh, Yan Zhang, Suzanne L Jacques, Rocco Coli, Joseph W Worrall, Wade Diehl, Jessie M English, and Joe C Wu. Function of activation loop tyrosine phosphorylation in the mechanism of c-Kit auto-activation and its implication in sunitinib resistance. *Journal of Biochemistry*, 147(November):601–609, 2010.

[58] Anshuman Dixit and Gennady M Verkhivker. Hierarchical modeling of activation mechanisms in the ABL and EGFR kinase domains: thermodynamic and mechanistic catalysts of kinase activation by cancer mutations. *PLoS computational biology*, 5(8):e1000487, aug 2009.

[59] Anshuman Dixit and Gennady M Verkhivker. Structure-functional prediction and analysis of cancer mutation effects in protein kinases. *Computational and mathematical methods in medicine*, 2014:653487, jan 2014.

[60] Nicole Dölker, Maria W. Górna, Ludovico Sutto, Antonio S. Torralba, Giulio Superti-Furga, and Francesco L. Gervasio. The SH2 Domain Regulates c-Abl Kinase Activation by a Cyclin-Like Mechanism and Remodulation of the Hinge Motion. *PLoS Computational Biology*, 10(10):23–25, 2014.

[61] Arianna Donella-deana, Oriano Marin, Luca Cesaro, Rosalind H Gunby, Anna Ferrarese, Addolorata M L Coluccia, Carmen J Tartari, Luca Mologni, Leonardo Scapozza, Carlo Gambacorti-passerini, and Lorenzo A Pinna. Unique Substrate Specificity of Anaplastic Lymphoma Kinase (ALK): Development of Phosphoacceptor Peptides for the Assay of ALK Activity †. *Biochemistry*, 23:8533–8542, 2005.

[62] Sudharshan Eathiraj, Rocio Palma, Marscha Hirschi, Erika Volckova, Enkeleda Nakuci, Jennifer Castro, Chang-rung Chen, Thomas C K Chan, Dennis S France, and Mark A Ashwell. A Novel Mode of Protein Kinase Inhibition Exploiting Hydrophobic Motifs of Autoinhibited Kinases. *Journal of Biological Chemistry*, 286(23):20677–20687, 2011.

[63] David C Edwards and Gordon N Gill. Structural Features of LIM Kinase That Control Effects on the Actin Cytoskeleton. *Journal of Biological Chemistry*, 274(16):11352–11361, 1999.

[64] Darrell L. Ellsworth, Heather L. Blackburn, Craig D. Shriver, Shahrooz Rabizadeh, Patrick Soon-Shiong, and Rachel E. Ellsworth. Single-cell sequencing and tumorigenesis: improved understanding of tumor evolution and metastasis. *Clinical and Translational Medicine*, 6(1):15, 2017.

[65] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 1995.

[66] Ying-xin Fan, Lily Wong, Jinhui Ding, Nikolay A Spiridonov, Richard C Johnson, and Gibbes R Johnson. Mutational Activation of ErbB2 Reveals a New Protein Kinase Autoinhibition Mechanism. *Journal of Biological Chemistry*, 283(3):1588–1596, 2008.

[67] JEAN-LUC FAUCHÈRE, MARVIN CHARTON, LEMONT B. KIER, ARIE VERLOOP, and VLADIMIR PLISKA. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International Journal of Peptide and Protein Research*, 32(4):269–278, 1988.

[68] P. Feyerabend. *Against Method*. Verso, 1993.

[69] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016.

[70] Irving Fisher. *The Scope and Importance to the State of National Eugenics*. 1930.

[71] Zachariah H Foda, Yibing Shan, Eric T Kim, David E Shaw, and Markus A Seeliger. A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nature Communications*, 5:1–10, 2015.

[72] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai YinKok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017.

[73] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W Teague, Michael R Stratton, Ultan Mcdermott, and J Campbell. COSMIC : exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43:805–811, 2014.

[74] Filip F Fratev and Svava Osk Jónsdóttir. An in silico study of the molecular basis of B-RAF activation and conformational stability. *BMC structural biology*, 9:47, jan 2009.

[75] Ketan S Gajiwala, Joe C Wu, James Christensen, Gayatri D Deshmukh, Wade Diehl, Jonathan P DiNitto, Jessie M English, Michael J Greig, You-Ai He, Suzanne L Jacques, Elizabeth a Lunney, Michele McTigue, David Molina, Terri Quenzer, Peter a Wells, Xiu Yu, Yan Zhang, Aihua Zou, Mark R Emmett, Alan G Marshall, Hui-Min Zhang, and George D Demetri. KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1542–7, feb 2009.

[76] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J Milano, Graham R Bignell, Ah T Tam, Helen Davies, Jesse a Stevenson, Syd Barthorpe, Stephen R Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey a Engelman, Sreenath V Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S Gray, Jeffrey Settleman, P Andrew Futreal, Daniel a Haber, Michael R Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–5, mar 2012.

[77] Michael G Gartside, Huaibin Chen, Omar a Ibrahimi, Sara a Byron, Amy V Curtis, Candice L Wellens, Ana Bengston, Laura M Yudt, Anna V Eliseenkova, Jinghong Ma, John a Curtin, Pilar Hyder, Ursula L Harper, Erica Riedesel, Graham J Mann, Jeffrey M Trent, Boris C Bastian, Paul S Meltzer, Moosa Mohammadi, and Pamela M Pollock. Loss-of-Function Fibroblast Growth Factor Receptor-2 Mutations

in Melanoma Loss-of-Function Fibroblast Growth Factor Receptor-2 Mutations in Melanoma. *Molecular Cancer Research*, 113(January):41–54, 2009.

[78] Paraskevi Gkeka, Thomas Evangelidis, Maria Pavlaki, Vasiliki Lazani, Savvas Christoforidis, Bogos Agianian, and Zoe Cournia. Investigating the Structure and Dynamics of the PIK3CA Wild-Type and H1047R Oncogenic Mutant. *PLoS Computational Biology*, 10(10), 2014.

[79] Florian Gnad, Albion Baucom, Kiran Mukhyala, Gerard Manning, and Zemin Zhang. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics*, 14 Suppl 3(Suppl 3):S7, jan 2013.

[80] Richard J Gowers, Max Linke, Jonathan Barnoud, Tyler J E Reddy, Manuel N Melo, Sean L Seyler, David L Dotson, Sébastien Buchoux, Ian M Kenney, and Oliver Beckstein. MDAnalysis : A Python Package for the Rapid Analysis of Molecular Dynamics Simulations MDAnalysis. *Proc. 15th Python in Science Conference*, (Scipy):102–109, 2016.

[81] R Grantham. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, 185(4154):862–864, 1974.

[82] Paul R Graves, Karen M Winkfield, Timothy A J Haystead, and North Carolina. Regulation of Zipper-interacting Protein Kinase Activity in Vitro and in Vivo by Multisite Phosphorylation *. *Journal of Biological Chemistry*, 280(10):9363–9374, 2005.

[83] Chris Greenman, Richard Wooster, P Andrew Futreal, Michael R Stratton, and Douglas F Easton. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–98, aug 2006.

[84] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, Sarah Edkins, Sarah O'Meara, Imre Vastrik, Esther E Schmidt, Tim Avis, Syd Barthorpe, Gurpreet Bhamra, Gemma Buck, Bhudipa Choudhury, Jody Clements, Jennifer Cole, Ed Dicks, Simon Forbes, Kris Gray, Kelly Halliday, Rachel Harrison, Katy Hills, Jon Hinton, Andy Jenkinson, David Jones, Andy Menzies, Tatiana Mironenko, Janet Perry, Keiran Raine, Dave Richardson, Rebecca Shepherd, Alexandra Small, Calli Tofts, Jennifer Varian, Tony Webb, Sofie West, Sara Widaa, Andy Yates, Daniel P Cahill, David N Louis, Peter Goldstraw, Andrew G Nicholson, Francis Brasseur, Leendert Looijenga, Barbara L Weber, Yoke-Eng Chiew, Anna DeFazio, Mel F Greaves, Anthony R Green, Peter Campbell, Ewan Birney, Douglas F Easton, Georgia Chenevix-Trench, Min-Han Tan, Sok Kean Khoo, Bin Tean Teh, Siu Tsan Yuen, Suet Yi Leung, Richard Wooster, P Andrew Futreal, and Michael R Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, mar 2007.

[85] Ian J Griswold, Mary MacPartlin, Thomas Bumm, Valerie L Goss, Thomas O'Hare, Kimberly a Lee, Amie S Corbin, Eric P Stoffregen, Caitlyn Smith, Kara Johnson, Erika M Moseson, Lisa J Wood, Roberto D Polakiewicz, Brian J Druker, and

Michael W Deininger. Kinase domain mutants of Bcr-Abl exhibit altered transformation potency, kinase activity, and substrate utilization, irrespective of sensitivity to imatinib. *Molecular and cellular biology*, 26(16):6082–93, aug 2006.

[86] Xin Guo, Michael D Ward, Jessica B Tiedebohl, Yvonne M Oden, Julius O Nyalwidhe, and O John Semmes. Interdependent Phosphorylation within the Kinase Domain. *Journal of Biological Chemistry*, 285(43):33348–33357, 2010.

[87] Jacob R Haling, Jawahar Sudhamsu, Ivana Yen, Steve Sideris, Wendy Sandoval, Wilson Phung, Brandon J Bravo, Anthony M Giannetti, Ariana Peck, Alexandre Masselot, Tony Morales, Darin Smith, Barbara J Brandhuber, Sarah G Hymowitz, and Shiva Malek. Structure of the BRAF-MEK Complex Reveals a Kinase Activity Independent Role for BRAF in MAPK Signaling. *Cancer Cell*, 26(3):402–413, 2014.

[88] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100:57–70, 2000.

[89] Douglas Hanahan and Robert a Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, mar 2011.

[90] Steven K Hanks and Tony Hunter. The eukaryotic protein kinase superfamily : kinase (catalytic) domam structure and classification. *The FASEB Journal*, 9(8):576–596, 1995.

[91] D. Haridas, M. P. Ponnusamy, S. Chugh, I. Lakshmanan, P. Seshacharyulu, and S. K. Batra. MUC16: molecular analysis and its functional implications in benign and malignant conditions. *The FASEB Journal*, 28(10):4183–4199, 2014.

[92] Tomio Hashimoto, Minoru Kato, Takeshi Shimomura, and Naomi Kitamura. TMPRSS13, a type II transmembrane serine protease, is inhibited by hepatocyte growth factor activator inhibitor type 1 and activates pro-hepatocyte growth factor. *FEBS Journal*, 277(23):4888–4900, 2010.

[93] Camilla Hauge, Torben L Antal, Daniel Hirschberg, Ulrik Doehn, Katrine Thorup, Leila Idrissova, Klaus Hansen, Ole N Jensen, Thomas J Jørgensen, Ricardo M Biondi, and Morten Frödin. Mechanism for activation of the growth factor-activated AGC kinases by turn motif phosphorylation. *The EMBO journal*, 26(9):2251–61, 2007.

[94] Sonja J Heidorn, Carla Milagre, Steven Whittaker, Arnaud Nourry, Ion Niculescu-duvas, Nathalie Dhomen, Jahan Hussain, Jorge S Reis-filho, Caroline J Springer, Catrin Pritchard, and Richard Marais. Kinase-Dead BRAF and Oncogenic RAS Cooperate to Drive Tumor Progression through CRAF. *Cell*, 140:209–221, 2010.

[95] Michael C Heinrich, Christopher L Corless, Anette Duensing, Laura McGreevey, Chang-Jie Chen, Nora Joseph, Samuel Singer, Diana J Griffith, Andrea Haley, Ajia Town, George D Demetri, Christopher D M Fletcher, and Jonathan a Fletcher. PDGFRA activating mutations in gastrointestinal stromal tumors. *Science (New York, N.Y.)*, 299(5607):708–10, jan 2003.

[96] Michael C Heinrich, Diana Grif, Arin Mckinley, Janice Patterson, Ajia Presnell, and Abhijit Ramachandran. Crenolanib Inhibits the Drug-Resistant PDGFRA D842V Mutation Associated with Imatinib-Resistant Gastrointestinal Stromal Tumors. *Clinical Cancer Research*, 18(16):4375–4385, 2012.

[97] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(November):10915–10919, 1992.

[98] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[99] Masayoshi Higashiguchi, Izumi Nagatomo, Takashi Kijima, and Osamu Morimura. Clarifying the biological significance of the CHK2 K373E somatic mutation discovered in The Cancer Genome Atlas database. *FEBS letters*, 590:4275–4286, 2016.

[100] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D.M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A. Margolin, Laura J. Van't Veer, Nuria Lopez-Bigas, Peter W. Laird, Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, Christopher C. Benz, Charles M. Perou, Joshua M. Stuart, Rachel Abbott, Scott Abbott, B. Arman Aksoy, Kenneth Aldape, Adrian Ally, Samirkumar Amin, Dimitris Anastassiou, J. Todd Auman, Keith A. Baggerly, Miruna Balasundaram, Saianand Balu, Stephen B. Baylin, Stephen C. Benz, Benjamin P. Berman, Brady Bernard, Ami S. Bhatt, Inanc Birol, Aaron D. Black, Tom Bodenheimer, Moiz S. Bootwalla, Jay Bowen, Ryan Bressler, Christopher A. Bristow, Angela N. Brooks, Bradley Broom, Elizabeth Buda, Robert Burton, Yaron S.N. Butterfield, Daniel Carlin, Scott L. Carter, Tod D. Casasent, Kyle Chang, Stephen Chanock, Lynda Chin, Dong Yeon Cho, Juok Cho, Eric Chuah, Hye Jung E. Chun, Kristian Cibulskis, Giovanni Ciriello, James Cleland, Melisssa Cline, Brian Craft, Chad J. Creighton, Ludmila Danilova, Tanja Davidsen, Caleb Davis, Nathan D. Dees, Kim Delehaunty, John A. Demchok, Noreen Dhalla, Daniel DiCara, Huyen Dinh, Jason R. Dobson, Deepti Dodda, Harsha Vardhan Doddapaneni, Lawrence Donehower, David J. Dooling, Gideon Dresdner, Jennifer Drummond, Andrea Eakin, Mary Edgerton, Jim M. Eldred, Greg Eley, Kyle Ellrott, Cheng Fan, Suzanne Fei, Ina Felau, Scott Frazer, Samuel S. Freeman, Jessica Frick, Catrina C. Fronick, Lucinda L. Fulton, Robert Fulton, Stacey B. Gabriel, Jianjiong Gao, Julie M. Gastier-Foster, Nils Gehlenborg, Myra George, Gad Getz, Richard Gibbs, Mary Goldman, Abel Gonzalez-Perez, Benjamin Gross, Ranabir Guin, Preethi Gunaratne, Angela Hadjipanayis, Mark P. Hamilton, Stanley R. Hamilton, Leng Han, Yi Han, Hollie A. Harper, Psalm Haseley, David Haussler, D. Neil Hayes, David I. Heiman, Elena Helman, Carmen Helsel, Shelley M. Herbrich, James G. Herman, Toshinori Hinoue, Carrie Hirst, Martin Hirst, Robert A. Holt, Alan P. Hoyle, Lisa Iype, Anders Jacobsen, Stuart R. Jeffreys,

Mark A. Jensen, Corbin D. Jones, Steven J.M. Jones, Zhenlin Ju, Joonil Jung, Andre Kahles, Ari Kahn, Joelle Kalicki-Veizer, Divya Kalra, Krishna Latha Kanchi, David W. Kane, Hoon Kim, Jaegil Kim, Theo Knijnenburg, Daniel C. Koboldt, Christie Kovar, Roger Kramer, Richard Kreisberg, Raju Kucherlapati, Marc Ladanyi, Eric S. Lander, David E. Larson, Michael S. Lawrence, Darlene Lee, Eunjung Lee, Semin Lee, William Lee, Kjong Van Lehmann, Kalle Leinonen, Kristen M. Leraas, Seth Lerner, Douglas A. Levine, Lora Lewis, Timothy J. Ley, Haiyan I. Li, Jun Li, Wei Li, Han Liang, Tara M. Lichtenberg, Jake Lin, Ling Lin, Pei Lin, Wenbin Liu, Yingchun Liu, Yuexin Liu, Philip L. Lorenzi, Charles Lu, Yiling Lu, Lovelace J. Luquette, Singer Ma, Vincent J. Magrini, Harshad S. Mahadeshwar, Elaine R. Mardis, Marco A. Marra, Michael Mayo, Cynthia McAllister, Sean E. McGuire, Joshua F. McMichael, James Melott, Shaowu Meng, Matthew Meyerson, Piotr A. Mieczkowski, Christopher A. Miller, Martin L. Miller, Michael Miller, Richard A. Moore, Margaret Morgan, Donna Morton, Lisle E. Mose, Andrew J. Mungall, Donna Muzny, Lam Nguyen, Michael S. Noble, Houtan Noushmehr, Michelle O'Laughlin, Akinyemi I. Ojesina, Tai Hsien Ou Yang, Brad Ozenberger, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Joel S. Parker, Evan Paull, Chandra Sekhar Pedamallu, Todd Pihl, Craig Pohl, David Pot, Alexei Protopopov, Teresa Przytycka, Amie Radenbaugh, Nilsa C. Ramirez, Ricardo Ramirez, Gunnar Rätsch, Jeffrey Reid, Xiaojia Ren, Boris Reva, Sheila M. Reynolds, Suhn K. Rhie, Jeffrey Roach, Hector Rovira, Michael Ryan, Gordon Saksena, Sofie Salama, Chris Sander, Netty Santoso, Jacqueline E. Schein, Heather Schmidt, Nikolaus Schultz, Steven E. Schumacher, Jonathan Seidman, Yasin Senbabaoglu, Sahil Seth, Samantha Sharpe, Ronglai Shen, Margi Sheth, Yan Shi, Ilya Shmulevich, Grace O. Silva, Janae V. Simons, Rileen Sinha, Payal Sipahimalani, Scott M. Smith, Heidi J. Sofia, Artem Sokolov, Mathew G. Soloway, Xingzhi Song, Carrie Sougnez, Paul Spellman, Louis Staudt, Chip Stewart, Petar Stojanov, Xiaoping Su, S. Onur Sumer, Yichao Sun, Teresa Swatloski, Barbara Tabak, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Barry S. Taylor, Nina Thiessen, Vesteinn Thorsson, Timothy Triche, David J. Van Den Berg, Fabio Vandin, Richard J. Varhol, Charles J. Vaske, Umadevi Veluvolu, Roeland Verhaak, Doug Voet, Jason Walker, John W. Wallis, Peter Waltman, Yunhu Wan, Min Wang, Wenyi Wang, Zhining Wang, Scot Waring, Nils Weinhold, Daniel J. Weisenberger, Michael C. Wendl, David Wheeler, Matthew D. Wilkerson, Richard K. Wilson, Lisa Wise, Andrew Wong, Chang Jiun Wu, Chia Chin Wu, Hsin Ta Wu, Junyuan Wu, Todd Wylie, Liu Xi, Ruibin Xi, Zheng Xia, Andrew W. Xu, Da Yang, Liming Yang, Lixing Yang, Yang Yang, Jun Yao, Rong Yao, Kai Ye, Kosuke Yoshihara, Yuan Yuan, Alfred K. Yung, Travis Zack, Dong Zeng, Jean Claude Zenklusen, Hailei Zhang, Jianhua Zhang, Nianxiang Zhang, Qunyuan Zhang, Wei Zhang, Wei Zhao, Siyuan Zheng, Jing Zhu, Erik Zmuda, and Lihua Zou. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[101] Eran Hodis, Ian R Watson, Gregory V Kryukov, Stefan T Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, Daniel Auclair, Liren Li, Chelsea Place, Daniel Dicara, Alex H Ramos, Michael S Lawrence, Kristian Cibulskis, Andrey Sivachenko, Douglas Voet, Gordon Saksena, Nicolas Stransky, Robert C Onofrio, Wendy Winckler, Kristin Ardlie, Nikhil Wagle, Jennifer Wargo, Kelly Chong, Don-

ald L Morton, Katherine Stemke-Hale, Guo Chen, Michael Noble, Matthew Meyerson, John E Ladbury, Michael a Davies, Jeffrey E Gershenwald, Stephan N Wagner, Dave S B Hoon, Dirk Schadendorf, Eric S Lander, Stacey B Gabriel, Gad Getz, Levi a Garraway, and Lynda Chin. A landscape of driver mutations in melanoma. *Cell*, 150(2):251–63, jul 2012.

[102] Matthew Holderfield, Hanne Merritt, John Chan, Marco Wallroth, Laura Tandeske, Huili Zhai, John Tellew, Stephen Hardy, Mohammad Hekmat-nejad, Darrin D Stuart, Frank Mccormick, and Tobi E Nagel. RAF Inhibitors Activate the MAPK Pathway by Relieving Inhibitory Autophosphorylation. *Cancer Cell*, 23(5):594–602, 2013.

[103] Jiancheng Hu, Lalima G Ahuja, Hiruy S Meharena, Natarajan Kannan, Alexandr P Kornev, Susan S Taylor, and Andrey S Shaw. Kinase regulation by hydrophobic spine assembly in cancer. *Molecular and cellular biology*, 35(1):264–76, 2015.

[104] Jiancheng Hu, Edward C Stites, Haiyang Yu, Elizabeth A Germino, Hiruy S Meharena, Philip J S Stork, Alexandr P Kornev, Susan S Taylor, and Andrey S Shaw. Allosteric Activation of Functionally Asymmetric RAF Kinase Dimers. *Cell*, 154(5):1036–1046, 2013.

[105] Ruili Huang, Anders Wallqvist, and David G Covell. Targeting changes in cancer: assessing pathway stability by comparing pathway gene expression coherence levels in tumor and normal tissues. *Molecular cancer therapeutics*, 5(9):2417–27, sep 2006.

[106] Zhifeng Huang, Huaibin Chen, Steven Blais, Thomas a Neubert, Xiaokun Li, and Moosa Mohammadi. Structural Mimicry of A-Loop Tyrosine Phosphorylation by a Pathogenic FGF Receptor 3 Mutation. *Structure (London, England : 1993)*, 21(10):1889–96, oct 2013.

[107] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD : Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[108] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[109] Morgan Huse and John Kuriyan. The Conformational Plasticity of Protein Kinases. *Cell*, 109(3):275–282, may 2002.

[110] Roxana E Iacob, Teodora Pene-dumitrescu, Jianming Zhang, Nathanael S Gray, Thomas E Smithgall, and John R Engen. Conformational disturbance in Abl kinase upon mutation and deregulation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1386–1391, 2009.

[111] Tsuneo Ikenoue, Yohko Hikiba, Fumihiko Kanai, Jun Aragaki, Yasuo Tanaka, Jun Imamura, Takaaki Imamura, Miki Ohta, Hideaki Ijichi, Keisuke Tateishi, Takayuki Kawakami, Masayuki Matsumura, Takao Kawabe, and Masao Omata. Different Effects of Point Mutations within the B-Raf Glycine-Rich Loop in Colorectal Tumors on Mitogen-Activated Protein / Extracellular Signal- Regulated Kinase Kinase / Extracellular Signal-Regulated Kinase and Nuclear Factor B Pathway and Cellular T. (30):3428–3435, 2004.

[112] Tsuneo Ikenoue, Yohko Hikiba, Fumihiko Kanai, Yasuo Tanaka, Jun Imamura, Takaaki Imamura, Miki Ohta, Hideaki Ijichi, Keisuke Tateishi, Takayuki Kawakami, Jun Aragaki, Masayuki Matsumura, Takao Kawabe, and Masao Omata. Advances in Brief Functional Analysis of Mutations within the Kinase Activation Segment of B-Raf in Human Colorectal Tumors. *Cancer research*, 63:8132–8137, 2003.

[113] Tsuneo Ikenoue, Fumihiko Kanai, Yohko Hikiba, Toshiyuki Obata, Yasuo Tanaka, Jun Imamura, Miki Ohta, Amarsanaa Jazag, Bayasi Guleng, Keisuke Tateishi, Yoshinari Asaoka, Masayuki Matsumura, Takao Kawabe, and Masao Omata. Functional analysis of PIK3CA gene mutations in human colorectal cancer. *Cancer research*, 65(11):4562–7, jun 2005.

[114] I. Illich. *Medical Nemesis: The Expropriation of Health*. Ideas in progress. Calder & Boyars, 1975.

[115] John P A Ioannidis. Why Most Published Research Findings Are False. 2(8), 2005.

[116] K Isozaki, B Terris, J Belghiti, S Schiffmann, S Hirota, and J M Vanderwinden. Germline-activating mutation in the kinase domain of KIT gene in familial gastrointestinal stromal tumors. *The American journal of pathology*, 157(5):1581–5, nov 2000.

[117] Jose M G Izarzugaza, Angela del Pozo, Miguel Vazquez, and Alfonso Valencia. Prioritization of pathogenic mutations in the protein kinase superfamily. *BMC genomics*, 13 Suppl 4:S3, jan 2012.

[118] Kevin A James and Gennady M Verkhivker. Structure-Based Network Analysis of Activation Mechanisms in the ErbB Family of Receptor Tyrosine Kinases : The Regulatory Spine Residues Are Global Mediators of Structural Stability and Allosteric Interactions. *PLOS ONE*, 9(11):1–46, 2014.

[119] Sunhwan Jo, Taehoon Kim, and Wonpil Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2(9), 2007.

[120] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. CHARMM-GUI : A Web-Based Graphical User Interface for CHARMM. *Journal of computational chemistry*, 29:1859–1865, 2008.

[121] Thorsten Joachims. A Support Vector Method for Multivariate Performance Measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1–8, Bonn, Germany, 2005.

[122] Thorsten Joachims. Training linear SVMs in linear time. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 217, 2006.

[123] Thorsten Joachims and Chun Nam John Yu. Sparse kernel SVMs via cutting-plane training. *Machine Learning*, 76(2-3):179–193, 2009.

[124] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[125] Hye Sook Jung, Dong Wook Kim, Young Suk Jo, Hyo Kyun Chung, Jung Hun Song, Jong Sun Park, Ki Cheol Park, Su Hyeon Park, Jung Hwan Hwang, Ki-won Jo, and Minho Shong. Regulation of Protein Kinase B Tyrosine Phosphorylation by Thyroid-Specific Oncogenic. *Molecular endocrinology (Baltimore, Md.)*, 19(11):2748–2759, 2005.

[126] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[127] Joshua S Kaminker, Yan Zhang, Allison Waugh, Peter M Haverty, Brock Peters, Dragan Sebisanovic, Jeremy Stinson, William F Forrest, J Fernando Bazan, Somasekar Seshagiri, and Zemin Zhang. Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer research*, 67(2):465–73, jan 2007.

[128] Rick Kamps, Rita D Brandão, Bianca J van den Bosch, Aimee D C Paulussen, Sofia Xanthoulea, Marinus J Blok, and Andrea Romano. Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *International Journal of Molecular Sciences*, 18(2), 2017.

[129] Rama Krishna Kancha, Christian Peschel, and Justus Duyster. The Epidermal Growth Factor Receptor-L861Q Mutation Increases Kinase Activity without Leading to Enhanced Sensitivity Toward Epidermal Growth Factor Receptor. *Journal of Thoracic Oncology*, 6(2):387–392, 2011.

[130] Andrew L Kau and Phillip E Korenblat. Kepler WebView: A Lightweight, Portable Framework for Constructing Real-time Web Interfaces of Scientific Workflows. *Procedia Computer Science*, 14(6):570–575, 2015.

[131] Shyam M Kavuri, Naveen Jain, Francesco Galimi, Francesca Cottino, Simonetta M Leto, Giorgia Migliardi, Adam C Searleman, Wei Shen, John Monsey, Livio Trusolino, Samuel A Jacobs, Andrea Bertotti, and Ron Bose. HER2 Activating Mutations Are Targets for Colorectal Cancer Treatment. *Cancer Discovery*, (August), 2015.

[132] Bee Luan Khoo, Parthiv Kant Chaudhuri, Naveen Ramalingam, Daniel Shao Weng Tan, Chwee Teck Lim, and Majid Ebrahimi Warkiani. Single-cell profiling approaches to probing tumor heterogeneity. *International Journal of Cancer*, 139(2):243–255, 2016.

[133] Phillip P Knowles, Judith Murray-rust, Svend Kjær, Rizaldy P Scott, Sarah Hanrahan, Massimo Santoro, and Carlos F Iba. Structure and Chemical Inhibition of the RET Tyrosine Kinase Domain. *Journal of Biological Chemistry*, 281(44):33577–33587, 2006.

[134] Sherry L. Kochanek, Kenneth D.and Murphy, Jiaquan Xu, and Betzaida Tejada-Vera. National Vital Statistics Report. 65(4):1–122, 2016.

[135] Alexandr P Kornev, Nina M Haste, Susan S Taylor, and Lynn F Ten Eyck. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17783–8, nov 2006.

[136] Alexandr P Kornev and Susan S Taylor. Dynamics-Driven Allostery in Protein Kinases. *Trends in Biochemical Sciences*, 40(11):628–647, 2015.

[137] Pamela K Kreeger and Douglas a Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, jan 2010.

[138] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[139] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, 1970.

[140] Ambuj Kumar and Rituraj Purohit. Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs. *PLoS computational biology*, 10(4):1–14, 2014.

[141] Ambuj Kumar and Vidya Rajendran. Computational Investigation of Cancer-Associated Molecular Mechanism in Aurora A (S155R) Mutation. *Cell Biochemistry and Biophysics*, 66(3):787–796, 2013.

[142] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[143] Shenshen Lai, Steven Pelech, and Jonathan Chernoff. Regulatory roles of conserved phosphorylation sites in the activation T-loop of the MAP kinase. *Molecular biology of the cell*, 27:1040–1050, 2016.

[144] Elodie Laine, Isaure Chauvot de Beauchêne, David Perahia, Christian Auclair, and Luba Tchertanov. Mutation D816V alters the internal structure and dynamics of c-KIT receptor cytoplasmic region: implications for dimerization and activation mechanisms. *PLoS computational biology*, 7(6):e1002068, jun 2011.

[145] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[146] Christian C Lee, Yong Jia, Nanxin Li, Xiuying Sun, Kenneth Ng, Eileen Ambing, Muyun Gao, Su Hua, Connie Chen, Sungjoon Kim, Pierre-yves Michellys, Scott A Lesley, Jennifer L Harris, and Glen Spraggon. Crystal structure of the ALK (anaplastic lymphoma kinase) catalytic domain. *Biochemical Journal*, 437:425–437, 2010.

[147] Mark a Lemmon and Joseph Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, 141(7):1117–34, jun 2010.

[148] Yen-Lin Lin, Yilin Meng, Wei Jiang, and Benoît Roux. Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 110(5):1664–9, jan 2013.

[149] Kresten Lindorff-larsen and David E Shaw. How Fast-Folding Proteins Fold. *Science*, 334(October):517–521, 2011.

[150] Peter Littlefield, Lijun Liu, Venkatesh Mysore, Yibing Shan, David E Shaw, and Natalia Jura. Structural analysis of the EGFR / HER3 heterodimer reveals the molecular basis for activating HER3 mutations. *Science Signaling*, 7(354):1–13, 2014.

[151] Lawrence a Loeb. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature reviews. Cancer*, 11(6):450–7, jun 2011.

[152] Silvia Lovera, Ludovico Sutto, Ralitza Boubeva, Leonardo Scapozza, Nicole Do, and Francesco L Gervasio. The Different Flexibility of c-Src and c-Abl Kinases Regulates the Accessibility of a Druggable Inactive Conformation. *Journal of the American Chemical Society*, 134:2496–2499, 2012.

[153] Donald A. MacKenzie. *An Engine, Not a Camera: How Financial Models Shape Markets*. Inside technology. MIT Press, 2006.

[154] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins <sup>†</sup>. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[155] Tomasz Makarewicz and Rajmund Kaźmierkiewicz. Molecular Dynamics Simulation by GROMACS Using GUI Plugin for PyMOL. *Journal of Chemical Information and Modeling*, 53(5):1229–1234, 2013.

[156] G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The Protein Kinase Complement of the Human Genome. *Science*, 298(5600):1912–1934, 2002.

[157] Kristen A Marino, Ludovico Sutto, and Francesco Luigi Gervasio. The E ff ect of a Widespread Cancer-Causing Mutation on the Inactive to Active Dynamics of the B-Raf Kinase. *Journal of the American Chemical Society*, 137:5280–5283, 2015.

[158] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H de Vries. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[159] Larry R Masterson, Cecilia Cheng, Tao Yu, Marco Tonelli, Alexandr Kornev, Susan S Taylor, and Gianluigi Veglia. Dynamics connect substrate recognition to catalysis in protein kinase A. *Nature chemical biology*, 6:821–828, 2010.

[160] Christopher P Mattison, William M Old, Estelle Steiner, Brenda J Huneycutt, Katheryn A Resing, Natalie G Ahn, and Mark Winey. Mps1 Activation Loop Autophosphorylation Enhances Kinase Activity. *Journal of Biological Chemistry*, 282(42):30553–30561, 2007.

[161] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.

[162] Daniel Ian McSkimmin. Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics*, pages 1–15, 2017.

[163] Rosa Marina Melillo, Anna Maria Cirafici, Valentina De Falco, Marie Bellantoni, Gennaro Chiappetta, Alfredo Fusco, Francesca Carlomagno, Antonella Picascia, Donatella Tramontano, Giovanni Tallini, and Massimo Santoro. The Oncogenic Activity of RET Point Mutants for Follicular Thyroid Cells May Account for the Occurrence of Papillary Thyroid Carcinoma in Patients Affected by Familial Medullary Thyroid Carcinoma. 165(2):511–521, 2004.

[164] Yilin Meng, Yen-lin Lin, and Benoît Roux. Computational Study of the “ DFG-Flip ” Conformational Transition in c-Abl and c-Src Tyrosine Kinases. *The journal of physical chemistry. B*, 119:1443–1456, 2015.

[165] Naveen Michaud-agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. MDAnalysis : A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of computational chemistry*, 32:2319–2327, 2011.

[166] C Miething, S Feihl, C Mugler, R Grundler, N Von Bubnoff, F Lordick, C Peschel, and J Duyster. The Bcr-Abl mutations T315I and Y253H do not confer a growth advantage in the absence of imatinib. *Leukemia*, 20:650–657, 2006.

[167] T. Mitchell. *Rule of Experts: Egypt, Techno-Politics, Modernity.* XIII, 413 p. University of California Press, 2002.

[168] Kensaku Mizuno. Rho-associated Kinase ROCK Activates LIM-kinase 1 by Phosphorylation at Threonine 508 within the Activation Loop. *Journal of Biological Chemistry*, 275(5):3577–3582, 2000.

[169] G Monsel, N Ortonne, M Bagot, A Bensussan, and N Dumaz. c-Kit mutants require hypoxia-inducible factor 1 a to transform melanocytes. *Oncogene*, 29(2):227–236, 2009.

[170] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008.

[171] Maria Agnese Morando, Giorgio Saladino, Nicola D Amelio, Encarna Pucheta-martinez, Silvia Lovera, Moreno Lelli, and Blanca López-méndez. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Scientific Reports*, (December 2015):1–9, 2016.

[172] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–4, apr 2011.

[173] Nathan L Nehrt, Thomas a Peterson, DoHwan Park, and Maricel G Kann. Domain landscapes of somatic mutations in cancer. *BMC genomics*, 13 Suppl 4(Suppl 4):S9, jan 2012.

[174] P. C. Ng. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, jul 2003.

[175] P C Ng and S Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–74, may 2001.

[176] Nicola Normanno, Anna Maria Rachiglio, Cristin Roma, Francesca Fenizia, Claudia Esposito, Raffaella Pasquale, Maria Libera La Porta, Alessia Iannaccone, Filippo Micheli, Michele Santangelo, Francesca Bergantino, Susan Costantini, and Antonella De Luca. Molecular diagnostics and personalized medicine in oncology: challenges and opportunities. *Journal of cellular biochemistry*, 114(3):514–24, mar 2013.

[177] Peter Nowell and David A. Hungerford. *Science*, pages 1488–1501 title = A minute chromosome in human chronic granulocytic leukemia, number = 3438, volume = 132, year = 1960.

[178] Peter C Nowell. The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection. *Science*, 194(4260):23–28, 1976.

[179] Ikuko Omori, Hiroki Yamaguchi, Koichi Miyake, Noriko Miyake, and Tomoaki Kitano. D816V mutation in the KIT gene activation loop has greater cell-proliferative and anti-apoptotic ability than N822K mutation in core-binding factor acute myeloid leukemia. *Experimental Hematology*, 52(exon 8):56–64, 2017.

[180] Himanshu Paliwal and Michael R Shirts. Using Multistate Reweighting to Rapidly and Efficiently Explore Molecular Simulation Parameters Space for Nonbonded Interactions. *Journal of Chemical Theory and Computation*, 9(11):4700–4717, 2013.

[181] Szilárd Páll, Carsten Kutzner, Mark James Abraham, Berk Hess, and Erik Lindahl. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. *Proc. of EASC 2015 LNCS*, 8579:3–27, 2015.

[182] Jin H Park, Yingting Liu, Mark a Lemmon, and Ravi Radhakrishnan. Erlotinib binds both inactive and active conformations of the EGFR tyrosine kinase domain. *The Biochemical journal*, 448(3):417–23, dec 2012.

[183] Wendy T Parker, Rebecca M Lawrence, Musei Ho, Darryl L Irwin, Hamish S Scott, and Timothy P Hughes. Sensitive Detection of BCR-ABL1 Mutations in Patients With Chronic Myeloid Leukemia After Imatinib Resistance Is Predictive of Outcome During Subsequent Therapy. *Journal of Clinical Oncology*, 29(32):4250–4259, 2011.

[184] Wendy T Parker, David T O Yeung, Alexandra L Yeoman, Haley K Altamura, Bronte A Jamison, Chani R Field, J Graeme Hodgson, Stephanie Lustgarten, Victor M Rivera, Timothy P Hughes, and Susan Branford. The impact of multiple low-level BCR-ABL1 mutations on response to ponatinib. *Blood*, 127(15):1870–1881, 2016.

[185] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[186] Harshnira Patani, Tom D Bunney, Nethaji Thiyagarajan, Richard A Norman, Derek Ogg, Jason Breed, Paul Ashford, Andrew Potterton, Mina Edwards, Sarah V Williams, Gary S Thomson, Camilla S M Pang, Margaret A Knowles, Alexander L Breeze, Christine Orengo, Chris Phillips, and Matilda Katan. Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use. *Oncotarget*, 7(17):24252–24268, 2016.

[187] Amish J Patel and Shekhar Garde. Efficient method to characterize the context-dependent hydrophobicity of proteins. *The journal of physical chemistry. B*, 118(6):1564–73, feb 2014.

[188] R.K. Pathria and P.D. Beale. *Statistical Mechanics*. Elsevier Science, 2011.

[189] Karl Pearson. *The Genetical Theory of Natural Selection*. 1909.

[190] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[191] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007.

[192] Audrey Petitjean, Ewy Mathe, Shunsuke Kato, Chikashi Ishioka, Sean V. Tavtigian, Pierre Hainaut, and Magali Olivier. TP53 Impact of Mutant p53 Functional Properties on Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database. *Human mutation*, 28(6):622–629, 2007.

[193] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–802, dec 2005.

[194] Stefano Piana, Kresten Lindorff-larsen, and David E Shaw. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5915–5920, 2013.

[195] Stefano Piana and David E Shaw. Atomistic Description of the Folding of a Dimeric Protein. *Journal of Physical Chemistry B*, 117:12935–12942, 2013.

[196] Iván Plaza-Menacho, Karin Barnouin, Rachael Barry, Pascal Meier, Neil Q Mcdonald, Karin Barnouin, Rachael Barry, Annabel Borg, Mariam Orme, and Rakhee Chauhan. RET Functions as a Dual-Specificity Kinase that Requires Allosteric Inputs from Juxtamembrane Article RET Functions as a Dual-Specificity Kinase that Requires Allosteric Inputs from Juxtamembrane Elements. *Cell*, 17:3319–3332, 2016.

[197] Iván Plaza-Menacho, Karin Barnouin, Kerry Goodman, Rubén J Martínez-Torres, Annabel Borg, Judith Murray-Rust, Stephane Mouilleron, Phillip Knowles, and Neil Q McDonald. Oncogenic RET kinase domain mutations perturb the autophosphorylation trajectory by enhancing substrate presentation in trans. *Molecular cell*, 53(5):738–51, mar 2014.

[198] Iván Plaza-Menacho, Andrea Morandi, Luca Mologni, Piet Boender, Carlo Gambacorti-passerini, Anthony I Magee, Robert M W Hofstra, Phillip Knowles, Neil Q Mcdonald, and Clare M Isacke. Focal Adhesion Kinase (FAK) Binds RET Kinase via Its FERM Domain , Priming a Direct and Reciprocal RET-FAK Transactivation Mechanism. *Journal of Biological Chemistry*, 286(19):17292–17302, 2011.

[199] Steve Plimpton. Fast Parallel Algorithms for Short – Range Molecular Dynamics. *Journal of computational Physics*, 117:1–19, 1995.

[200] Tirso Pons, Miguel Vazquez, María Luisa Matey-hernandez, Søren Brunak, and Alfonso Valencia. KinMutRF : a random forest classifier of sequence variants in the human protein kinase superfamily. *BMC Genomics*, 17(Suppl 2):207–217, 2016.

[201] Poulikos I Poulikakos, Chao Zhang, Gideon Bollag, Kevan M Shokat, and Neal Rosen. RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. *Nature*, 464(7287):427–430, 2010.

[202] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not : how much can we rely on published data on potential drug targets ? *Nature reviews. Drug discovery*, 10(712), 2011.

[203] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David Van Der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[204] Shweta Purawat, Pek U Ieong, Robert D Malmstrom, Garrett J Chan, Alan K Yeung, Ross C Walker, Ilkay Altintas, and Rommie E Amaro. Computational Tool A Kepler Workflow Tool for Reproducible AMBER GPU Molecular Dynamics. *Biophysj*, 112(12):2469–2474, 2017.

[205] Jens Rauch, Natalia Volinsky, David Romano, and Walter Kolch. The secret life of kinases : functions beyond catalysis. *Cell Communications and Signalling*, 22:1–28, 2011.

[206] Monica Red Brewer, Sung Hee Choi, Diego Alvarado, Katarina Moravcevic, Ambra Pozzi, Mark A Lemmon, and Graham Carpenter. The juxtamembrane region of the EGF receptor functions as an activation domain. *Molecular cell*, 34(6):641–51, jun 2009.

[207] E Premkumar Reddy and Aneel K Aggarwal. The ins and outs of bcr-abl inhibition. *Genes & cancer*, 3(5-6):447–54, may 2012.

[208] Jüri Reimand, Omar Wagih, and Gary D. Bader. The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, 3(1):2651, 2013.

[209] Peter Rellos, Frank J Ivins, Joanne E Baxter, Ashley Pike, Timothy J Nott, Donna-marie Parkinson, Sanjan Das, Steven Howell, Oleg Fedorov, Qi Yu Shen, Andrew M Fry, Stefan Knapp, and Stephen J Smerdon. Structure and Regulation of the Human Nek2. *Journal of Biological Chemistry*, 282(9):6833–6842, 2007.

[210] Julie Y Reuther, Gary W Reuther, David Cortez, Ann Marie Pendergast, and Albert S Baldwin. A requirement for NF-κB activation in Bcr – Abl-mediated transformation. *Genes & development*, 12:968–981, 1998.

[211] Boris Reva, Yevgeniy Antipin, and Chris Sander. Determinants of protein function revealed by combinatorial entropy optimization. *Genome biology*, 8(11):R232, jan 2007.

[212] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118, sep 2011.

[213] João V Ribeiro, Rafael C Bernardi, Till Rudack, John E Stone, James C Phillips, Peter L Freddolino, and Klaus Schulten. QwikMD — Integrative Molecular Dynamics Toolkit for Novices and Experts. *Scientific Reports*, 6(26536), 2016.

[214] Sanjit Roopra, Bernhard Knapp, Ulrich Omasits, and Wolfgang Schreiner. jSimMacs for GROMACS: A Java Application for Advanced Molecular Dynamics Simulations with Remote Access Capability. *Journal of Chemical Information and Modeling*, 49(10):2412–2417, 2009.

[215] Burkhard Rost, Chris Sander, and Reinhard Schneider. PHD – an Automatic Mail Server for Protein Secondary Structure Prediction. *Computer applications in the biosciences : CABIOS*, 10:53–60, 1994.

[216] A Sali and TL Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Protein structure by distance analysis*, 234(3):779–815, 1994.

[217] Vassiliki Saloura, Theodore Vougiouklakis, Makda Zewde, Xiaolan Deng, Kazuma Kiyotani, Jae-Hyun Park, Yo Matsuo, Mark Lingen, Takehiro Suzuki, Naoshi Dohmae, Ryuji Hamamoto, and Yusuke Nakamura. WHSC1L1-mediated EGFR mono-methylation enhances the cytoplasmic and nuclear oncogenic activity of EGFR in head and neck cancer. *Scientific reports*, 7(January):40664, 2017.

[218] K. Scheffzek. The Ras-RasGAP Complex: Structural Basis for GTPase Activation and Its Loss in Oncogenic Ras Mutants. *Science*, 277(5324):333–338, jul 1997.

[219] Sjors H W Scheres. RELION : Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012.

[220] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[221] Diamantis Sellis, Dimitrios Vlachakis, and Metaxia Vlassi. Gromita: A fully integrated graphical user interface to gromacs 4. *Bioinformatics and Biology Insights*, 3:99–102, 2009.

[222] A. Sen. *Development as Freedom.* Oxford India paperbacks. Oxford University Press, 1999.

[223] Yibing Shan, Michael P Eastwood, Xuewu Zhang, Eric T Kim, Anton Arkhipov, Ron O Dror, John Jumper, John Kuriyan, and David E Shaw. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 149(4):860–70, may 2012.

[224] Yibing Shan, Michael P Eastwood, Xuewu Zhang, Eric T Kim, Anton Arkhipov, Ron O Dror, John Jumper, John Kuriyan, and David E Shaw. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 149(4):860–70, may 2012.

[225] Yibing Shan, Markus a Seeliger, Michael P Eastwood, Filipp Frank, Huafeng Xu, Morten Ø Jensen, Ron O Dror, John Kuriyan, and David E Shaw. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):139–44, jan 2009.

[226] S. Shapin and S. Schaffer. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life.* Princeton Classics. Princeton University Press, 2011.

[227] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, New York, NY, USA, 2004.

[228] Fumin Shi, Shannon E Telesco, Yingting Liu, Ravi Radhakrishnan, and Mark A Lemmon. ErbB3/HER3 intracellular domain is competent to bind ATP and catalyze autophosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17):7692–7, apr 2010.

[229] Andrew J Shih, Shannon E Telesco, Sung-Hee Choi, Mark a Lemmon, and Ravi Radhakrishnan. Molecular dynamics analysis of conserved hydrophobic and hydrophilic bond-interaction networks in ErbB family kinases. *The Biochemical journal*, 436(2):241–51, jun 2011.

[230] Brian J Skaggs, Mercedes E Gorre, Ann Ryvkin, Michael R Burgess, Yongming Xie, Yun Han, Evangelia Komisopoulou, Lauren M Brown, Joseph A Loo, Elliot M Landaw, Charles L Sawyers, and Thomas G Graeber. Phosphorylation of the ATP-binding loop directs oncogenicity of drug-resistant BCR-ABL mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51):19466–19471, 2006.

[231] Matthew J Smith, Benjamin G Neel, and Mitsuhiko Ikura. NMR-based functional profiling of RASopathies and oncogenic RAS mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(12):4574–9, mar 2013.

[232] Simona Soverini, Antonella Vitale, Angela Poerio, Alessandra Gnani, Sabrina Colarossi, Ilaria Iacobucci, Giuseppe Cimino, Loredana Elia, Annalisa Lonetti, Marco Vignetti, Stefania Paolini, Giovanna Meloni, Valeria Maio, Cristina Papayannidis, Marilina Amabile, Anna Guarini, Michele Baccarani, Giovanni Martinelli, Robin Foà, and San Paolo. Philadelphia-positive acute lymphoblastic leukemia patients already harbor BCR-ABL kinase domain mutations at low levels at the time of diagnosis. *Haematologica*, 96(4):552–557, 2011.

[233] Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, Lucy R Yates, Elli Papaemmanuil, David Beare, Adam Butler, Angela Cheverton, John Gamble, Jonathan Hinton, Mingming Jia, Alagu Jayakumar, David Jones, Calli Latimer, King Wai Lau, Stuart McLaren, David J McBride, Andrew Menzies, Laura Mudie, Keiran Raine, Roland Rad, Michael Spencer Chapman, Jon Teague, Douglas Easton, Anita Langerød, Ming Ta Michael Lee, Chen-Yang Shen, Benita Tan Kiat Tee, Bernice Wong Huimin, Annegien Broeks, Ana Cristina Vargas, Gulisa Turashvili, John Martens, Aquila Fatima, Penelope Miron, Suet-Feung Chin, Gilles Thomas, Sandrine Boyault, Odette Mariani, Sunil R Lakhani, Marc van de Vijver, Laura van 't Veer, John Foekens, Christine Desmedt, Christos Sotiriou, Andrew Tutt, Carlos Caldas, Jorge S Reis-Filho, Samuel a J R Aparicio, Anne Vincent Salomon, Anne-Lise Børresen-Dale, Andrea L Richardson, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–4, jun 2012.

[234] Eric a Stone and Arend Sidow. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome research*, 15(7):978–86, jul 2005.

[235] Jianmin Sun, Malin Pedersen, and Lars Rönnstrand. The D816V mutation of c-Kit circumvents a requirement for Src family kinases in c-Kit signal transduction. *The Journal of biological chemistry*, 284(17):11039–47, apr 2009.

[236] Ilknur Sur, Sascha Neumann, and Angelika A Noegel. Nesprin-1 role in DNA damage response. *Nucleus*, 5(2):173–191, 2014.

[237] Ludovico Sutto and Francesco Luigi. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proceedings of the National Academy of Sciences of the United States of America*, pages 0–5, 2013.

[238] Mario L Suvà, Nicolo Riggi, and Bradley E Bernstein. Epigenetic reprogramming in cancer. *Science (New York, N.Y.)*, 339(6127):1567–70, mar 2013.

[239] Susan S Taylor, Andrey S Shaw, Natarajan Kannan, and Alexandr P Kornev. Integration of signaling in the kinome : Architecture and regulation of the α C Helix. *BBA - Proteins and Proteomics*, 1854(10):1567–1574, 2015.

[240] The HDF Group. Hierarchical data format version 5, 2000-2010.

[241] Tianhai Tian, Sarah Olson, James M Whitacre, and Angus Harding. The origins of cancer robustness and evolvability. *Integrative biology : quantitative biosciences from nano to macro*, 3(1):17–30, jan 2011.

[242] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355:1330–1334, 2017.

[243] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 345:78–81, 2015.

[244] Ali Torkamani and Nicholas J Schork. Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics (Oxford, England)*, 23(21):2918–25, nov 2007.

[245] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*, 16(5):62–74, 2014.

[246] Torsten Trowe, Sotiria Boukouvala, Keith Calkins, Richard E. Cutler, Ryan Fong, Roel Funke, Steven B. Gendreau, Yong D. Kim, Nicole Miller, John R. Woolfrey, Valentina Vysotskaia, Ping Yang Jing, Mary E. Gerritsen, David J. Matthews, Peter Lamb, and Timothy S. Heuer. EXEL-7647 inhibits mutant forms of ErbB2 associated with lapatinib resistance and neoplastic transformation. *Clinical Cancer Research*, 14(8):2465–2475, 2008.

[247] Daniela Ungureanu, Jinhua Wu, Tuija Pekkala, Yashavanthi Niranjan, Clifford Young, Ole N Jensen, Chong-Feng Xu, Thomas a Neubert, Radek C Skoda, Stevan R Hubbard, and Olli Silvennoinen. The pseudokinase domain of JAK2 is a dual-specificity protein kinase that negatively regulates cytokine signaling. *Nature structural & molecular biology*, 18(9):971–6, sep 2011.

[248] Stéfan van der Walt, S Chris Colbert, and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[249] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1st edition, 1995.

[250] Harish Vashisth, Luca Maragliano, and Cameron F Abrams. "DFG-flip" in the insulin receptor kinase is facilitated by a helical intermediate state of the activation loop. *Biophysical journal*, 102(8):1979–87, apr 2012.

[251] R S K Vijayan, Peng He, Vivek Modi, Krisna C Duong-ly, Haiching Ma, R Peterson, Roland L Dunbrack, and Ronald M Levy. Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. *Journal of Medicinal Chemistry*, 58:466–479, 2015.

[252] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, and Luis A. Diaz Jr. Cancer Genome Landscapes. *Science*, 339:1546, 2013.

[253] Martin Vogtherr, Krishna Saxena, Swen Hoelder, Susanne Grimme, Marco Betz, Ulrich Schieborr, Barbara Pescatore, Michel Robin, Laure Delarbre, Thomas Langer, K Ulrich Wendt, and Harald Schwalbe. NMR Characterization of Kinase p38 Dynamics in Free and Ligand-Bound Forms**. *Angewandte Chemie*, 45:993–997, 2006.

[254] Paul T C Wan, Mathew J Garnett, S Mark Roe, Sharlene Lee, Dan Niculescu-Duvaz, Valerie M Good, C Michael Jones, Christopher J Marshall, Caroline J Springer, David Barford, and Richard Marais. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, 116(6):855–67, mar 2004.

[255] Shunzhou Wan and Peter V Coveney. Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. *Journal of computational chemistry*, 32(13):2843–52, oct 2011.

[256] Shunzhou Wan, David W Wright, and Peter V Coveney. Mechanism of Drug Efficacy Within the EGF Receptor Revealed by Microsecond Molecular Dynamics Simulation. *Molecular Cancer Therapeutics*, 11(11):2394–2401, 2012.

[257] Jiguang Wang, Emanuela Cazzato, Erik Ladewig, Veronique Frattini, Daniel I S Rosenbloom, Sakellarios Zairis, Francesco Abate, Zhaoqi Liu, Oliver Elliott, Yong-Jae Shin, Jin-Ku Lee, In-Hee Lee, Woong-Yang Park, Marica Eoli, Andrew J Blumberg, Anna Lasorella, Do-Hyun Nam, Gaetano Finocchiaro, Antonio Iavarone, and Raul Rabadan. Clonal evolution of glioblastoma under therapy. *Nature Genetics*, 48(7):768–776, 2016.

[258] N Wang, H Ding, C Liu, X Li, L Wei, J Yu, M Liu, M Ying, W Gao, H Jiang, and Y Wang. A novel recurrent CHEK2 Y390C mutation identified in high-risk Chinese breast cancer patients impairs its activity and is associated with increased breast cancer risk. *Oncogene*, 34:5198–5205, 2015.

[259] Qiong Wei and Roland L. Dunbrack. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, 8(7):e67863, jul 2013.

[260] Qiong Wei, Liqun Wang, Qiang Wang, Warren D Kruger, and Roland L Dunbrack. Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase. *Proteins*, 78(9):2058–74, jul 2010.

[261] Andrew D Whale, Anna Dart, Mark Holt, Gareth E Jones, and Claire M Wells. PAK4 kinase activity and somatic mutation promote carcinoma cell motility and influence inhibitor sensitivity. *Oncogene*, 32(16):2114–2120, 2013.

[262] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P. Balcazar Vargas, Shoaib Sufi, and Carole Goble. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic acids research*, 41(Web Server issue):557–561, 2013.

[263] Qifang Xu and Roland L Dunbrack. The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic acids research*, 39(Database issue):D761–70, jan 2011.

[264] Xiaowei Xu, Carmine De Angelis, Kathleen A. Burke, Agostina Nardone, Huizhong Hu, Lanfang Qin, Jamunarani Veeraraghavan, Vidyalakshmi Sethunath, Laura M. Heiser, Nicholas Wang, Charlotte K.Y. Ng, Edward S. Chen, Alexander Renwick, Tao Wang, Sarmistha Nanda, Martin Shea, Tamika Mitchell, Mahitha Rajendran, Ian Waters, Daniel J. Zabransky, Kenneth L. Scott, Carolina Gutierrez, Chandandeep Nagi, Felipe C. Geyer, Gary C. Chamness, Ben H. Park, Chad A. Shaw, Susan G. Hilsenbeck, Mothaffar F. Rimawi, Joe W. Gray, Britta Weigelt, Jorge S. Reis-Filho, C. Kent Osborne, and Rachel Schiff. HER2 Reactivation through Acquisition of the HER2 L755S Mutation as a Mechanism of Acquired Resistance to HER2-targeted Therapy in HER2 <sup>+</sup> Breast Cancer. *Clinical Cancer Research*, (i):5123–5135, 2017.

[265] Y. Yamamoto. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. *Blood*, 97(8):2434–2439, apr 2001.

[266] Hiroyuki Yasuda, Eunyoung Park, Cai-Hong Yun, Natasha J Sng, Antonio R Lucena-Araujo, Wee-Lee Yeo, Mark S Huberman, David W Cohen, Sohei Nakayama, Kota Ishioka, Norihiro Yamaguchi, Megan Hanna, Geoffrey R Oxnard, Christopher S Lathan, Teresa Moran, Lecia V Sequist, Jamie E Chaft, Gregory J Riely, Maria E Arcila, Ross A Soo, Matthew Meyerson, Michael J Eck, Susumu S Kobayashi, and Daniel B Costa. Structural, biochemical and clinical characterization of epidermal growth factor receptor (EGFR) exon 20 insertion mutations in lung cancer. *Science translational medicine*, 5(216):1946–6234, 2014.

[267] Cai Hong Yun, Titus J. Boggon, Yiqun Li, Michele S. Woo, Heidi Greulich, Matthew Meyerson, and Michael J. Eck. Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. *Cancer Cell*, 11(3):217–227, 2007.

[268] Cai-Hong Yun, Kristen E Mengwasser, Angela V Toms, Michele S Woo, Heidi Greulich, Kwok-Kin Wong, Matthew Meyerson, and Michael J Eck. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2070–5, feb 2008.

[269] Bao-hong Zhang and Kun-liang Guan. Activation of B-Raf kinase requires phosphorylation of the conserved residues Thr598 and Ser601. 19(20):5429–5439, 2000.

[270] Chunzhi Zhang, Lynette M Moore, Xia Li, W K Alfred Yung, and Wei Zhang. IDH1 / 2 mutations target a key hallmark of cancer by deregulating cellular metabolism in glioma. 15(9):1114–1126, 2013.

[271] Xuewu Zhang, Jodi Gureasko, Kui Shen, Philip a Cole, and John Kuriyan. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell*, 125(6):1137–49, jun 2006.

[272] Yan Zhang, Yoshiko Hiraishi, Hua Wang, Ken-saku Sumi, Yasutaka Hayashido, Shigeaki Toratani, and Mikio Kan. Constitutive activating mutation of the FGFR3b in oral squamous cell carcinomas. *International Journal of Cancer*, 117:166–168, 2005.

[273] Wen-jia Zuo, Yi-zhou Jiang, Yu-jie Wang, Xiao-en Xu, Xin Hu, Guang-yu Liu, Jiong Wu, Gen-hong Di, and Ke-da Yu. Dual Characteristics of Novel HER2 Kinase Domain Mutations in Response to HER2-Targeted Therapies in Human Breast Cancer. *Clinical Cancer Research*, 22(19):4859–4869, 2016.